A Hierarchial Model for Visual Perception

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 computational model for visual perception: the visual pathway is considered as a functional process of dimensionality reduction for input data, that is, to search for the *intrinsic* dimensionality of natural scene images. This model is hierarchically constructed, and final leads to the formation of a low-dimensional space called *perceptual manifold*. Further analysis of the perceptual manifold reveals that scene images which share similar perceptual similarities stay nearby in the manifold space, and the dimensions of the space could describe the spatial layout of scenes, which are like the degree of naturalness, openness supervised trained in [1]. Moreover, the implementation of scene retrieval task validates the topographic property of the perceptual manifold space.

1 Introduction

From photoreceptor of retina to the unified scene perception in higher level cortex, the transformations of input signal in visual cortex are very complicated. The neurophysiological studies [2] indicate that through the hierarchical processing of information flow in visual cortex, the extremely high-dimensional input signal is gradually represented by fewer active neurons, which is believed to achieve a sparse coding [3] or efficient coding [4].

Meanwhile, the efficient coding theory proposed by Horace Barlow in 1961 gradually becomes the theoretical principles for nervous system and explain away much neuronal behaviors. One of the key predictions of efficient coding theory is that the visual cortex relies on the environmental statistics to encode and represent the signal [5], that is to say, the neuronal representation is closely related to the intrinsic properties of the input signal. Moreover, studies on the natural image statistics show that the natural images are usually embedded in a relatively low dimensional manifold of pixel space [7], and there is a large amount of information redundancy within the natural images. According to the efficient coding theory, it is assumed that the structure of neural computations should fully adapt to extract the intrinsic low dimensionality of natural image to form the unified scene perception, in spite of the extremely high-dimensional raw sensory input from the retina [2].

Under the efficient coding theory and natural image statistics, we propose a computational model for visual perception: the visual pathway is considered as a functional process of dimensionality reduction for input signal, that is, to remove the information redundancy and to search for the *intrinsic* dimensionality of natural scene images. By pooling together the activity of local low-level feature detectors across large regions of the visual field, we build the population feature representation which is the statistical summary of the input scene image. Then, thousands of population feature representations of scene images are extracted, and to be mapped *unsupervised* to a low-dimensional space called perceptual manifold. Further analysis of this manifold reveal that scene images which share similar perceptual similarity stay nearby in the manifold space, and the dimensions of the manifold could describe continuous changes within the spatial layout of scenes, which are similar to the degree of naturalness and openness supervised trained in [1]. In addition, the implementation of scene retrieval task validates the topographic property of the perceptual manifold space. In the following section, the Perceptual Manifold model is extended in detail.

2 The Hierarchical Model

One of the fundamental properties of visual cortex concerns the ability to integrate the local components of a visual image into a unified perception [2]. Based on the neurobiological mechanism of hierarchical signal processing in visual system, we build a computational architecture called Perceptual Manifold. The architecture includes three hierarchical layers: 1) local feature encoding, 2) population feature encoding and 3) perceptual manifold embedding (refer to Fig.1a).

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 [5]. The standard efficient coding method [9] assumes that the image patch is transformed by a set of linear filters w_i to output response s_i. In matrix form,

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

Or equivalently in terms of a basis set, $\mathbf{x}$=$\mathbf{A}\mathbf{s}$=$\mathbf{W}^{-1}\mathbf{s}$. Then, the filter response s_i are assumed to be statistically independent,

$$p(\mathbf{s}) = \prod_i p(s_i) \propto exp(-\sum_i |s_i|) \tag{2}$$

Specifically, 150000 20×20 gray image patches from natural scenes are used in training. A set of 384 $20{\times}20$ basis function is obtained. These basis functions

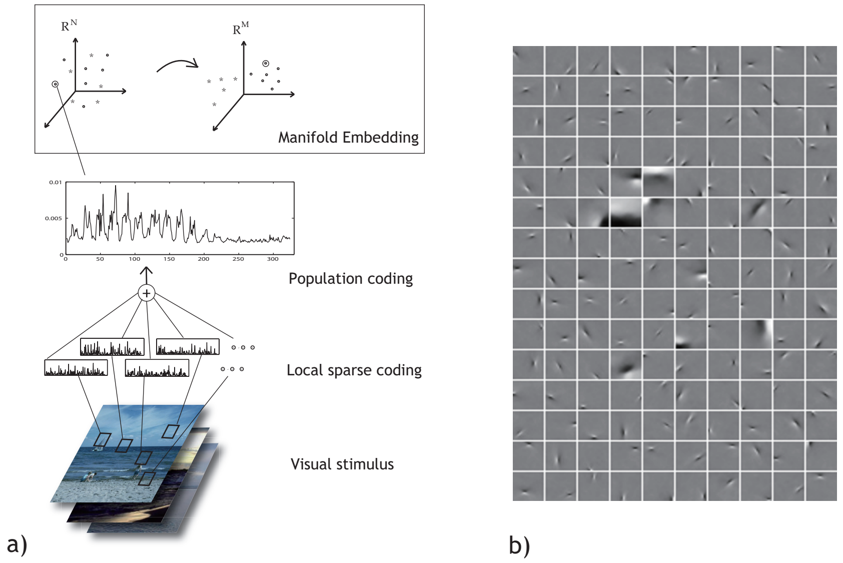

Fig. 1. a) Schematic diagram of the Perceptual Manifold model. There are three hierarchical layers of sensory computation before the final formation of the perceptual space. **b)** A subset of filter basis $\mathbf{W}$ trained from natural image patches, they resemble receptive fields of the simple cells in V1.

resemble the receptive field properties of simple cells, i.e., they are spatially localized, oriented, and band-pass in different spatial frequency bands (Fig.1b). Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{384}]$ be the filter function. 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 the architecture.

2.2 Population Feature Encoding

Higher processing stages are not influenced by single V1 neuron alone but by the activity of a larger population [10]. The neurophysiological study [2] implies that on the population level, a linear pooling mechanism might be used by the visual cortex to extract the global response of the stimulus. In view of this evidence, our computational model includes a second layer of population encoding to sum up feature's response over the local visual fields.

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 scene image. The population feature component for the i^{th} feature is:

$$p_i = \frac{\sum_n |\mathbf{w}_i^\top \mathbf{x}_n|}{\sum_i \sum_n |\mathbf{w}_i^\top \mathbf{x}_n|} \tag{3}$$

Thus, $\mathbf{p} = [p_1, p_2, ..., p_{384}]^\top$ is the population response of scene image.

2.3 Perceptual Space Embedding

Furthermore, neurophysiological studies have often found that the firing rate of each neuron in a population can be written as a smooth function of a small number of variables [10], which supports the idea that the population activity might be constrained to lie on a low-dimensional manifold. To search for the underlying meaningful dimensionality of percept, the Local Linear Embedding [11] is applied as the nonlinear dimensionality reduction method for thousands of the population feature responses of scene images:

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

$$\varepsilon(W) = \sum_i |\mathbf{p}_i - \sum_j W_{ij}\mathbf{p}_j|^2 \tag{4}$$

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

$$\phi(\mathbf{y}) = \sum_i |\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j|^2 \tag{5}$$

The resulting embedding space $\mathbb{R}^M$ is called perceptual space, where vector $\mathbf{y} = [y_1, y_2, ..., y_M]^\top \in \mathbb{R}^M$. The topographic properties of this space would be analyzed in the following experiment section.

3 The Experiment

The dataset of natural images implemented here is from [12], which contains 3890 images from 13 semantic categories of natural scenes, like coast, forest, etc. And image size is normalized as 128×128 pixels.

All 3890 images are used in the embedding process. The dimensionality of manifold space M is tuned as 20, so that the space embedding is $\mathbb{R}^{384} \rightarrow \mathbb{R}^{20}$. Fig.2 shows the embedded sample points described by first three coordinates of the perceptual space, in which points of images from four categories are visualized, and are colored according to the images' scene category. We can clearly see the clustering and nonlinear geometric property of the data points in the space. Moreover, the image retrieval task as follows validates its property.

3.1 Image Retrieval

Image retrieval task is to retrieve the perceptually similar sets of images when given a target image. This task is implemented in our embedded perceptual space and achieves impressive results.

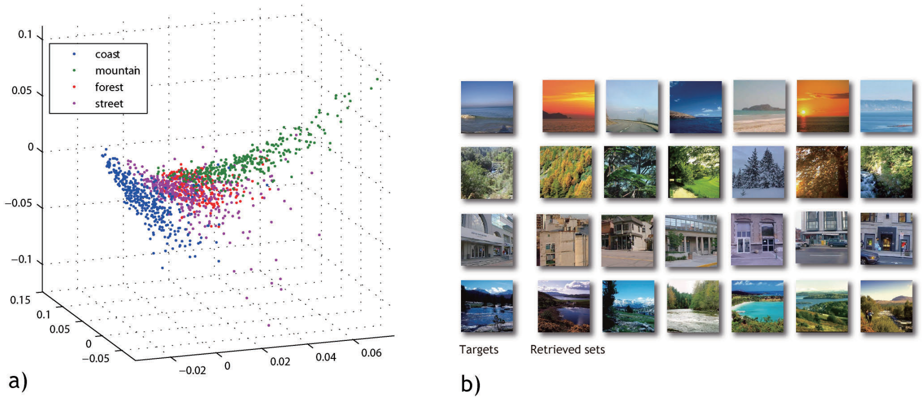

Fig. 2. a) Colored points of scene images from four categories are visualized by the first three coordinates of the perceptual space, the geometric structure among those points of scene images is clearly nonlinear. **b)**Some examples of target images with the sets of retrieved images. Despite the simplicity of the similarity metric used in the experiment, these pairs share close perceptual similarity.

To show the topographic property of the perceptual space, image similarity metric is approximated simply as the Euclidean distance between the target i and retrieved ones j in the perceptual space, that is,

$$D^2(i,j) = \sum_{h=1}^{M} (y_{ih} - y_{jh})^2 \tag{6}$$

Fig. 2 shows examples of target images and retrieved sets of K least metric images, here K is 6. Despite the simplicity of the similarity metric used in the experiment, the set of retrieved images are very similar to the target image.

4 Discussion and Conclusion

The efficiency of the neural code depends both on the transformation that maps the input to the neural responses and on the statistics of the input signal [5].

In this paper, under the theory of efficient coding, a hierarchical framework is developed to explore the intrinsic dimensions of visual perception. By pooling together the activity of local low-level feature detectors across large region of the visual field, we build the population feature representation which is the statistical summary of the input scene image. Then, thousands of population feature representations of scene images are extracted, and mapped *unsupervised* to the low dimensional perceptual space. Analysis of the perceptual space reveals the topographic property that scene images with similar perceptual similarity stay nearby in the embedded space. Moreover, the implementation of scene retrieval task validates the topographic property of the perceptual space.

In addition, it is noteworthy that in the image retrieval task, most of the retrieved images belong to the same scene category of the target image hand-labored in [12]. This would imply the topographic property of the embedded perceptual space corresponds to the semantic concept of scene image, and this interesting property would be discussed in our further work.

5 Acknowledgements

The work was supported by the National Basic Research Program of China (Grant No. 2005CB724301) and the Science and Technology Commission of Shanghai Municipality (Grant No. 08511501701).

References

1. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42** (2001)
2. Weliky, M., Fiser, J., Hunt, R.H., Wagner, D.N.: Coding of natural scenes in primary visual cortex. Neuron **37**(4) (February 2003) 703–718
3. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583) (June 1996) 607–609
4. Olshausen, B.A., Field, D.J.: Sparse coding of sensory inputs. Current Opinion in Neurobiology **14**(4) (August 2004) 481–487
5. Simoncelli, E.P., Olshausen, B.: Natural image statistics and neural representation. Annual Review of Neuroscience **24** (may 2001) 1193–1216
6. Chklovskii, D.B., Mel, B.W., Svoboda, K.: Cortical rewiring and information storage. Nature **431**(7010) (October 2004) 782–788
7. Srivastava, A., Lee, A.B., Simoncelli, E.P., c. Zhu, S.: On advances in statistical modeling of natural images. Journal of Mathematical Imaging and Vision **18** (2003) 17–33
8. Seung, S.H., Lee, D.D.: Cognition: The manifold ways of perception. Science **290**(5500) (December 2000) 2268–2269
9. Bell, A.J., Sejnowski, T.J.: The ”independent components” of natural scenes are edge filters. Vision Res **37**(23) (December 1997) 3327–3338
10. Averbeck, B.B., Latham, P.E., Pouget, A.: Neural correlations, population coding and computation. Nature Reviews Neuroscience **7**(5) 358–366
11. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (December 2000) 2323–2326
12. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR ’05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 2, Washington, DC, USA, IEEE Computer Society (2005) 524–531