Learning Overcomplete Spatiotemporal Bubbles from Natural Image Sequences

Libo Ma, and Liqing Zhang
Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai, China
malibo@sjtu.edu.cn
zhang-lq@cs.sjtu.edu.cn

Abstract

Recently, bubble coding for natural image sequences has been proposed. This method unified three important statistical properties: sparseness, temporal coherence, and topographic dependencies. However, this approach does not consider the overcomplete case. It is widely believed that the overcomplete representation is more efficient than the complete representation. In this paper, we use Bayesian estimation to extend the bubble coding into overcomplete case. Based on a quasi-orthogonality in a high-dimensional space, the prior probability of the mixing matrix is derived. Instead of examining basis coefficient, we investigate the dot product between basis functions and whitened observed data vectors for their sparseness and the advantage in the Bayesian model. Based on the bubble detector definition, an approximation of the prior probability of this dot product is given. Simulation results suggest that the overcomplete bubble coding can be achieved by a Bayesian inference. The model is promising in a wide variety of applications, such as image processing and pattern recognition.

1 Introduction

It is widely assumed that neurons in the visual cortex are better tuned to the stimuli that they are more often exposed to. This property, known as the efficient coding hypothesis, has become an important computational principle for the design of sensory systems [1, 2]. Motivated by this hypothesis, many computational models investigated the statistical properties of natural signals. Among these discoveries, three statistical properties have been extensively discussed: sparseness, temporal coherence, and topographic dependencies. By applying sparseness or temporal coherence criteria on natural images (or sequences), the extracted features indicate similarities, such as shape and orientations to the classical receptive fields (CRFs) in V1 in the human visual cortex[11, 5]. By using topography criterion, the energy topographic dependencies of responses of simple cells are modelled by a two-layer network for natural images. This criterion leads to topographic independent component analysis (TICA) model [6]. When TICA model is performed on natural images, the emerging topography is qualitatively similar to the observed properties of complex cells in V1. Furthermore, Hyvärinen et al. based on the concept of bubble-like spatiotemporal activities of neurons, developed a unified framework that combines all these three properties [7]. The bubble in the model means that the responses of simple cells are contiguous both in space and time.

However, the limitation of the bubble coding is that it does not allow for an overcomplete representation – a case where the number of basis functions is larger than the dimensionality of input signals. An important assumption of bubble coding is the existence of the inverse matrix of mixing matrix, which is only suitable for the complete case.

In this paper we propose a Bayesian method to generalize the bubble coding into an overcomplete basis. This method extends bubble coding and ordinary overcomplete ICA model in several ways that are relevant to the efficient coding. First, we reformulate the ordinary overcomplete ICA into a hierarchical fashion, which extends the basic single-layer sparse coding scheme. In comparison, ordinary ICA methods only capture linear structures of the input data since only one stage is involved. In the overcomplete case, the basis functions learned by our method can promise the appearance of this topographic organization. Second, our model yields overcomplete representations while ordinary bubble coding only produces complete representations for natural image sequences. Overcomplete representations generally provide more efficient representations than the complete case, and have been widely used in fields of com-

putational perceptions and pattern recognitions [10]. Previously, some computational models have been proposed for the overcomplete representations [12, 9, 8].

The paper is organized as follows. In the next section, we propose a Bayesian method to estimate basis functions and bubble activity in an overcomplete case for natural image sequences. The gradient descent algorithm for learning the mixing matrix $\mathbf{A}$ is given. In section 3, we apply the model to natural image sequences. In this section, several properties of our basis functions are analyzed. Finally, we discuss the contribution of our work in section 4.

2 Overcomplete Spatiotemporal Bubbles

A simple model for natural images is the linear generative model, where the input data $\mathbf{x}$ are assumed to be generated as a linear transformation of basis functions:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_M)^T$ is a vector of observed data, $\mathbf{a}_i$ is the i^{th} column of the mixing matrix $\mathbf{A}$. $\mathbf{s} = (s_1, s_2, ..., s_N)^T$ is a vector of basis coefficient. In a cortical interpretation, the coefficient s_i models the response of a simple cell, and $\mathbf{A}$ is closely related to the classical receptive fields (CRFs) of neurons [11].

To make the model hierarchical, we consider a two-layer neural network as in bubble model [7]. The simple cells are assumed to be arranged in a 2-D grid. The squared outputs of simple cells are pooled to complex cells in the second layer. The pooling weight between i^{th} complex cell and j^{th} simple cell is described by a neighborhood function $h(i, j)$. Typically, if the cells are close enough to each other, $h(i, j) = 1$; otherwise, $h(i, j) = 0$. In order to incorporate spatiotemporal pooling into consideration, we further formulate the neighborhood function to $\tilde{h}(i, j, \tau) = h(i, j)\phi(\tau)$, where τ is a time lag (delay) and $\phi(\tau)$ is a temporal smoothing kernel. Thus, the output of a bubble detector at point i during time t can be formulated as:

$$b_i(t) = \sum_{\tau} \sum_{j=1}^{N} \tilde{h}(i, j, \tau) s_i^2. \tag{2}$$

2.1 Bayesian Inference

In this paper, we estimate the model by maximum a posteriori (MAP) approach. From a Bayesian viewpoint, the purpose of our model is to estimate most probable basis functions. In other word, we want to maximize the posterior probability of basis functions given natural images sequences. We first whiten the observed data $\mathbf{x}$ to $\mathbf{z} = \mathbf{V}\mathbf{x}$, where $\mathbf{V}$ is the whitening matrix. To factorize the posterior probability of the parameters, we have:

$$P(\mathbf{A}|\mathbf{z}) = \frac{P(\mathbf{z}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{z})}. \tag{3}$$

Note that $P(\mathbf{z})$ does not depends on $\mathbf{A}$. Now we turn to the problem of learning likelihood $P(\mathbf{z}|\mathbf{A})$ and the prior probability of mixing matrix $P(\mathbf{A})$.

2.1.1 Likelihood of The Model

The likelihood $P(\mathbf{z}|\mathbf{A})$ can be derived with several assumptions: (1) the norms of basis vectors are set to unity; (2) the variance of basis coefficients can differ from unity. Now, we examine the vector $\mathbf{y} = (y_1, ..., y_i, ..., y_N)^T = \mathbf{A}^T\mathbf{z}$, where y_i is the dot product between the i^{th} basis vector and the whitened data vector:

$$y_i = \mathbf{a}_i^T\mathbf{z} = \mathbf{a}_i^T\mathbf{A}\mathbf{s} = s_i + \sum_{j \neq i} \mathbf{a}_i^T\mathbf{a}_j s_j. \tag{4}$$

The first item s_i is the i^{th} basis coefficient and the second term is Gaussian especially in the overcomplete case. Therefore, the dot product is very likely to have sparse marginal distributions. We can then place factorable sparse on dot product: $P(\mathbf{y}) \approx C \prod_{i=1}^{N} P_{y_i}(y_i)$, where C is a constant. Thus, the probability of $\mathbf{z}(t), t = 1, ..., T$ for T observations given $\mathbf{A}$ can be approximated as follows:

$$P(\mathbf{z}(t)|\mathbf{A}) = P(\mathbf{y}) \approx C \prod_{i=1}^{N} P_{y_i}(y_i) = C \prod_{i=1}^{N} p_{y_i}(\mathbf{a}_i^T\mathbf{z}(t)). \tag{5}$$

Maximizing the sparseness of dot product y_i is sufficient to provide an approximation of the basis function. Now, we can examine the dot product y_i instead of the basis coefficient s_i. Clearly, better approximations of the prior probability of dot product would allow the model to capture more accurate structures in images. The accuracy of the prior probability is more important especially in overcomplete cases [10]. Based on the two-layer network, we consider an approximation of the prior probability of dot product, which is derived in the Bubble model. A approximation of the prior probability in the temporal bubble model is derived as:

$$\tilde{P}(\mathbf{y}) = \prod_{i} \exp\big(G(b_i(t))\big). \tag{6}$$

where $b_i(t)$ is the output of bubble detector given by Eq. (2). The function $G(\xi)$ has a similar role as the log-density of components in basic ICA, and it should be convex for non-negative variable ξ to enforce sparseness of bubbles. Many heuristically chosen functions can be used, such as the form: $G(\xi) = -\alpha\sqrt{\xi + \varepsilon} + \beta$, where α is the scaling constant and β is the normalization constant. The bubble

pooling given by $\tilde{h}(i,j,\tau)$ is considered fixed, and only the first-layer connections $\mathbf{A}$ are estimated, so this likelihood is a function of the $\mathbf{a}_i$ only.

2.1.2 Prior Probability of Mixing Matrix

Obviously, an overcomplete representation means that the number of basis functions are large. In other words, the basis vectors are randomly distributed into a high-dimensional space. In high-dimensional space, there is a useful property called quasi-orthogonality [8]. This property is previously presented by Hecht-Nielsen [4]: there exists a much larger number of almost orthogonal than orthogonal directions in a high-dimensional space. Therefore, in high-dimensional space even vectors having random directions might be sufficiently close to be orthogonal. The probability for the dot product between two randomly drawn basis vectors: $P(\mathbf{a}_i^T\mathbf{a}_j)$ can be obtained in terms of this quasi-orthogonality. Then the prior probability of mixing matrix $\mathbf{A}$ can be conducted as follows

$$p(\mathbf{A}) = c_m \prod_{i<j} \left(1-(\mathbf{a}_i^T\mathbf{a}_j)^2\right)^{\frac{m-3}{2}}, \tag{7}$$

where c_m is a constant. The detailed derivation of Eq. (7) can be obtained in [8].

2.1.3 Posterior Probability of Mixing Matrix

According to the Eq. (3), (5), (6) and (7), we obtain the log-probability of posterior $\mathcal{L} = \log P(\mathbf{A}|\mathbf{z}(t))$ for T observations $\mathbf{z}(t), t = 1, ..., T$ as follows

$$\mathcal{L} \propto \sum_{t=1}^{T}\sum_{i=1}^{N} G(b_i(t)) + \alpha T \sum_{i<j} \log(1-(\mathbf{a}_i^T\mathbf{a}_j)^2) + C, \tag{8}$$

where α is a constant that is related not only to c_m, but also to the approximations we have made.

2.2 Learning Rule

Our goal is to minimization of objective function $\mathcal{L}$ with respect to $\mathbf{A}$. In practice, it is not necessary to compute the value of the bubble detector for all values of t. For simplicity, only one output of the bubble detector for each sampled spatiotemporal patch $\mathbf{z}(t)$ from the image sequence is computed. Then a simple version of $b_i(t)$ is derived as follows:

$$b_{ik} = \sum_{j=1}^{N} h(i,j) \sum_{t=1}^{T} \phi(T/2-t)(\mathbf{a}_i^T\mathbf{z}_k(t))^2 \tag{9}$$

Note that since the sampling of a spatiotemporal patch automatically introduces limits for temporal integration, the temporal smoothing kernel is defined to $\phi = 1$ in this case.

The learning algorithm of mixing matrix can be derived by maximizing the log-posterior of Eq. (8) and using gradient ascent method

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{a}_r} \propto & \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbf{z}_k(t)\left(\mathbf{a}_r^T\mathbf{z}_k(t)\right) \sum_{i=1}^{N} h(i,j) g(b_{ik}) \\ & + \alpha T \sum_{i<j} \frac{-2\mathbf{a}_i^T\mathbf{a}_j}{1-(\mathbf{a}_i^T\mathbf{a}_j)^2}\mathbf{d}_r,\end{aligned} \tag{10}$$

where the function g is the derivative of function G. $\mathbf{d}_r$ is the r-th column vector of matrix $\mathbf{D} = [0, ..., \mathbf{a}_j, ..., \mathbf{a}_i, ...0]$, where $\mathbf{a}_j$ is the i-th column vector and $\mathbf{a}_i$ is the j-th column vector. Note that we estimate the overcomplete basis functions under a generative model, however, it is unnecessary for an additional step to make the filter $\mathbf{w}_i$ to be orthogonal. After each iteration during learning process on the basis functions, only the norm of the basis functions $\mathbf{a}_r$ need to be set to unity.

3 Simulations

We test the overcomplete spatiotemporal bubble model on natural image sequences. The data are obtained from a video of forest, which is available on http://bcmi.sjtu.edu.cn/~malibo/data/. The training set consists of 16×16 pixels at 20 consecutive time points. Principle component analysis is used to reduce the dimensions of the input data from 256 to 100. Hyvärinen's bubble model is applied to learn complete basis functions. In this case, a set of 100 basis functions are arranged on a 10×10 2-D torus grid (i.e. opposite side are connected to each other) to avoid the border effects. The learned complete basis functions with neighborhoods size of 3×3 are shown in Fig. 1.

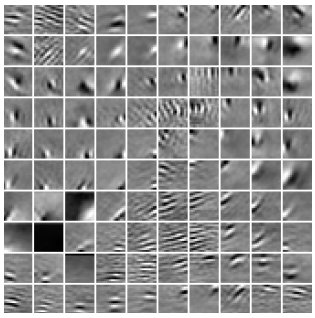


Figure 1. Basis functions learned by ordinary bubble model from natural image sequences in complete case.

The effects of overcomplete cases are investigated in depth. For $2\times$ overcomplete case, a set of 200 basis functions are arranged on a 10×20 2-D torus grid. Whereas

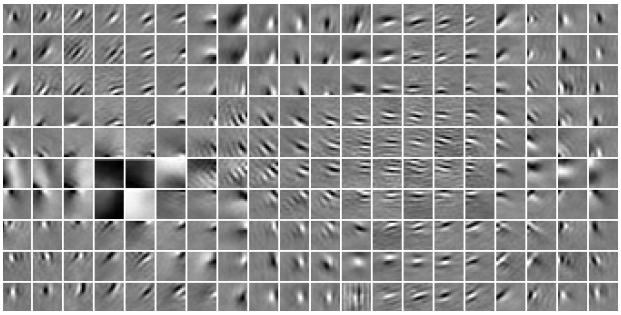


Figure 2. Basis functions learned by our model from natural image sequences in 2× overcomplete case.

for 4× overcomplete case, a set of 400 basis functions are arranged on a 20 × 20 2-D torus grid. The basis functions are initialized to random values and are updated as Eq. (10). After each updating of basis functions, the norm of the basis functions $\mathbf{a}_r$ need to be set to unity.

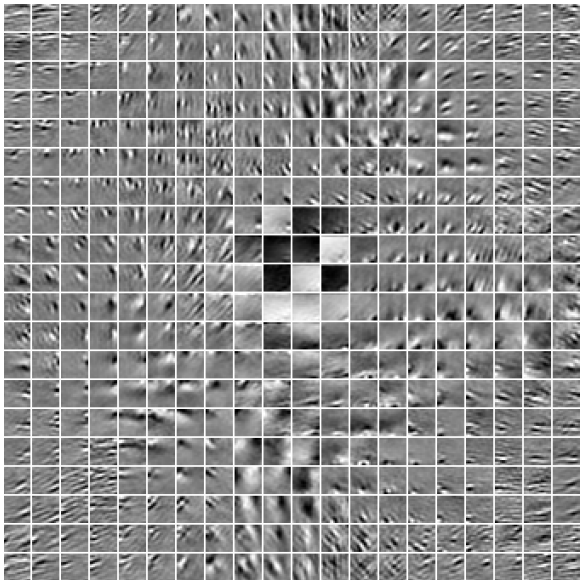


Figure 3. Basis functions learned by our model from natural image sequences in 4× overcomplete case.

The learned basis functions of 2× and 4× overcomplete case with neighborhoods size of 3 × 3 are shown in Fig. 2 and Fig. 3. We can see that overcomplete basis functions are well learned and they are quite similar to the complete ones obtained by Hyvärinen's bubble model. They demonstrate a clear topographic organization for location, orientation and frequency. These three parameters of two nearby basis vectors are similar and mostly change smoothly in the topographic map.

To analyze the tiling properties of the estimated basis vectors, we fitted each basis vector with a Gabor function by minimizing the squared error between the estimated basis vectors and the model Gabor. Figure 4 shows distributions of parameters obtained by fitting Gabor functions in complete, 2× overcomplete and 4× overcomplete case. We can see that the distribution of the centers is quite uniform inside the sampling window. Orientations and spatial frequencies are quite independent from each other. With the increasing of the level of overcompleteness, the scattering points in the plot of location, orientation and spatial frequency become denser and more uniform. And the distribution of phase is much closer to be uniform.

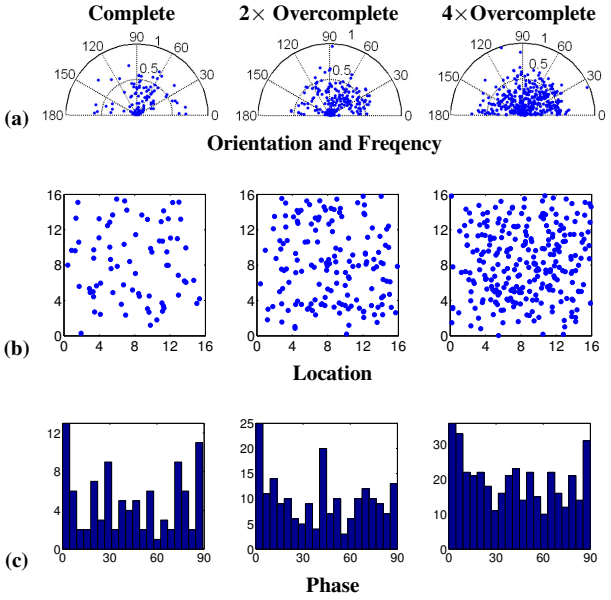
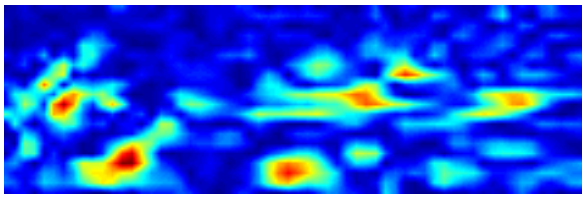


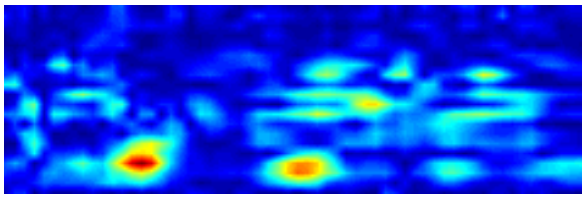
Figure 4. Distributions of parameters derived by fitting Gabor function with some overcomplete levels. The leftmost column (A-C) is a complete case, the middle column is 2× overcomplete case and the rightmost column is 4× overcomplete case. (a) Center location within a patch. (b) Joint distribution of orientation and spatial frequency (plotted in the upper-half plane) (c) Histograms of phase of Gabor fitted (mapped to range $0\,^\circ$˜ $90\,^\circ$)

We also illustrate the responses of complex cells (the outputs of bubble detectors) for a short image sequences (60 frames) in Fig. 5. We use the learned basis functions in 2× overcomplete case, in which a set of 200 basis functions are arranged on a 10× 20 2-D torus grid. The responses of ninth and tenth row of complex cells for a short period are shown.

We can see that the image sequences indeed produce spatiotemporal bubble activities. Surely, the clusters of activity are both spatially and temporally contiguous. And complex cells show more intense sparse topographic activations.



(a) The responses of tenth row of complex cells



(b) The responses of ninth row of complex cells

Figure 5. Bubble activities of complex cells. We use the learned basis functions in 2× overcomplete case, in which a set of 200 basis functions are arranged on a 10×20 2-D torus grid. The responses of two nearby row of complex cells are shown. The horizontal axis is the time, the vertical axis is the cell number. Blue color is negative value. Red color is positive value.

4 Summary and Discussion

We have proposed a Bayesian method for learning overcomplete bubbles from natural image sequences. This is based on two useful properties: (1) quasi-orthogonality of basis vectors in a high-dimensional space; (2) the dot product between basis function and whitened data vector is certain to have sparse marginal distributions. Simulation results suggest that overcomplete bubble coding can be achieved by a Bayesian inference. An important concern in the our model is how we can generalize the ordinary bubble model into an overcomplete case by a Bayesian inference. The bubble coding is an extension of basic ICA and they have the same assumption of existence of the inverse matrix of mixing matrix ($\mathbf{W} = \mathbf{A}^{-1}$). Only complete representations for natural signals can be produced.

Another issue presented in this paper is the relevance of the learned basis functions to neurobiological interpretation. The examination of overcomplete basis functions for natural image sequences suggest that some of the similarity of the properties of complex cells in V1 can be derived by efficient coding principle. A spatiotemporal bubble is more similar to the activity of a complex cell with a space-time-separable receptive field [3]. This model may offer new insights into other aspects of the response properties of neurons at a higher level of cortical processing. In addition, overcomplete bubbles model is also promising in a wide range of fields, such as signal processing and pattern recognition.

Acknowledgments

The work was supported by the National Basic Research Program of China (Grant No. 2005CB724301) and the National High-Tech Research Program of China (Grant No.252006AA01Z125).

References

[1] F. Attneave. Some informational aspects of visual perception. *Psychol Rev*, 61(3):183–93, 1954.

[2] H. B. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961.

[3] R. Emerson, J. Bergen, and E. Adelson. *Directionally Selective Complex Cells and the Computation of Motion Energy in Cat Visual Cortex*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[4] R. Hecht-Nielsen. Context vectors: general purpose approximate meaning representations self-organized from raw data. *Computational Intelligence: Imitating Life*, pages 43–56, 1994.

[5] J. Hurri and A. Hyvarinen. Simple-Cell-Like Receptive Fields Maximize Temporal Coherence in Natural Video, 2003.

[6] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic Independent Component Analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[7] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, 20(7):1237–1252, 2003.

[8] A. Hyvärinen and M. Inki. Estimating Overcomplete Independent Component Bases for Image Windows. *Journal of Mathematical Imaging and Vision*, 17(2):139–152, 2002.

[9] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations. *IEEE SIGNAL PROCESSING LETTERS*, 6(4):87, 1999.

[10] M. Lewicki and B. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587–1601, 1999.

[11] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[12] E. Simoncelli, W. Freeman, E. Adelson, and D. Heeger. Shiftable multiscale transforms. *Information Theory, IEEE Transactions on*, 38(2):587–607, 1992.