IEICE TRANS. INF. & SYST., VOL.E94–D, NO.7 JULY 2011

LETTER

Complex Cell Descriptor Learning for Robust Object Recognition

Zhe WANG[†a)], **Yaping HUANG**[†], *Nonmembers*, **Siwei LUO**[†], *Member*, *and* **Liang WANG**[†], *Nonmember*

SUMMARY An unsupervised algorithm is proposed for learning overcomplete topographic representations of nature image. Our method is based on Independent Component Analysis (ICA) model due to its superiority on feature extraction, and overcomes the weakness of traditional method in fast overcomplete learning. Besides, the learnt topographic representation, resembling receptive fields of complex cells, can be used as descriptors to extract invariant features. Recognition experiments on Caltech-101 dataset confirm that these complex cell descriptors are not only efficient in feature extraction but achieve comparable performances to traditional descriptors.

key words: *Independent Component Analysis, overcomplete, complex cell, invariant feature, object recognition, Caltech-101*

1. Introduction

Neurophysiological studies have shown that there are many topology structures existed in simple cells and complex cells of primary visual cortex (V1) [2]. Receptive fields of simple cells can be regarded as Gabor-like filters. Receptive fields of complex cells are probably organized by receptive fields of simple cells and the input of complex cells is the output of simple cells. Complex cells have good properties, such as phase invariance and some shift invariance. Recently, how to model and utilize the information process of complex cells has been the subject of intense study. Hyvärinen et al. developed the independent Subspace analysis (ISA) and the topographic independent component analysis (TICA) to model the properties of complex cell in V1 [3]. However, ISA and TICA, based on the classic ICA, are complete models that only learn a limited number of filters, thus, they are restricted in the application of feature extraction. Moreover, Osindero et al. proposed a generalized model by extending ICA and TICA to the overcomplete case [4]. However, their models are in general extremely difficult to learn, e.g. using Markov chain Monte Carlo sampling. The training is unacceptably slow, since it often requires several days or weeks.

Because of the weakness of ICA, recent state-of-art methods often use sparse coding with its overcomplete basis to extract rich representations of objects [5], [8]. However, the procedure of feature extractions in sparse coding requires running some sort of iterative algorithms that are always computationally expensive. To avoid such heavy computation, Kavukcuoglu et al. proposed Invariant Predictive Sparse Decomposition (IPSD) to extend sparse coding by adding a feed-forward prediction function to approximate the optimal representations and make inference efficient. Besides, they introduced the nonlinear operation used in ISA and TICA and obtained complex cell-like basis to extract invariant features [1]. However, due to its incorporating many extra parameters in sparse coding, it leads to an increased computing burden in training, and multiple factors need to be adjusted.

In this paper, we propose a new method for learning the overcomplete filter maps from nature image. We will show that these complex cell-like filters can be obtained by a simple iterative algorithm, and will give better discriminative power than traditional descriptors.

2. Learning Overcomplete Topographic Representation from Nature Image

ICA is a generative model for low-level features of many types of natural data [8]. The classical version of the model can be expressed as

$$\mathbf{s}_i = \mathbf{w}_i^T \mathbf{x} \tag{1}$$

where $\mathbf{w}_i \in \mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$. In a neuroscientific interpretation, the variables s_i can model the responses of simple cells from nature image $\mathbf{x}$ and filters $\mathbf{w}_i$ are their receptive fields. The problem of estimating matrix W in Eq. (1) can be resolved by maximizing the sparseness of coefficient s_i, e.g. $\arg\max E\{s^4\}$. However, four order cumulant, being the measurement of sparseness, has the bad property that it is susceptible to noise. In practice other functions or high order cumulant may have to be used.

Here, we introduce the definition of "pairwise cumulant" extended from high order cumulant for modeling the binary relations among these complex cells. The form of pairwise cumulant is defined as:

$$E\left\{\left(g(s_i) + g(s_j)\right)^2\right\} \tag{2}$$

where the nonlinearity $g(.)$ is strictly convex, even (rectifying), and differentiable. The nonlinearity $g(.)$ measures the strength of coefficients, and a good choice is

$$g(s) = \ln\cosh(s) \tag{3}$$

which is a more robust nonlinearity unlike the adverse statistical properties of four order cumulant. Then the form (2)

Manuscript received January 26, 2011.
Manuscript revised March 22, 2011.

[†]The authors are with School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044 China.
a) E-mail: wangzhe908@gmail.com

DOI: 10.1587/transinf.E94.D.1502

gives $E\{(\ln\cosh s_i)^2 + (\ln\cosh s_j)^2 + 2\ln\cosh s_i \ln\cosh s_j\}$. The expectations of the first two terms measure sparseness just as four order cumulant, and the expectation of the last term measures nonlinear correlations.

In addition, a topography function is defined for representing spatial adjacency structures in complex cells. It is based on topography of spatial organization of the cells, and is given the form as

$$h(i,j) = \begin{cases} 1 & if\ \|d(i) - d(j)\| \le r\ \ and\ \ i \ne j \\ 0 & otherwise \end{cases} \tag{4}$$

where $d(i)$ and $d(j)$ denote the location of s_i and s_j in the topography, and r is a constant represented the width of the neighborhood. In this paper, the function won't be learnt, and is fixed for simplicity.

For estimating the filters W in our work, we maximize the pairwise cumulants among these components in topography, and obtain the following objective function

$$\arg\max_{\mathbf{w}_i} \sum_{k,k\ne i} h(i,k) E\{(g(\mathbf{w}_i^T\mathbf{x}) + g(\mathbf{w}_k^T\mathbf{x}))^2\} \quad Subject\ to\ \|\mathbf{w}\| = 1 \tag{5}$$

Then some simple and frequently-used method, e.g. stochastic gradient descent, can be employed to maximize the object function. Besides, for making filters $\mathbf{w}_i$ under the constraint of unit norm and linear independence, an orthogonalization procedure must be performed after every iteration. In classic ICA, the orthogonalization procedures require the filter set $\mathbf{W}$ is a square matrix, and are only adequate for the complete learning. In overcomplete case — the number of filters $\mathbf{w}_i$ exceeds their self-dimensionality, we combine the quasi-orthogonal estimation used in [8], [9].

The quasi-orthogonal estimation assumes that there is much more place for vectors in high dimension data spaces. It is possible to have more than n vectors that are practically orthogonal in the n-dimensional space. When n is larger, the number of quasi-orthogonal vectors grow and the angles between these vectors are as close as 90 degrees. So the following quasi-orthogonalization procedure will be performed to maximize the angles between vectors $\mathbf{w}_i$ after every iteration:

$$\begin{aligned} &1.\ Let\ \mathbf{W} \leftarrow \frac{3}{2}\mathbf{W} - \frac{1}{2}\mathbf{W}\mathbf{W}^T\mathbf{W} \\ &2.\ Normalize\ each\ \mathbf{w}\ to\ unit\ norm \end{aligned} \tag{6}$$

Note that it is approximate estimations for filters $\mathbf{W}$ because of the assumption of quasi-orthogonality. We will show that these overcomplete filters can give a good performance on recognition.

3. Complex Cell Feature Extraction

Once parameter W are learnt, these filters will emerge topographic representation similar to the properties of complex cells, and constitute unsupervised complex cell (UCC) descriptors that are robust to minor variations of input data. In accordance with information process of complex cell, the invariant features can be extracted by our UCC descriptors with the three steps:

(1) Computing the liner coefficients $\mathbf{s}$ of input by Eq. (1).

(2) Rectifying these coefficients s_i by the sigmoid function and absolute value function.

(3) Computing a max or an average of the coefficients s_i from the same descriptor to generate the final representations $\mathbf{u}$.

Thus, the complex cell feature obtained by a UCC descriptor Π_i with max or average operations can be described as

$$u_i = \max_{\mathbf{w}_j \in \Pi_i} (|sigmoid(\mathbf{w}_j^T\mathbf{x})|) \tag{7}$$

or

$$u_i = \underset{\mathbf{w}_j \in \Pi_i}{avg} (|sigmoid(\mathbf{w}_j^T\mathbf{x})|) \tag{8}$$

Note that liner features can be obtained by the inner products (Eq. (1)) because of the characteristic of ICA, and no iterative algorithms need to be performed. So UCC descriptors is more simple and efficient than traditional methods, e.g. sparse coding based descriptors.

4. Experiments

Our experiments include three parts. First, we use the proposed algorithm to learn overcomplete the topographic representation from nature image. Second, we study the invariance of these learnt representation to object transformations. Finally, Caltech-101 database is used to test their discriminative power on object recognition.

4.1 Overcomplete Topographic Representation

The 50000 image patches of size 16×16 pixels sampled from 13 nature images[†] are used as training data for our algorithm. In the preprocess, these patches are removed mean and whitened by Principal Component Analysis (PCA), and the dimension of the data vectors is down to 196. In the results shown below, the inverse of these preprocessing steps is performed. Besides, the topography function $h(i,k)$ labels the neighbors of component s_i to 1 in the 5×5 neighborhood.

We train 512 filters, which is overcomplete, by the stochastic gradient descent with 600 iterations. Learning is quite rapid — less than half an hour to reach convergence. Figure 1 shows the learnt filters organized a topographic map. First, they are localized, oriented, and bandpass filters resembling Gabor-like filters. In addition, it demonstrates a clear topography of filters with local continuity of orientation, frequency, and location, whereas the phases seem to be random. These representations conform to the properties of complex cell spatial receptive fields in V1.

[†]Available on www.cis.hut.fi/projects/ica/data/images/

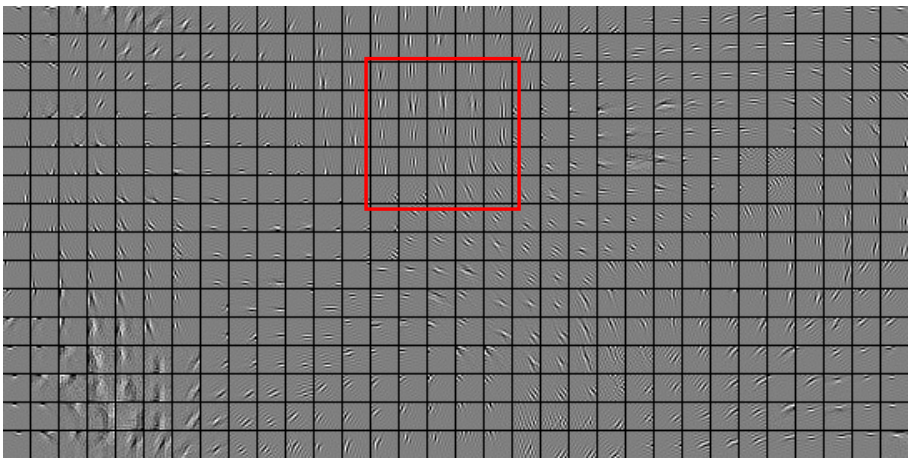

Fig. 1 A learnt 32×16 filter map. These filters located in the red box group a UCC descriptor Π_i.

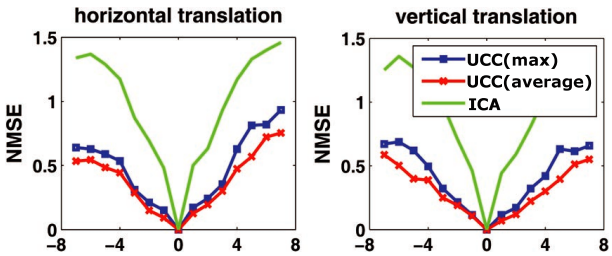

Fig. 2 NMSE results between the features of generated stimulus and its translation. The operations used in UCC descriptors are specified in parentheses.

4.2 UCC Descriptor and Its Invariance

We set 5×5 subregion which is identical to the neighborhood size used in learning, and choose 128 subregions overlapped by three filters both horizontally and vertically from the learnt filters map. These filters, located in a same subregion, have similar properties and belong to a same UCC descriptor as shown in Fig. 1. So we obtain 128 UCC descriptors which can produce 128-dimensional features similar to Scale-Invariant Feature Transform (SIFT) [7] and IPSD descriptors.

Then, Gabor functions are used as stimulus to test the invariance of UCC descriptors under horizontal and vertical translations. There are 2925 Gabor stimulus randomly generated in the center location, and then normalized mean squared errors (NMSE) are compared between the features of generated stimulus and features of its translation, averaged over all of these generated stimulus. Figure 2 shows the NMSE results obtained by the 128 learnt UCC descriptors. For comparison, we employ 128 ICA filters as descriptors and their testing results are also given. It can be seen that the features produced by UCC descriptors with max or average operations are robust to minor variations of input data, whereas, liner features produced by ICA filters are not invariant — a small change in the input results in a large change in the representations.

4.3 Object Recognition

In this experiment, the performance of UCC descriptors on the recognition task is tested using the well known Caltech-101 database. The Caltech-101 database contains 102 classes (101 categories of objects as well as a background class) with high shape variations. We follow the common experiment setup for Caltech-101, training on 30 images per category and testing on 30 images per category randomly. Each image is converted to gray-scale and is preprocessed by removing the image mean and normalizing the pixel values so that their standard deviation is equal to 1.

By employing the 128 UCC descriptors obtained from the 32×16 topographic map with corresponding the feature extraction method described in Sect. 3, the detailed feature extraction procedure is as follows:
(1) Extracting complex cell features on 16×16 image patches spaced by 4 pixels over each image to produce 128 feature maps of size 32×32.
(2) A locally normalized step is performed by using a 5×5 Gaussian window and a 5×5 boxcar filter to smooth the feature maps.
(3) The dimensionality of the representations is further reduced to 3060 components by PCA.
(4) Finally, the 3060-dimensionality representations are fed to a linear SVM classifier with the "one against all" technique.

This architecture is similar to that of IPSD and SIFT used in [1], as shown in Fig. 3. Table 1 gives the average recognition results of different descriptors. It turns out that UCC descriptors achieve powerful performance on recognition, which confirms UCC descriptors give invariant features as well as good intrinsic representations of objects. Besides, UCC descriptors, which belong to unsupervised learned descriptors, embody the statistic characteristics of objects and outperform these hand-designed descriptors, e.g. SIFT and C2 features. Finally, the UCC descriptors with average operation achieve the better accuracy in

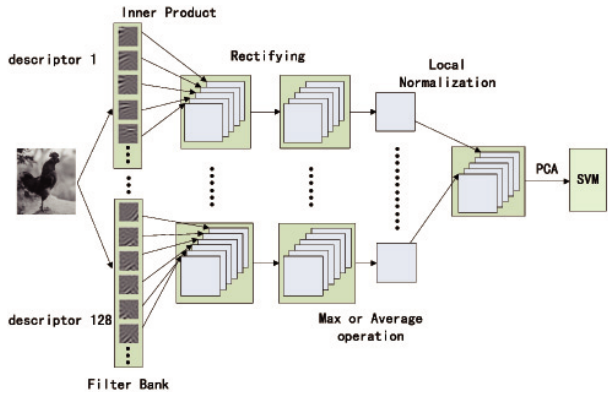

Fig. 3 The architecture of feature extraction for object recognition.

Table 1 Recognition rate comparison on Caltech-101.

Descriptor	Average Rate (%)
UCC (max/average)	51.4/**53.1**
IPSD [1]	50.9
SIFT [1], [7] (without orientation invariance)	51.2
SIFT [1], [7] (with orientation invariance)	45.7
C2 features [6]	47.1

this task. The excellent performance of average operation is also approved by Jarrett et al. [10].

5. Conclusion

A simple overcomplete algorithm is developed for learning topographic representations from nature images. Experiments confirm that these learnt filters resemble receptive fields of complex cells, and bring a measure of invariance against the change of input. Though fast and simple procedures both in training and inference, our approach gives the internal and invariance representations of objects. Classification results demonstrate state-of-the-art performances using SVM classifiers, and also show the superiority of complex cells modeling for solving recognition task.

Further work to improve the performance of recognition is shown as the following aspects: introducing spatial pyramid techniques, building a multi-stage architecture of feature extraction, and developing a supervised learning algorithm.

Acknowledgments

This work is supported by National Nature Science Foundation of China (60975078, 60902058, 60805041, 60872082, 60773016), Beijing Natural Science Foundation (4092033) and Doctoral Foundations of Ministry of Education of China (200800041049).

References

[1] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," Proc. Computer Vision and Pattern Recognition Conference, 2009.

[2] J.M. Alonso and L.M. Martinez, "Functional connectivity between simple cells and complex cells in cat striate cortex," Nature Neuroscience, vol.1, no.5, pp.395–403, 1998.

[3] A. Hyvärinen and P.O. Hoyer, "A Two-layer sparse coding model learn simple and complex cell receptive fields and topography from natural images," Vision Research, vol.41, no.18, pp.2413–2423, 2002.

[4] S. Osindero, M. Welling, and G.E. Hinton, "Topographic product models applied to natural scene statistics," Neural Comput., vol.18, pp.381–344, 2006.

[5] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," Vision Research, vol.37, pp.3311–3325, 1997.

[6] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," Proc. Computer Vision and Pattern Recognition, 2005.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, Wiley Interscience, 2001.

[9] Z. Wang, S. Luo, and L. Wang, "A fast algorithm for learning the overcomplete image prior," IEICE Trans. Inf. & Syst., vol.E93-D, no.2, pp.403–406, Feb. 2010.

[10] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," Proc. International Conference on Computer Vision, IEEE, 2009.