Spatiotemporal Feature Extraction Based on Invariance Representation

Wenlu Yang and Liqing Zhang

***Abstract*— This paper investigates spatiotemporal feature extraction from temporal image sequences based on invariance representation. Invariance representation is one of important functions of the visual cortex. We propose a novel hierarchical model based on invariance and independent component analysis for spatiotemporal feature extraction. Training the model from patches sampled from natural scenes, we can obtain image basis with properties of translational, scaling, and rotational features. Further experiments on TV videos and facial image sequences show different characteristics of spatiotemporal features are achieved by training the proposed model. All these computer simulations verify that our proposed model is successful for spatiotemporal feature extraction.**

I. INTRODUCTION

WE can recognize an object regardless of its transformation, such as translation, rotation or scaling. Such capability of our recognizing transformation-invariant objects is one of important functions in the brain. Many recent studies in the fields of neuroscience, neurophysiology and psychology show that such a transformation invariant preprocessing could be a necessary step towards achieving transformation-invariant classification or detection in a hierarchical system. In this paper, we focus on extracting spatiotemporal features from image sequences and videos, and propose a hierarchical model that simulates the mechanism in the visual pathway.

On the other hand, due to evolution from nature in the long term, this mechanism of neural representation has an important correlation with statistical properties of natural scenes. Following the way, Barlow[1] found that the role of early sensory neurons in the visual pathway is to reduce statistical redundancy in the sensory inputs, suggesting that Redundancy Reduction is an important processing principle in the neural system. Based on this concept, Gabor-like features resembling the receptive fields of V1 cells have been derived either by imposing sparse over-complete representations[2] or statistical independence as in the framework of independent component analysis[3].

However, these studies have not taken transformation invariance into account. In other words, what are the responses of simple cells and complex cells when one transformation is applied to an object within the receptive fields of these cells? Some researchers have addressed this problem. Van Hateren[4] obtained spatiotemporal receptive fields of complex cells and Hyvarinen and Hoyer[5] modelled the receptive fields of complex cells. Grimes and Rao[6] proposed a bilinear generative model to study the translation-invariance. However, there are few models in the literatures of physiologically and neurophysiologically extracting simultaneously multi-temporal features from image sequences or videos. To investigate this problem, we combine the visual perception invariance mechanism and Independent Component Analysis(ICA) to learn spatiotemporal features, and these spatiotemporal features can be used to construct a model for transformation-invariant perception.

The rest of the paper is organized as follows. Section II introduces our model and learning algorithm for learning spatiotemporal features. In section III, we will demonstrate computer simulation results to show the basic characteristics in the trained model. In the final section IV, we provide some discussions and conclusions.

II. THE PROPOSED MODEL AND LEARNING ALGORITHM

The visual information received by the visual system is very complex and the resources of the optic nerve are limited. How does the visual system compromise between them? Barlow[1] found that the role of early sensory neurons is to remove statistical redundancy in the sensory inputs, provided a criterion called Redundancy Reduction. Olshausen and Field[2] presented the Sparse Coding method based on that only minor neurons respond to the stimulus from the natural environment, whereas the major neurons weakly respond. To verify it, they provided experimental results that natural images are able to be reconstructed by linearly combining the basis functions and the corresponding coefficients, which are considered as the receptive fields and responses of simple cells, respectively. The coefficients are described as the supergaussian probability distribution. An alternative method is (ICA)[3][4][5] that imposes the mutual independent constraint on the responses of neurons, and the similar results are obtained.

Using ICA model, we propose a hierarchical model combining the ICA and invariance representation. In the following section we will introduce the model and the corresponding learning algorithm.

A. The model for spatiotemporal feature extraction

In this section, we propose a model for spatiotemporal feature extraction, as shown in figure 1. The model is a four-layer network which includes the input layer L_1, the sparse representation layer L_2, the integrated layer L_3, and the final invariance representation layer L_4. In the first layer L_1, each

Wenlu Yang and Liqing Zhang are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China (email:wenluyang@online.sh.cn, zhang-lq@cs.sjtu.edu.cn). And Wenlu Yang is with the Department of Electronic Engineering, 1550 Pudong Rd. Shanghai Maritime University, Shanghai 200135, China

978-1-4244-1821-3/08/$25.00©2008 IEEE

neuron in the retina receives gray value of one pixel as its activity. L_2 is a layer for sparse representation, and its main function is to represent the input images with features and the corresponding independent components. The layer L_3 is the integrated layer at which neurons average all activities of neurons connected from layer L_2. And the final layer L_4 is a senior representation layer at which each neuron receives responses of connected neurons in the layer L_3 to extract the transformation invariance. In other words, the L_4 layer possesses the capability of representing transformation invariance.

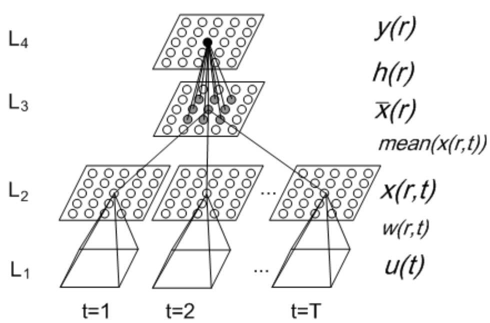

Fig. 1. Model for spatiotemporal features. L_1 is the input layer, $u(t)$ denotes a stimuli at time t. L_2 is a layer for sparse representation, and $x(r,t)$ denotes the response of neuron r at time t, $w(r,t)$ denotes the connect weights with L_1. The layer L_3 is the integrated layer, and their activities are denoted by $\overline{x}(r)$. And the final layer L_4 is a senior representation layer, here, *h(r)* denotes neighborhood connect, and *y(r)* denotes responses.

B. The learning algorithm of the model

To derive the learning algorithm of the model, we introduce briefly the standard ICA algorithm. For the standard ICA model $u(t) = \mathbf{A}(t)x(t) = \mathbf{W}^{-1}(t)x(t)$, Cichocki et al.[7] used the Kullback-Leibler divergence between the distribution $p(\mathbf{x}(t);\mathbf{W}(t))$ of obtained by the actual value $\mathbf{W}(t)$ and the reference distribution $q(\mathbf{x}(t))$ to give the cost function as

$$\begin{aligned} R(\mathbf{x}(t),\mathbf{W}(t)) &= -\frac{1}{2}\log|det(\mathbf{W}(t)\mathbf{W}(t)^T)| \\ &\quad -\sum_{r=1}^{n} E\log q(r)(x(r,t)). \end{aligned} \tag{1}$$

Applying the Natural Gradient rule to the cost function, the learning algorithm of $\mathbf{W}(t)$(the corresponding basis functions $\mathbf{A}(t)=\mathbf{W}^{-1}(t)$) can be described[8] as

$$\begin{aligned} \triangle\mathbf{W}(t) &= -\eta(t)\frac{\partial R}{\partial \mathbf{W}(t)}\mathbf{W}^T(t)\mathbf{W}(t) \\ &= \eta(k)[\mathbf{I}-\varphi[\mathbf{x}(t)]\mathbf{x}^T(t)]\mathbf{W}(t), \end{aligned} \tag{2}$$

where, $\varphi_r(x(r)) = -\frac{q_r^{'}(x(r))}{q_r(x(r))}$, $q(x(r))$ is the prior probability distribution over the coefficients $x(r)$ are highly peaked at zero with heavy tails as compared to a Gaussian distribution of the same variance (i.e., the Laplace probability distribution function).

The standard ICA estimation methods constrain the independent components to be uncorrelated. These components have the properties of higher-order correlation, which can be interpreted biologically as simultaneous activation of neurons at the same time when neurons receive a stimulus. Thus, we can use this mechanism to analyze the higher-order correlation of neural responses.

Suppose $\overline{x}(r_1)$ and $\overline{x}(r_2)$ are responses of two neurons in the layer L_3, if $\overline{x}(r_1)$ and $\overline{x}(r_2)$ are topographical neighborhood, the covariance between $\overline{x}(r_1)$ and $\overline{x}(r_2)$ satisfies

$$\begin{aligned} cov(\overline{x}^2(r_1),\overline{x}^2(r_2)) &= E\{\overline{x}^2(r_1),\overline{x}^2(r_2)\} \\ &\quad -E\{\overline{x}^2(r_1)\}\{\overline{x}^2(r_2)\} \\ &\neq 0. \end{aligned} \tag{3}$$

Due to the higher-order correlation, the response of each complex cell in the layer L_4 is described as $|y(r_1)| = (\sum_{r_2=1}^{n} h(r_1,r_2)\overline{x}^2(r_2))^{1/2}$, $h(r_1,r_2)$ is the connect weights between a complex cell r_1 and a simple cell r_2 in its vicinity. That is, the receptive field of the complex cell consists of that of its neighborhood simple cells and is bigger than that of simple cells. Its probability distribution function is described as equation (4).

$$\begin{aligned} q(y(r_1)) &= \frac{1}{\sqrt{2}\sigma}\exp\left(-\frac{\sqrt{2}|y(r_1)|}{\sigma}\right) \\ &= \frac{1}{\sqrt{2}\sigma}\exp\left(-\frac{\sqrt{2}}{\sigma}\sqrt{\sum_{r_2=1}^{n} h(r_1,r_2)\overline{x}(r_2)^2}\right), \end{aligned} \tag{4}$$

where, σ^2 is the variance of responses. Therefore, we obtain

$$\begin{aligned} \varphi_{r_1}(y(r_1)) &= -\frac{q_{r_1}^{'}(y(r_1))}{q_{r_1}(y(r_1))} \\ &= \left(\sum_{r_2=1}^{n}\frac{\sqrt{2}h(r_1,r_2)\overline{x}(r_2)^2}{\sigma}\sqrt{\sum_{r_2=1}^{n} h(r_1,r_2)\overline{x}(r_2)^2}\right). \end{aligned} \tag{5}$$

As vectors in the matrix $\varphi(\mathbf{y})$, all $\varphi(y(r_1))$s are rewritten as

$$\varphi(\mathbf{y}) = [\varphi(y(1)),\varphi(y(2)),\cdots,\varphi(y(n))]. \tag{6}$$

Combining equations (6) and (2), the learning algorithm of topographically self-organized receptive fields of simple cells is described as

$$\triangle\mathbf{W}(r,t) = \eta(k)[\mathbf{I}-\varphi(\mathbf{y}(r))\mathbf{x}(r,t)^T]\mathbf{W}(r,t). \tag{7}$$

According to the equation (7), we are able to learn $\mathbf{W}(r,t)$ with topographical characteristics of receptive fields of simple cells. The steps of learning spatiotemporal features is given as the following pseudocode.

1) **Input:** The training data **U**=$\{u(i,j,t)\}$ is centered and whitened using PCA, the whitened data is denoted by $\mathbf{Z}(l,j,t) \in \mathbf{R}^{L\times N\times T}$ and the whitening matrix denoted by $\mathbf{V}$, here, L is the dimensionality of data, N is number of samples, and T is length of sequences.

2) **Initialization:** Randomly initialize the weights $\mathbf{W}_Z(M,L,t)$, where, M denotes the number of neurons in each group in the layer L_3 in figure 1, L and t is same as the above.
3) **Iteration:**
   a) for each t (=1,2,$\cdots$,T)
   b) Update $\mathbf{X}(m,j,t)$=$\mathbf{W}_Z(m,l,t)$*$\mathbf{Z}(l,j,t)$;
   c) end
   d) Calculate the mean of $\mathbf{X}$ for all t, denoted by $\overline{\mathbf{X}}(m,j)$;
   e) Replace each $\mathbf{X}(m,j,t)$ with $\overline{\mathbf{X}}(m,j)$ for all t;
   f) for each t
   g) Update the $\mathbf{W}_Z(t)$ according to the eqution (7);
   h) Normalize $\mathbf{W}_Z(m,l,t)$ to the unity.
   i) end
   j) Calculate the $\Delta\mathbf{W}_Z(t)$;
   k) until $\Sigma\Delta\mathbf{W}_Z(t) < \varepsilon$.
4) **Output:** Calculate $\mathbf{W}(t) = \mathbf{W}_Z(t)\mathbf{Z}$, the inverse of $\mathbf{W}(t)$ is denoted by $\mathbf{A}(t)$

III. Simulation results

In this section, we will show three experiments to verify our proposed model and the learning algorithm. The first one is about learned spatiotemporal features from natural image sequences, the second from TV videos, and the third from multi-view faces.

A. Spatiotemporal features from natural image sequences

Training Set: Randomly select patches of size 12×12 from natural images, and then translate them by {2,4,6,8,10,12} pixels along horizontal and vertical direction respectively, rotate them by angles {0,15,30,45,60,75} in degree. Vectorize them into column vectors as samples denoted by $u(i,j,t) \in \mathbf{R}^{I\times N\times T}$. Here, I(=144) is the dimensionality of samples, T (=6×6×6=216) transformation sequences, and j the number of samples. The training data **U**={$u(i,j,t)$} consists of all these samples.

We use the data to train our proposed model and the resulting features $\mathbf{A}(t)(=\mathbf{W}^{-1}(t))$ are shown in figure 2. For the simplicity of analyzing feature regularity, we show a spatiotemporal feature in figure 3. From the figure, we can easily find that there are 216 features including six horizontal translations, six vertical translations, and six rotations. All the features can be considered as one spatiotemporal feature of a complex cell which is able to detect transformations of translation and rotation.

By the way, our extra experiments show that, taking no account of rotating angle, features repeat in the range of {90, 105, $\cdots$, 345} in degree.

From the figures 2 and 3, the features are localized, oriented, and bandpass, which resemble the receptive fields of simple cells. In figure 3, all the 216 features can be considered as a spatiotemporal feature which has the properties of horizontal translation, vertical translation, and rotation. These characteristics are same as that of training data. In other words, when stimuli like the training data are presented

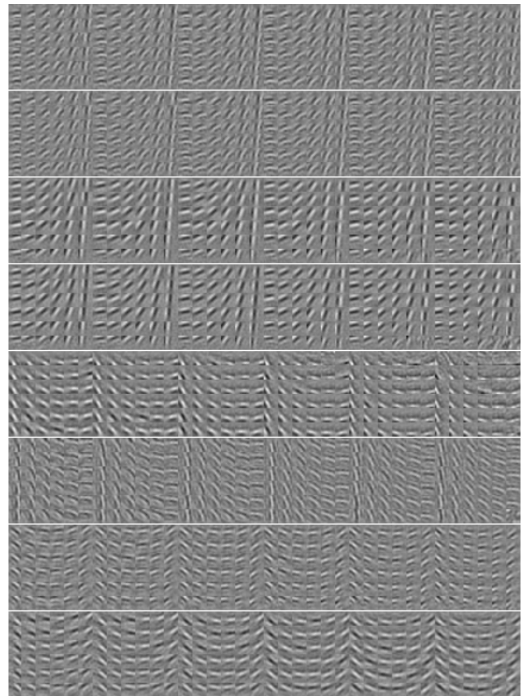

Fig. 2. Subsets of spatiotemporal features.

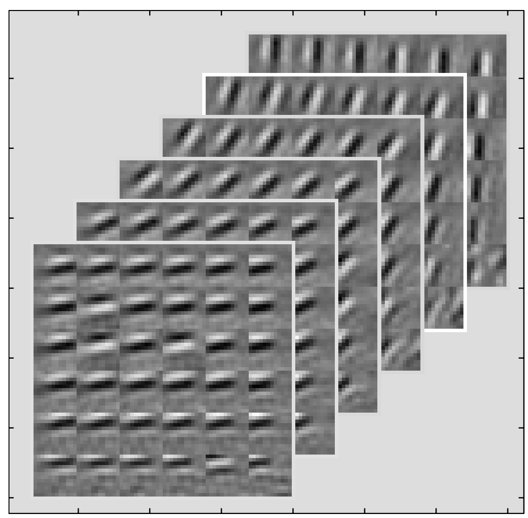

Fig. 3. An example of spatiotemporal features. The spatiotemporal feature is composed of 216 features, including six intervals of translation ($\Delta x = 2, \Delta y = 2$) pixels in a slice and six rotational angles in the range of {0,15,30,45,60,75} in degree. Each row in a slice represents one translation in horizontal direction and each column in vertical direction. Between two neighborhood slices, features are rotating by angle interval of fifteen in degree.

within the receptive field of neuron, the neuron will be firing all the time. In the neurophysiological term, the neuron has an ability of invariance-transformation representation, which is the fundamentals of cognitive ability.

B. Structure of Spatiotemporal features

The learned spatiotemporal features can be considered as 3-D data set, denoted by $A(X,Y,T)$, where, X and Y are the 2 spatial dimensions and T is the temporal dimension. Integrating this 3-D data set along the Y-axis yields a simplified spatiotemporal profile, or X-T plot[9]. Figure 4 shows the contour of integrating 36 features with a same rotating angle.

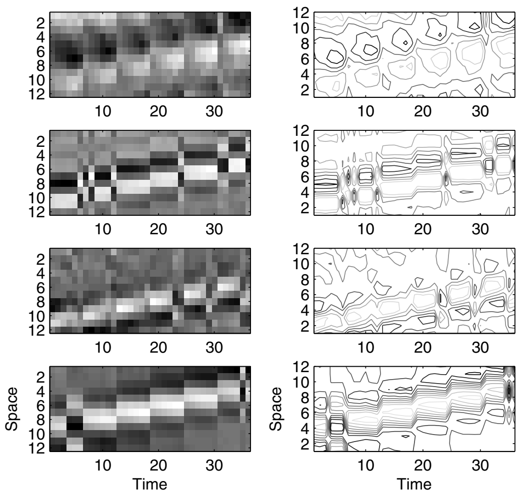

Fig. 4. Spatiotemporal structure. Integrating the 3-D data set of spatiotemporal features along the Y-axis yields a simplified spatiotemporal profile, or X-T plot.

This space-time profile resemble the spatiotemporal receptive-field profiles for simple cells[9]. Because of integrating 3-D data along Y-axis, horizontal translation is eliminated between time 1-12, 13-24, and so forth, whereas the vertical translation is very obvious with time.

C. Spatiotemporal features from TV video sequences

Training Set: Each video sequence consists of a stack of 10 time frames of 12×12 image patches. Vectorize image patches into column vectors as samples denoted by $u(i,j,t) \in \mathbf{R}^{I\times N\times T}$. Here, I(=144) denotes the dimensionality of samples, T (=10) denotes the length of video sequences, and j denotes the number of samples. All these samples form the training data $\mathbf{U}$=$\{u(i,j,t)\}$.

The video data comes from the situation video 'Six Friends', in which there are almost indoor shows. Sampling frequency is 25 frames per second. We resample one out of five frames from the videos. Then we randomly select successive video sequences of 12×12×10 as a spatiotemporal sample. The total 20000 spatiotemporal samples are used to train the model. The resulting spatiotemporal features are shown in figure 5.

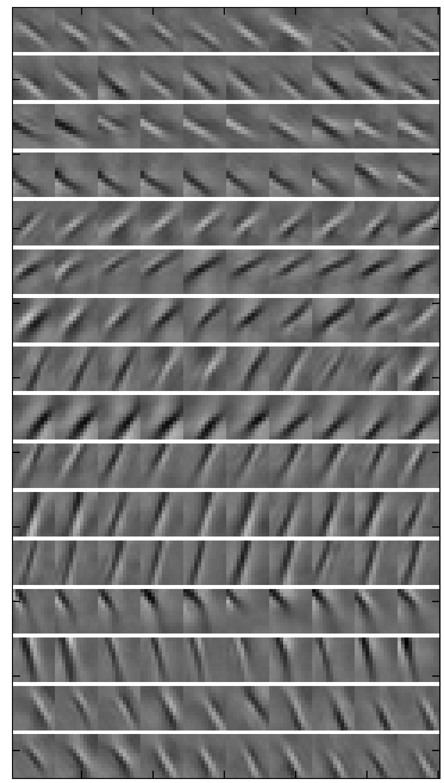

Fig. 5. Subsets of spatiotemporal features for Videos.

Spatiotemporal features: The figure 5 tells us five points: (1) the features resemble the receptive fields of simple cells in the primary visual cortex; (2) each row has ten features that can be considered as a spatiotemporal features; (3) between features, there are slightly horizontal or vertical movement; (4) the neighborhood two features in a spatiotemporal feature have scale changes; (5) there are little rotation transformation.

All the characteristics consist with that of videos. For (1), applying the spatiotemporal ICA to the video sequences yields edge-detectors similar to the receptive fields of simple cells. For (2), ten successive frames have strong correlation and continuous movements. For (3) and (4), when capturing two successive frames, the vidicon only move slightly its view in the horizontal, vertical, and relative direction, or objects shift a little in the scene in the opposite direction. For (5), because there are little rotating view in this TV video, there are not obvious rotational features.

Further more experiments show that different sampling rates will result in different results. If a sampling rate is smaller, the two neighborhood features in a spatiotemporal feature is less correlation. The higher the sampling rate is, the much similar the features. If sampling is appropriate, the characteristics mentioned above are obtained. In summary,

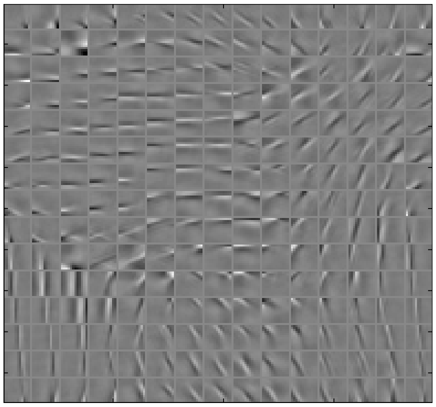

Fig. 6. Topographical maps of features.

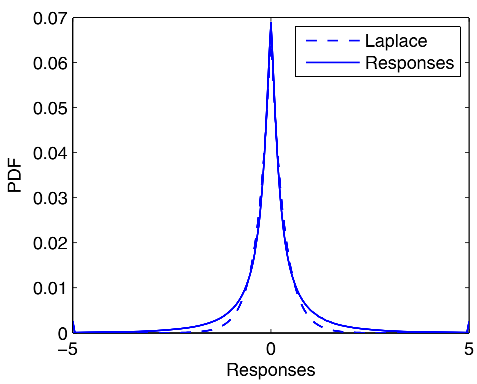

Fig. 7. PDF of responses and fitted Laplace function.

spatiotemporal features must be learned in so far as training sequences have spatiotemporal properties of transformation.

Topographical map: To examine the topographical map of features, we show subset of all features at the first column in figure 5. As shown in figure 6, most of features are similar to that in their vicinity. The orientations of features are gradually changing in the neighborhood. The characteristics resemble results found in the neurophysiological experiments.

Analysis of sparseness of responses: A neuron fires sparsely through its receptive field, a feature obtained by ICA. That is consistent with the results of neurophysiological experiments: when neurons are stimulated, a few neurons are active and most of them are inactive. In other words, the probability distribution of neuronal activity will be highly peaked around zero with heavy tails.

To analyze the sparseness of responses, we fit the probability distribution with Laplace function. The fitted results are shown in figure 7, where, the solid line denotes the probability distribution of responses and the dashed line does the that of fitted Laplace function. From this figure, the responses are sparse and can be described as the supergaussian probability distribution.

D. spatiotemporal facial features

Further to verify the learning algorithm for extracting spatiotemporal feature, we have made an experiment on facial images with view changes.

Training Set: The training set is selected from the pose subset, images of 1040 subjects across 7 different poses, included in the CASE-PEAL-R1 face database[10]. The training data are generated as follows: for any pose t facial image, detect and crop the face, resize it to size of 36×30 pixels. Then these seven cropped faces are reshaped to a column vector as a sample $u(i, j, t)$, size of 1080×1×7. Here, $u(i, j, t) \in \mathbf{R}^{1080\times1040\times7}(t = 1, 2, \cdots, 7)$. The training data **U**=$\{u(i, j, t)\}$ consists of all these samples.

Fig. 8. Subsets of spatiotemporal features for facial squences.

The learned features is given in figure 8. For faces of an individual, there are only different poses without more such as lighting and expression, and so the learned spatiotemporal features only possess the characteristics of pose changes.

IV. Discussions and Conclusions

By imposing a constraint of invariance representation, our proposed model extends ICA model [4][5][11] to generate spatiotemporal features. Experiments on natural image sequences, TV video, and facial images with view changes demonstrate that this model has some characteristics: (1) extracting spatiotemporal features, (2) overcomplete features, (3) theoretically limitless length of sequences, and (4) easier implementation.

Furthermore, compared with the bilinear generative model[6] proposed by Grimes and Rao, our model has some advantages such as easy realization and less computing cost, and more rich properties for transformation-invariant features. They only explored the model for learning translation-invariant basis functions.

Compared to the previous methods [4][5][11] in which a sample sequence , for example, size of $12 \times 12 \times 10$, is vectorized into one column as a sample, size of 1440×1, we divide it into 10 subsets, size of 144×1. In the proposed model, we add an integrated layer L_3 with a constraint of invariance representation to learn spatiotemporal features. Our method has some advantages such as: (1) sequences are not to be limited anymore, any length of sequences of data can be used to learned spatiotemporal features and then decide right length; (2) data preprocessing becomes easier, if sequences are long enough, the previous methods are difficult in dimensionality reduction and de-correlation of training data to promote the convergence of ICA algorithm.

Compared to the recent multi-linear methods such as MICA[12] and Tensor Factorization (TF)[13], our method is better for performing a great data set and learning more multi-way, whereas MICA and TF is limited to the computing resources for large number of multi-way and samples. Meanwhile, our method can learn, with less computing resources, more spatiotemporal features than that obtained from MICA and TF.

From the viewpoint of models, we compare our model with Neocognitron proposed by Fukushima [14]. The Neocognitron was actually a relative classifier which recognized patterns from given testing data. The model cannot perceive object motions or transformations. However, these are important functions in the visual pathways of the brain. Different from Neocognitron, the goal of our model focuses on extracting spatiotemporal features which are used to construct more complicated networks for perceiving input patterns and object motions.

Our further work will focus on applying learned spatiotemporal features to visual perception of transformations such as translation, rotation, view change, and scaling, and object recognition.

ACKNOWLEDGMENTS

The work was supported by the National High-Tech Research Program of China (Grant No.2006AA01Z125) and the National Basic Research Program of China (Grant No. 2005CB724301).

REFERENCES

[1] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," *Sensory Communication*, pp. 217–234, 1961.

[2] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[3] A. J. Bell and T. J. Sejnowski, "The independent component of natural scenes are dege filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[4] J. H. V. Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proceedings: Biological Sciences*, vol. 265, no. 1412, pp. 2315–2320, 1998.

[5] P. O. Hoyer and A. Hyvarinen, "A multi-layer sparse coding network learns contour coding from natural image," *Vision Research*, vol. 42, no. 12, pp. 1593–1605, 2002.

[6] D. B. Grimes and R. P. N. RAO, "Bilinear sparse coding for invariant vision," *Neural computation*, vol. 17, no. 11, pp. 47–73, 2005.

[7] A. Cichocki and L. Q. Zhang, "Two-stage blind deconvolution using state-space models," *In proceedings of the Fifth International Conference on Neural Information Processing*, pp. 729–732, 1998.

[8] L. Q. Zhang, A. Cichocki, and S. Amari, "Self-adaptive blind source separation based on activation function adaptation," *IEEE Transactions on Neural Networks*, vol. 15, no. 2, pp. 233–244, 2004.

[9] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. i. general characteristics and postnatal development," *Journal of Neurophysiology*, vol. 69, no. 4, pp. 1091–1117, 1993.

[10] W. Gao, B. Cao, S. G. Shan, D. L. Zhou, X. H. Zhang, and D. B. Zhao, "The cas-peal large-scale chinese face database and evaluation protocols," *Technical Report No. JDL_TR_04_FR_001, Joint Research & Development Laboratory, CAS*, 2004.

[11] W. L. Yang and L. Q. Zhang, "Perception of transformation - invariance in the visual pathway," *Lecture Notes In Computer Science*, vol. 4666, pp. 657–664, 2007.

[12] M. Alex, O. Vasilescu, and D. Terzopoulos, "Multilinear independent components analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 547–553, 2005.

[13] M. Mørup, L. K. Hansen, J. Parnas, and S. M. Arnfred, "Decomposing the time-frequency representation of EEG using non-negative matrix and multi-way factorization," *Technical reports*, 2006.

[14] K. Fukushima, "Visual pattern recognition with neural networks," *Lecture Notes in Computer Science*, vol. 654, pp. 16–31, 1992.