Available online at www.sciencedirect.com

SCIENCE DIRECT

NEUROCOMPUTING

Neurocomputing 69 (2006) 878–883

www.elsevier.com/locate/neucom

Letters

Nonnegative independent component analysis based on minimizing mutual information technique☆

Chun-Hou Zheng[a,b], De-Shuang Huang[a,*], Zhan-Li Sun[a,b], Michael R. Lyu[c], Tat-Ming Lok[d]

[a] *Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China*
[b] *Department of Automation, University of Science and Technology of China, China*
[c] *Computer Science & Engineering Department, The Chinese University of Hong Kong, Shatin, Hong Kong*
[d] *Information Engineering Department, The Chinese University of Hong Kong, Shatin, Hong Kong*

Received 15 May 2005; received in revised form 20 June 2005; accepted 20 June 2005
Available online 10 October 2005
Communicated by R.W. Newcomb

Abstract

A novel neural network technique for nonnegative independent component analysis is proposed in this letter. Compared with other algorithms, this method can work efficiently even when the source signals are not well grounded. Moreover, this method is insensitive to the particular underlying distribution of the source data. Experimental results demonstrate the advantages of our approach in achieving satisfactory results regardless of whether the source data are well grounded or not.

Keywords: Independent component analysis (ICA); Nonnegative approach; Mutual information; Neural network

1. Introduction

The task of nonnegative independent component analysis (ICA) is to estimate the source vector $\mathbf{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$ and mixing matrix $\mathbf{A}$ in the linear generative model $\mathbf{x} = \mathbf{As}$ given an observation vector $\mathbf{x} = (x_1, \ldots, x_{\mathrm{n}})^{\mathrm{T}}$, where the sources are nonnegative, i.e., $\Pr(s_i < 0) = 0$, and independent, i.e., $p(s_i s_j) = p(s_i)p(s_j)$ if $i \neq j$, here $\Pr(\bullet)$ is probability function and $p(\bullet)$ is probabilistic density function (pdf). Several authors have introduced some algorithms for nonnegative ICA [7–10], yet these algorithms are all based on the assumption that the sources s_i are *well grounded* except for independence and nonnegativity. We call a source s_i *well grounded* if $\Pr(s_i < \varepsilon) > 0$ for any $\varepsilon > 0$, i.e., s_i has nonzero probabilistic density function all the way down to zero [8]. However, in practical applications, many real-world nonnegative sources are not well grounded, e.g., images. In this letter, we propose a new algorithm for nonnegative ICA even in the case that the sources are not well grounded. Essentially, our algorithm is one based on minimizing mutual information.

2. Algorithm for Nonnegative ICA

2.1. Algorithm architecture

Lemma 1. *Let* $\mathbf{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$ *be an n-dimensional random vector of real-valued independent sources which have non-Gaussian distributions*, $\mathbf{A}$ *and* $\mathbf{B}$ *be nonsingular* $n \times n$ *real matrices*, $\mathbf{x} = \mathbf{As}$ *be a linear mixing model of* $\mathbf{s}$, *and* $\mathbf{y} = \mathbf{Bx} = \mathbf{BAs} = \mathbf{Rs}$ *be a linear unmixing model of* $\mathbf{x}$. *Then the mutual information* $I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{y})$ *is minimized if and only if* $\mathbf{R} = \mathbf{\Lambda P}$, *where*

$$H(y_i) = -\int p(y_i) \log p(y_i)\, \mathrm{d}y_i \tag{1}$$

denotes Shannon's differential entropy [3,6]. $H(\mathbf{y})$ *is the differential entropy of a multidimensional random variable* $\mathbf{y}$, $\mathbf{\Lambda}$ and $\mathbf{P}$ *are diagonal and permutation matrices*, *respectively*.

☆This work was supported by the National Science Foundation of China (Nos. 60472111 and 0405002), and partially supported by Hong Kong Research Grants Council (Nos. CUHK4205/04E).

*Corresponding author.

E-mail address: dshuang@iim.ac.cn (D.-S. Huang).

0925-2312/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.06.008

Proof. In the case where $\mathbf{R} = \mathbf{\Lambda P}$, $\mathbf{y}$ is simply a permutation of the independent source vector $\mathbf{s}$ with just sign and scale ambiguity, and the mutual information $I(\mathbf{y})$ is zero. The proof of the converse is relatively complicated, but interested readers can refer to the literature [3] or chapter 10 in the literature [6].

According to the theory given above, the unmixing system of nonnegative ICA can be constructed as shown in Fig. 1, where $\mathbf{B}$ is the unmixing matrix in ICA, y_i are the extracted independent components, and $\psi_i(\boldsymbol{\varphi}_i, y_i)$ are some nonlinear mappings, $\boldsymbol{\varphi}_i$ is the parameters contained in $\psi_i(\boldsymbol{\varphi}_i, y_i)$.

Assume that each function $\psi_i(\boldsymbol{\varphi}_i, y_i)$ is the cumulative probability function (CPF) of the corresponding component y_i, i.e.

$$z_i = \psi_i(\boldsymbol{\varphi}_i, y_i) = \int p(y_i)\,\mathrm{d}y_i \tag{2}$$

then

$$p(z_i) = \frac{p(y_i)}{|\partial z_i/\partial y_i|} = \frac{p(y_i)}{p(y_i)} = 1. \tag{3}$$

That is to say, z_i are uniformly distributed in [0, 1], Consequently, $H(z_i) = 0$ [2]. Moreover, because $\psi_i(\boldsymbol{\varphi}_i, y_i)$ are all continuous and monotonic increasing transformations (thus also invertible), then it can be easily shown that $I(\mathbf{z}) = I(\mathbf{y})$ [1]. Consequently, we can obtain

$$I(\mathbf{y}) = I(\mathbf{z}) = \sum_i H(z_i) - H(\mathbf{z}) = -H(\mathbf{z}). \tag{4}$$

Therefore, maximizing $H(\mathbf{z})$ is equivalent to minimizing $I(\mathbf{y})$.

It has been proved in the literature [1] that, given the constraints placed on $\psi_i(\boldsymbol{\varphi}_i, y_i)$, then z_i is bounded to [0, 1], and given that $\psi_i(\boldsymbol{\varphi}_i, y_i)$ is also constrained to be a continuous increasing function, then maximizing $H(\mathbf{z})$ will lead $\psi_i(\boldsymbol{\varphi}_i, y_i)$ to become the estimates of the CPFs of y_i.

Based on the above analysis, we can minimize $I(\mathbf{y})$ by maximizing $H(\mathbf{z})$ with appropriate constraints placed on $\psi_i(\boldsymbol{\varphi}_i, y_i)$ (see Section 2.2). As a result, $\mathbf{R} = \mathbf{BA} = \mathbf{\Lambda P}$ (according to Lemma 1). Without loss of generality, we shall assume $\mathbf{R}$ to be a diagonal matrix, i.e.

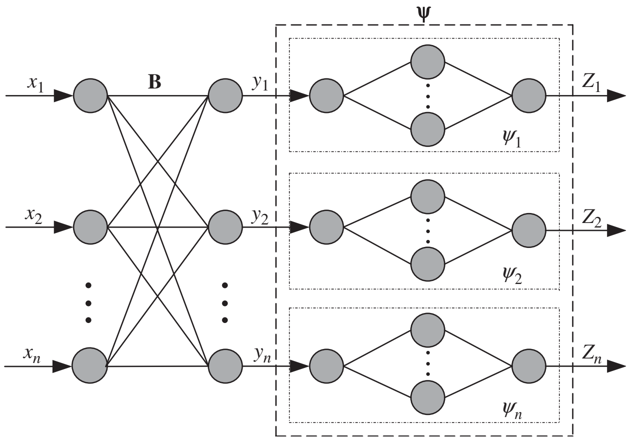


Fig. 1. Structure of nonnegative ICA unmixing system proposed in this letter.

$$\begin{aligned} r_{ij} &\neq 0 \quad \text{if } i = j, \\ r_{ij} &= 0 \quad \text{if } i \neq j, \end{aligned} \tag{5}$$

then we have $y_i = r_{ii}s_i$. So y_i is a duplicate of s_i with just sign and scale ambiguity. Moreover, by considering the sources s_i to be nonnegative in this letter, we will see that y_i is either nonnegative or nonpositive, corresponding respectively to a positive r_{ii} or a negative one. Consequently, we can eliminate the sign ambiguity by taking absolute value of y_i, i.e., $|y_i|$, as the recovered signals.

Now, the fundamental problem that we have to solve is to optimize the network by maximizing $H(\mathbf{z})$.

2.2. Learning algorithm

With respect to the separation structure of our interest, the joint probabilistic density function of $\mathbf{z}$ can be calculated as [2]:

$$p(\mathbf{z}) = \frac{p(\mathbf{x})}{|\det(\mathbf{B})|\prod_{i=1}^{n} |\psi_i'(\boldsymbol{\varphi}_i, y_i)|} \tag{6}$$

where $\psi_i'(\boldsymbol{\varphi}_i, y_i)$ is the derivative of $\psi_i(\boldsymbol{\varphi}_i, y_i)$ with respect to y_i. From Eq. (6), we can immediately obtain the following expression:

$$H(\mathbf{z}) = H(\mathbf{x}) + \log|\det(\mathbf{B})| + \sum_{i=1}^{n} E(\log|\psi_i'(\boldsymbol{\varphi}_i, y_i)|). \tag{7}$$

Minimizing $I(\mathbf{y})$, which is equivalent to maximizing $H(\mathbf{z})$ here, requires the computation of its gradient with respect to the separation structure parameters $\mathbf{B}$ and $\boldsymbol{\varphi}$.

Since $H(\mathbf{x})$ does not depend on $\mathbf{B}$ and $\boldsymbol{\varphi}$, we thus have the following gradient expressions:

$$\frac{\partial H(\mathbf{z})}{\partial \mathbf{B}} = \frac{\partial \log|\det(\mathbf{B})|}{\partial \mathbf{B}} + \frac{\partial\left(\sum_{i=1}^{n} E(\log|\psi_i'(\boldsymbol{\varphi}_i, y_i)|)\right)}{\partial \mathbf{B}}, \tag{8}$$

$$\frac{\partial H(\mathbf{z})}{\partial \boldsymbol{\varphi}_k} = E\left(\frac{\partial \log|\psi_k'(\boldsymbol{\varphi}_k, y_k)|}{\partial \boldsymbol{\varphi}_k}\right). \tag{9}$$

Of course, the above computation depends on the structure of the parametric nonlinear mapping function $\psi_i(\boldsymbol{\varphi}_i, y_i)$.

In this letter, we use multilayer perceptrons (MLP) [4] with a single hidden layer to model the nonlinear parametric functions $\psi_k(\boldsymbol{\varphi}_k, y_k)$, thus they can be written as

$$\psi_k(\boldsymbol{\varphi}_k, y_k) = \psi_k(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\mu}_k, y_k) = \sum_{j=1}^{M_k} \alpha_j^k \tau(\beta_j^k y_k - \mu_j^k), \tag{10}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the weight matrices of the input layer and the output layer, respectively, $\boldsymbol{\mu}$ is the hidden unit's bias term, and $\tau(\bullet)$ the activation function of the hidden layer.

From Eqs. (8)–(10), we can easily calculate the gradients of $H(\mathbf{z})$ with respect to each parameter, and then optimize the network accordingly.

The explanation we give for our method improving efficiency of the algorithm is that the momentum and adaptive step sizes with error control have been employed. In addition, in this letter, there are four neurons used in the hidden layer of $\psi_i(\boldsymbol{\varphi}_i, y_i)$ blocks. Finally, to implement the constraints on $\psi_i(\boldsymbol{\varphi}_i, y_i)$, which are increasing functions with values in a finite interval, the arctangent sigmoids of the hidden units of $\psi_i(\boldsymbol{\varphi}_i, y_i)$ blocks were chosen as increasing functions, and the vector of weights leading from the hidden units to the output units was normalized at the end of each epoch. All the weights in each $\psi_i(\boldsymbol{\varphi}_i, y_i)$ block were initialized to positive values, resulting in an increasing function. Furthermore, the interval of z_i was chosen as $[-1, 1]$ instead of $[0, 1]$, which still maintains the fact that the maximum of $H(\mathbf{z})$ corresponds to the minimum of $I(\mathbf{y})$. On the other hand, it allows the use of bipolar sigmoids in hidden units. Consequently, faster training results [1].

3. Experimental results and discussions

3.1. Simulating data

In this experiment, we generated three artificial nonnegative signals as the original signals, which can be expressed as

$$\mathbf{S} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} (\sin(600\pi t/10000 + 6\cos(120\pi t/10000)) + 1) + \lambda \\ (\sin(\pi t/10) + 1) + \lambda \\ ((\mathrm{rem}(t, 23) - 11)/9)^5 + 2.8)/2 + \lambda \end{bmatrix}, \tag{11}$$

where λ is a nonnegative constant used to control the well grounded degree of the source signals, the function $\mathrm{rem}(u, v)$ represents the remainder of u divided by v. The third one, s_3, is a supergaussian signal and the other two, s_1 and s_2, are supergaussian signals. Fig. 2 (a) shows the three source signals in the case of $\lambda = 0.2$. Clearly, they are all nonnegative and not well grounded. Yet they approximate well grounded when $\lambda = 0$. In this experiment, three source signals ($\lambda = 0.2$) are mixed by using a 3×3 mixing matrix:

$$A = \begin{bmatrix} 0.4452 & 0.4511 & -0.1234 \\ -0.2061 & 0.8013 & 0.3241 \\ 0.3113 & -0.2314 & 0.7125 \end{bmatrix}. \tag{12}$$

Fig. 2 (b) shows the unmixed three signals, while the correlations between these three recovered signals and the three original signals are reported in Table 1. In addition, the source-to-output matrix $\mathbf{R} = \mathbf{BA}$ was

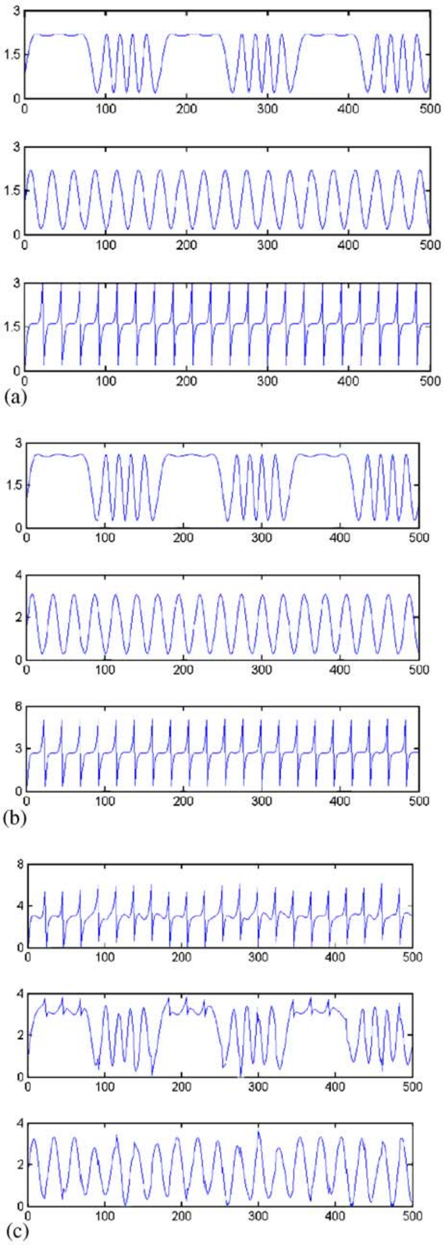


Fig. 2. The three set of signals ($\lambda = 0.2$). (a) Source signals. (b) Recovered signals using the method proposed in this letter. (c) Recovered signals using the method proposed in the literature [10].

$$\mathbf{R} = \begin{bmatrix} -1.1742 & -0.0094 & 0.0013 \\ 0.0021 & 1.3977 & -0.0012 \\ 0.0149 & 0.0014 & -1.7060 \end{bmatrix}. \tag{13}$$

From Fig. 2 (b) and matrix $\mathbf{R}$ we can see that the unmixed signals are all nonnegative and they are very similar to the original signals shown in Fig. 2 (a).

Further, for comparison, we also used another nonnegative ICA algorithm proposed in the literature [10] to conduct the same experiment. Fig. 2 (c) shows the recovered signals. The correlations between the recovered

Table 1
Correlations between the recovered signals and the original signals

		Method in this letter			Method in the literature [10]		
		y_1	y_2	y_3	y_1	y_2	y_3
$\lambda = 0.0$	s_1	1.0000	−0.0590	0.0223	0.9999	−0.0166	0.0044
	s_2	−0.0520	1.0000	0.0020	−0.0425	0.9991	0.0059
	s_3	0.0317	0.0025	1.0000	0.0267	0.0012	0.9996
$\lambda = 0.2$	s_1	1.0000	−0.0575	0.0186	−0.1397	0.9807	0.1367
	s_2	−0.0512	1.0000	0.0056	0.1372	−0.1765	0.9747
	s_3	0.0304	0.0054	0.9999	0.9779	0.1847	−0.0981

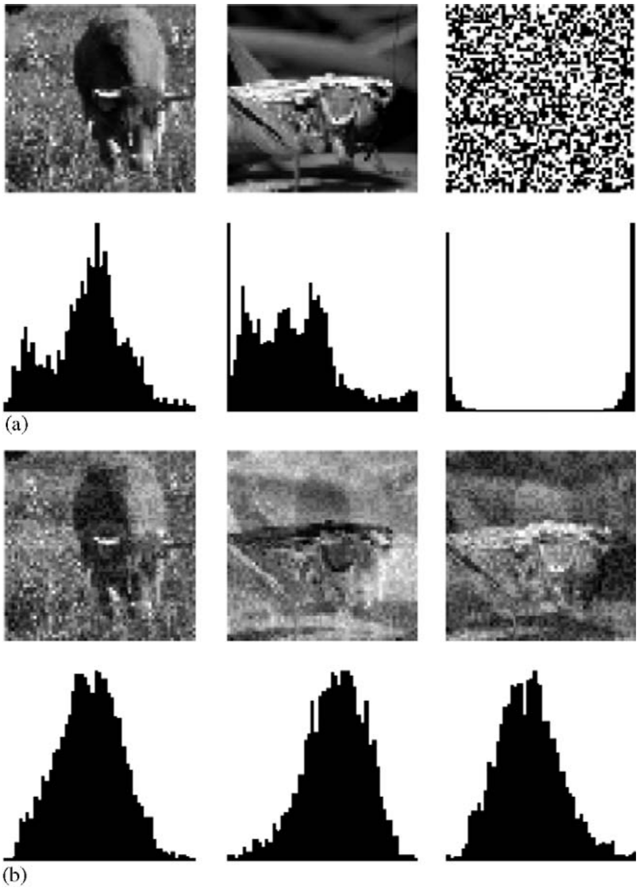


Fig. 3. (a) The original source images and their histograms. (b) The mixed images and their histograms.

signals and the original signals are also reported in Table 1. The results of the experiment when $\lambda = 0$ are shown in Table 1 too. From Fig. 2 and Table 1, it can be seen that if the source signals were not well grounded, our method would become significantly better than the other ones. Furthermore, the experimental results also show that our approach can achieve satisfactory results regardless of whether the source data are well grounded or not.

3.2. Image data

In order to further verify the efficacy of the proposed scheme, we finally applied the algorithm to unmix image data. In this experiment, two image patches of size 240 × 240 were selected from a set of images of natural scenes [5], and down-sampled by a factor of 4 in both directions to yield 60 × 60 images. The third image is an artificial one containing only noisy signals. Each of the images was treated as one source with its pixel values representing 60 × 60 = 3600 samples. Three sources are then mixed using a randomly chosen mixing matrix

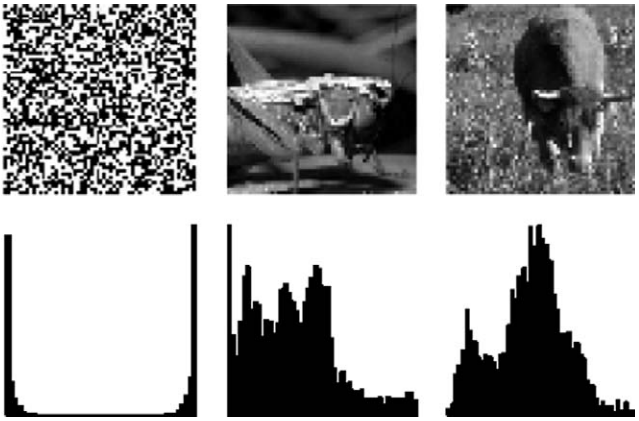


Fig. 4. The recovered images and their histograms.

$$\mathbf{A} = \begin{bmatrix} 0.8412 & 0.2513 & 0.3234 \\ 0.3864 & -0.8015 & 0.2241 \\ -0.3123 & 0.7314 & 0.3234 \end{bmatrix}. \tag{14}$$

Fig. 3 shows the original and mixed images as well as their histograms. No special pre-processing was performed on the mixed image data, other than dividing them by a constant, so these data can be appropriately analyzed with our network (The values of $\mathbf{x}$ are roughly between −2 and 2).

The recovered images and their histograms are shown in Fig. 4, while the correlations between these three recovered images and the three original images are reported in Table 2. In addition, the source-to-output matrix $\mathbf{R} = \mathbf{BA}$ was

$$\mathbf{R} = \begin{bmatrix} 0.0000 & -0.0000 & 1.0108 \\ -0.0335 & -1.0084 & -0.0074 \\ 1.0250 & 0.0300 & 0.0168 \end{bmatrix}. \tag{15}$$

Table 2
Correlations between the recovered images and the original images

	Method in this letter			Method in the literature [10]		
	y_1	y_2	y_3	y_1	y_2	y_3
s_1	−0.0107	−0.0154	0.9994	0.9963	0.0517	−0.0693
s_2	−0.0038	0.9994	−0.0194	0.0214	−0.0099	0.9991
s_3	1.0000	0.0032	0.0056	−0.0620	0.9980	0.0075

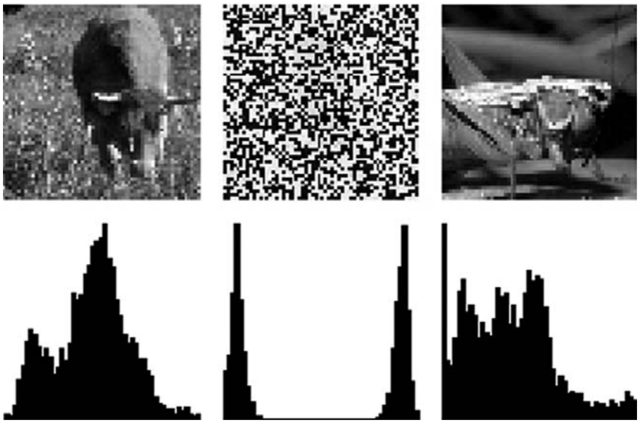


Fig. 5. The recovered images and their histograms using the method proposed in the literature [10].

Clearly, our proposed algorithm is able to separate the images reasonably well.

Further, Fig. 5 shows the recovered images and their histograms using the method proposed in the literature [10]. The correlations between the recovered images and the original sources are also reported in Table 2. From Figs. 4, 5 and Table 2, it can be seen that the separated images using the method proposed in this letter is more similar to the original images than the other ones. In addition, it can be found that the distinction between the two experimental results is not too great. The reason for this phenomenon is that the source images were approximately well grounded.

4. Conclusions

A novel algorithm based on minimizing mutual information for the nonnegative ICA was proposed in this letter. This approach is shown to be effective and feasible even when the source signals are not well grounded. Finally, experimental results supporting the evidence of our approach are provided.

References

[1] L.B. Almeuda, Linear and nonlinear ICA based on mutual information—the MISEP method, Signal Process. 84 (2) (2004) 231–245.

[2] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.

[3] P. Comon, Independent component analysis, a new concept?, Signal Process. 36 (3) (1994) 287–314.

[4] D.S. Huang, Systematic Theory of Neural Networks for Pattern Recognition, Publishing House of Electronic Industry of China, Beijing, September 1996.

[5] A. Hyvärinen, P. Hoyer, Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces, Neural Comput. 12 (7) (2000) 1705–1720.

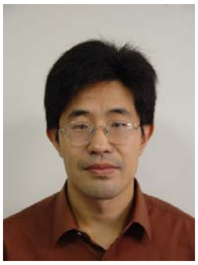
[6] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, March 2001.

[7] E. Oja, M.D. Plumbley, Blind separation of positive sources by globally convergent gradient search, Neural Comput. 16 (9) (2004) 1811–1825.

[8] M.D. Plumbley, Conditions for nonnegative independent component analysis, IEEE Signal Process. Lett. 9 (6) (2002) 177–180.

[9] M.D. Plumbley, Algorithms for nonnegative independent component analysis, IEEE Trans. Neural Networks 14 (3) (2003) 534–543.

[10] M.D. Plumbley, Optimization using Fourier expansion over a geodesic for non-negative ICA, in: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2004, Granada, Spain, September 2004, pp. 49–56.



Chun-Hou Zheng was born in Shandong province, China, in 1973. He received his M.Sc. degree in Control Theory & Control Engineering in 2001, from the Qu Fu Normal University, China. He is now in pursuit for his Ph.D degree in Pattern Recognition & Intelligent System at the University of Science and Technology of China. His research interests include artificial neural networks, intelligent computing, and intelligent information processing.

De-Shuang Huang (SM'98), a Professor and a Ph.D. Advisor in the University of Science and Technology of China (USTC), Hefei, China, and Professor of Graduate School of the Institute of Intelligent Machines, Chinese Academy of Sciences (CAS). From September 2000 to March 2001, he worked as a Research Associate at the Hong Kong Polytechnic University. From April 2002 to June 2003, he worked as a Research Fellow at the City University of Hong Kong. From October to December 2003, he worked as a Research Fellow at the Hong Kong Polytechnic University. From July to December 2004, he worked as the University Fellow in Hong Kong Baptist University. Dr. Huang is currently a senior member of the IEEE.

Zhan-Li Sun obtained Bachelor's degree in Machinery and Electronic Engineer of the Huainan Industrial University in 1997, obtained M.Sc. degree in Machinery and Electronic Engineer of the Hefei University of Technology in 2003. From March 2003 on, in pursuit for Doctor's degree in Pattern Recognition & Intelligent System in University of Science & Technology of China. His research interests include artificial neural networks, blind source separation, signal and image processing.

Michael R. Lyu received his B.Sc. in Electrical Engineering from the National Taiwan University in 1981, his M.Sc. in Computer Science from the University of California, Santa Barbara, in 1985, and his Ph.D. in Computer Science form the University of California, Los Angeles, in 1988. He is currently a Professor at the Computer Science and Engineering Department of the Chinese University of Hong Kong. Dr. Lyu's research interests include software reliability engineering, distributed systems, fault-tolerant computing, web technologies, mobile networks, digital video library, multimedia processing, and video searching and delivery.

Tat M. Lok received his B.Sc degree in Electronic Engineering from the Chinese University of Hong Kong, and his M.Sc and Ph.D degree in Electrical Engineering from the Purdue University. In 1996, he joined the Chinese University of Hong Kong, where he is currently an Associate Professor. His research interests include communication theory, signal processing for communications and CDMA systems.