IEICE TRANS. INF. & SYST., VOL.E88–D, NO.10 OCTOBER 2005

PAPER Special Section on Image Recognition and Understanding

A Nonlinear Principal Component Analysis of Image Data

Ryo SAEGUSA[†a)], **Hitoshi SAKANO**[††b)], ***and*** **Shuji HASHIMOTO**[†c)], ***Members***

SUMMARY Principal Component Analysis (PCA) has been applied in various areas such as pattern recognition and data compression. In some cases, however, PCA does not extract the characteristics of the data-distribution efficiently. In order to overcome this problem, we have proposed a novel method of Nonlinear PCA which preserves the order of the principal components. In this paper, we reduce the dimensionality of image data using the proposed method, and examine its effectiveness in the compression and recognition of images.

key words: *nonlinear PCA, neural network, dimensionality reduction, image*

1. Introduction

In the analysis of multi-dimensional data, it is important to reduce the dimensionality of the data, because it will help to extract new knowledge from the data and to decrease the computational cost. As a method of dimensionality reduction, Principal Component Analysis (PCA) has been applied in various areas such as pattern recognition and data compression [1]–[7].

PCA, especially, has been one of the conventional methods to extract features of image data. Eigenface by Turk extracts features of facial images for recognition [8]. The subspace method extracts a subspace of images for each category and classifies an input image into the category to which subspace has the greatest similarity.

However, it is reported that nonlinear characteristics exist in some data sets of images, such as facial images with emotional expressions and images of an object with variable orientation [9]. When we reduce the dimensionality of such data, a nonlinear method is considered to perform more effectively than a linear method such as PCA.

Recently, some methods of Nonlinear PCA (NLPCA) have been developed [10]–[18].

The method by Gnanadesikan [10] applies PCA to a vector of which components are polynomial terms generated from the components of an input vector. The method has difficulty with the NLPCA of high-dimensional data, because the number of the polynomial terms increases in the polynomial order of the dimensionality of the input vector.

Kernel PCA (KPCA) by Schölkopf et al. [11] is an effective method of NLPCA which applies linear PCA to the nonlinearly mapped image of an input vector. Studies of KPCA are widely developed. However, KPCA poses several problems for practical use in respect to computational costs, dimensionality reduction, and data-reconstruction.

The Sandglass-type Multi-Layered Perceptron (SMLP) by Irie and Kawato [14] and by DeMers and Cotterell [15] can construct adequate nonlinear mapping functions to extract a low-dimensional internal representation from given data. In this method, the dimensionality of the internal representation is fixed.

The training method proposed by Takahashi, et al. [16] enables an SMLP to extract the ordered principal components. The authors of the paper mathematically proved that, in a case where the training method is applied to the three layered linear MLP, the outputs of the units in the bottleneck layer converge to the ordered principal components obtained by PCA.

This hierarchical training is considered to be effective and interesting, while the network architecture of the SMLP is not hierarchical. Therefore, when we add a new principal component, the re-training of the whole networks is required.

We have proposed a novel method of NLPCA that preserves the order of the principal components [17], [18]. In the proposed method, hierarchically arranged neural networks corresponding to the ordered principal components are trained to build a set of nonlinear mapping functions to extract and reconstruct the data.

In the proposed method, the training and the architecture are both hierarchical so that user can add new principal components with only the training of the additional mapping functions (the additional sub-networks). We do not require the re-training of the entire mapping functions.

We have already examined the effectiveness of the proposed method in reconstructing the data and in preserving the distances among the data. In this paper, we apply the proposed method to some data sets of images and discuss the experimental results.

In Sect. 2, we formulate the proposed method. In Sect. 3, we demonstrate the extraction and reconstruction of images, and discuss the effectiveness of the proposed method. In Sect. 4, we discuss the complexity and the redundancy. We also compare the proposed method with other

Manuscript received September 27, 2004.
Manuscript revised February 7, 2005.

[†]The authors are with the Department of Applied Physics, School of Science and Engineering, Waseda University, Tokyo, 169–8555 Japan.
[††]The author is with NTT Data Corporation, Tokyo, 104–0033 Japan.
a) E-mail: ryos@ieee.org
b) E-mail: sakanoh@nttdata.co.jp
c) E-mail: shuji@shalab.phys.waseda.ac.jp

DOI: 10.1093/ietisy/e88–d.10.2242

methods. In Sect. 5, we present our conclusion and areas of feature work.

2. Methods

In this section, we describe a concept on dimensionality reduction of multi-dimensional data, and formulate a proposed method. This section is based on [17], [18].

2.1 Reconstruction Error in Dimensionality Reduction

Dimensionality reduction of multi-dimensional data means to map the data in a high-dimensional space into a low-dimensional space while preserving the characteristics of the distribution.

In this paper, we define the criterion of an optimization using Mean Square Error (MSE) in the original data space. A smaller MSE indicates the higher fidelity of the reconstruction.

Let $\boldsymbol{x} \in R^n$, $\boldsymbol{y} \in R^m$ and $\hat{\boldsymbol{x}} \in R^n$ be the coordinates of an input vector in a high-dimensional space, an extracted vector (principal components) in a low-dimensional space and a reconstructed vector in the original high-dimensional space, respectively, where $n, m \in N$ $(n \geq m)$ indicates the dimensionality of the given data vectors and the number of principal components, respectively. R and N represent a set of real numbers and natural numbers. The MSE is defined as

$$\mathrm{MSE} = E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{y})\|^2], \tag{1}$$

where $E[\cdot]$ represents the expectation and $\|\cdot\|$ represents L_2 norm. In the following, we assume $E[\boldsymbol{x}] = 0$ for simplicity.

2.2 Nonlinear Dimensionality Reduction

Figure 1 shows the concept of the dimensionality reduction from R^2 to R^1 by NLPCA.

In Fig. 1, NLPCA nonlinearly maps the data vector $\boldsymbol{x} \in R^2$ onto $y \in R^1$, and also nonlinearly maps $y \in R^1$ onto $\hat{\boldsymbol{x}} \in R^2$. If the nonlinear mapping functions are monotonous, any $\hat{\boldsymbol{x}}$ is mapped onto a curved line in R^2. Therefore, the MSE of NLPCA corresponds to the average of the squared distances between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ mapped onto the curved line.

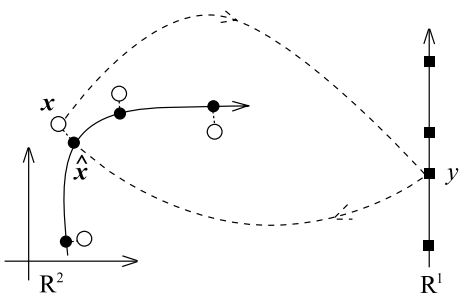

Fig. 1 A concept of the nonlinear dimensionality reduction of two-dimensional vectors onto one-dimensional vectors. In the figure, a white circle, a black square and a black circle represent the input vector $\boldsymbol{x}$, the principal component y of $\boldsymbol{x}$, the reconstructed vector $\hat{\boldsymbol{x}}$ from y, respectively.

In general, when we reduce the dimensionality of data in R^n to obtain their principal components in R^m, the nonlinear mapping function $\boldsymbol{\phi} : R^n \mapsto R^m$ is defined as

$$\boldsymbol{y} = \boldsymbol{\phi}(\boldsymbol{x}), \tag{2}$$

while the nonlinear mapping function $\boldsymbol{\psi} : R^m \mapsto R^n$ is defined as

$$\hat{\boldsymbol{x}} = \boldsymbol{\psi}(\boldsymbol{y}). \tag{3}$$

These mapping functions associate data with their principal components nonlinearly.

NLPCA is considered to decrease the MSE more than PCA does, because the mapping functions of NLPCA have a greater degree of freedom than the linear mapping functions of PCA.

2.3 The Formulation of the Proposed Method

In the proposed method, in order to construct principal components $y_1, y_2, \cdots, y_m$ in the order of their contributions to represent an input vector $\boldsymbol{x}$, the nonlinear mapping function $\phi_i : R^n \mapsto R^1$ is defined as

$$y_i = \phi_i(\boldsymbol{x}), \quad i = 1, \cdots, m, \tag{4}$$

while the nonlinear mapping function $\boldsymbol{\psi}_i : R^i \mapsto R^n$ from the product space $(y_1, \cdots, y_i) \in R^i$ onto the reconstructed vector $\hat{\boldsymbol{x}}_i \in R^n$ is defined as

$$\hat{\boldsymbol{x}}_i = \boldsymbol{\psi}_i(y_1, \cdots, y_i), \quad i = 1, \cdots, m. \tag{5}$$

The pairs of $(\phi_k, \boldsymbol{\psi}_k)_{k=1,\cdots,m}$ are adjusted in the increasing order of i. ϕ_i and $\boldsymbol{\psi}_i$ are optimized at the ith with the criterion:

$$\mathrm{MSE}_i = E[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_i(\boldsymbol{y}_i)\|^2] \tag{6}$$

$$= E[\|\boldsymbol{x} - \boldsymbol{\psi}_i(y_1, \cdots, y_{i-1}, \phi_i(\boldsymbol{x}))\|^2], \tag{7}$$

where $y_1, \cdots, y_{i-1}$ are given. Consequently, the pair of $(\phi_i, \boldsymbol{\psi}_i)$ is adjusted to perform the best extractor and reconstructer combined with the previous pairs of mapping functions: $(\phi_k, \boldsymbol{\psi}_k)_{k=1,\cdots,i-1}$.

We call the principal component of small i the higher component, and the component of large i the lower component.

In the proposed method, pairs of nonlinear mapping functions $(\phi_k, \boldsymbol{\psi}_k)_{k=1,\cdots,m}$ are implemented with neural networks that have a hierarchical structure as shown in Fig. 2. In this figure, the i-th principal component: y_i is generated in the i-th unit of the third layer (the extraction layer) of the i-th sub-network. The sub-networks are adjusted in the increasing order of i. A back propagation algorithm is applied for the adjustment.

In Fig. 2, the activation function of the units in the input layer (to which $\boldsymbol{x}$ is input), the extraction layer (which outputs y), and the output layer (which outputs $\hat{\boldsymbol{x}}$) is

$$f(u) = u, \tag{8}$$

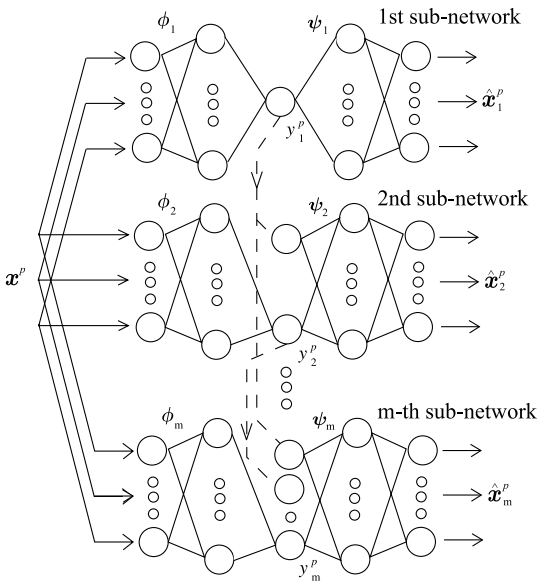

Fig. 2 The proposed neural network. The network is composed of the number of m sub-networks. In the i-th sub-network, the left three layers from the middle layer, and the right three layers play the role of ϕ_i and ψ_i, respectively. The i-th sub-network is given the values of principal components $y_1, \cdots, y_{i-1}$ from the higher $1, \cdots, (i-1)$-th sub-networks. The sub-networks are adjusted in the increasing order of i with a back propagation algorithm.

while the activation function of the units in the other layers is a monotonous nonlinear function such as a hyperbolic tangent

$$f(u) = \tanh\left(\frac{u}{T}\right), \tag{9}$$

where T is the constant value on nonlinearity and u is the weighted sum of inputs to the units.

We employed a threshold in the nonlinear units in the second and the fourth layer of the proposed network.

3. Experiments

We carried out some experiments to examine the effectiveness of the proposed method. In this section, we present two experimental results with facial images and hand-written numerals.

3.1 Dimensionality Reduction of Facial Images

We carried out a dimensionality reduction experiment with facial images sampled from the UMIST Face Database [19].

In this experiment, we sampled 750 training images and 250 test images of 20 peoples. The resolution of the image is 16×16 pixels in 256 gray scales.

We adjusted the parameters of the PCA and the proposed method with training samples. Next, we calculated the MSE of these methods with training and test samples, respectively. In the experiment, we reduced the dimensionality from 256 into 10, and employed the values shown in Table 1 for the parameters of the proposed method.

Table 1 Parameters of the proposed method.

no. of principal components	10
learning rate	0.001
parameter T	1.0
no. of hidden neurons	200
no. of training iterations	30,000
initial w	random over [-0.03, 0.03]

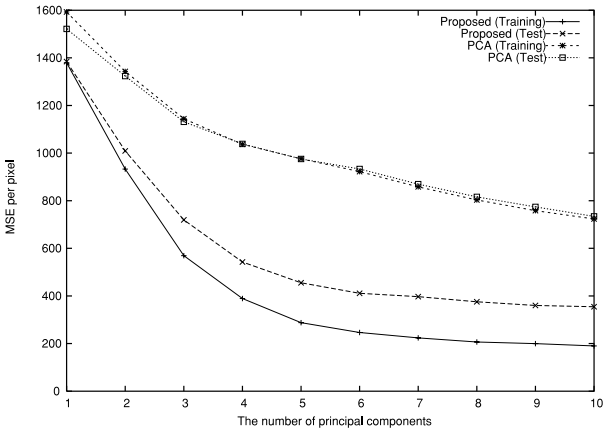

Fig. 3 The MSE of Principal Component Analysis (PCA) and the proposed method in reconstructing the training and test samples. The horizontal axis indicates the number of principal components employed in the reconstruction, and the vertical axis indicates the MSE per pixel. The principal components are applied to reconstruct the images in the order from the higher to the lower components.

Figure 3 shows the MSE with training and test samples by PCA and the proposed method. The horizontal axis indicates the number of principal components in the reconstruction, and the vertical axis indicates the MSE per pixel. In the experiment, the principal components are applied to reconstruct the image in order from the higher to the lower components.

As shown in Fig. 3, when the number of principal components increases, the MSE of the proposed method on training samples decreases more rapidly than that of the PCA. The MSE of the proposed method on test samples is also superior to that of the PCA, which shows the proposed method does not over-train. This result represents the effectiveness of the proposed method in dimensionality reduction.

Figure 4 shows the images which are extracted and reconstructed from the training samples and the test samples by PCA and the proposed method. The blocks of the images from top to bottom correspond to the results on the training samples by PCA, the results on the training samples by the proposed method, the results on the test samples by PCA, and the results on the test samples by the proposed method. The images in row (No) are the results of the (No)-th target image. The images in column (T) are the target images. The images in column (1), (3), $\cdots$, and (9) are reconstructed with one, three, $\cdots$, and nine principal components, respectively.

As shown in Fig. 4, the images reconstructed by the proposed method are high in fidelity with a small number

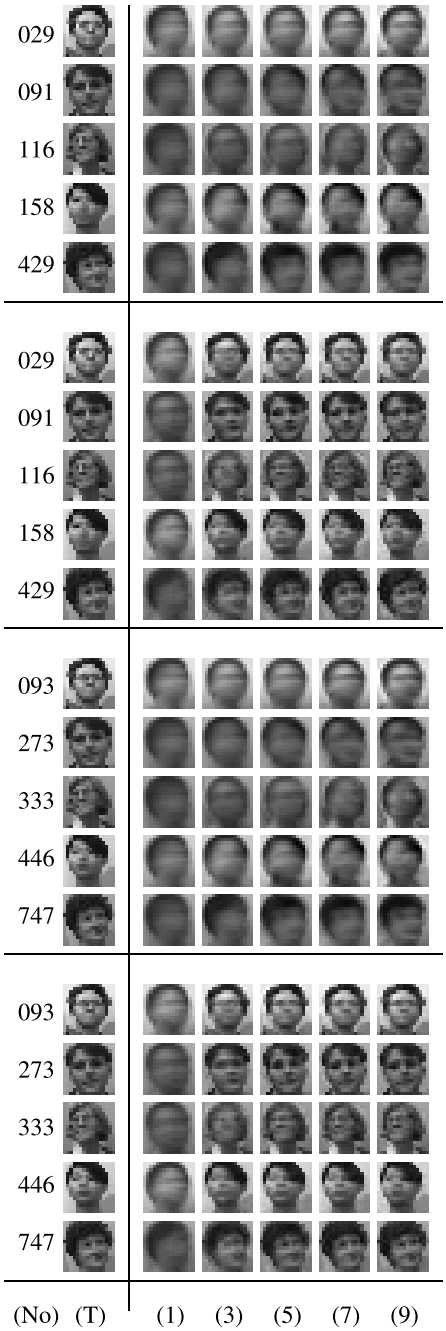

Fig. 4 The images reconstructed from the training samples and the test samples by PCA and the proposed method. The blocks of the images from top to bottom correspond to the results on the training samples by PCA, the results on the training samples by the proposed method, the results on the test samples by PCA, and the results on the test samples by the proposed method. The images in row (No) are the results of the (No)-th target image. The images in column (T) are the target images. The images in column (1), (3), $\cdots$, and (9) are reconstructed with one, three, $\cdots$, and nine principal components, respectively.

of principal components. The advantage of the proposed method over PCA is considered to come from the high representation ability of nonlinear mapping functions.

The set of facial images in this experiment are sampled from peoples whose sex and race are different. In this case, it is reported that the distribution of the data has nonlinearity [20]. In this experiment, nonlinearity of the proposed method is considered to contribute to the effective extraction and reconstruction of facial images.

3.2 Feature Extraction of Hand-Written Numerals

We demonstrated experiments on dimensionality reduction of hand-written numerals with PCA and the proposed method.

In character recognition, dimensionality reduction is important to reduce the computational cost and to prevent the so-called curse of dimensionality, which means that the number of samples required for the estimation increases in the order of the dimensionality of the feature space.

Conventionally, linear methods such as PCA and Discriminant Analysis are mainly used in dimensionality reduction of character recognition. However, the proposed method may be more effective for such recognition, because its nonlinear mapping functions can represent the samples in high fidelity. Moreover, since the proposed method preserves the order of principal components, we can easily select the dimensionality number of the feature vectors.

In this experiment, we used an image database of hand-written numerals, IPTP-CDROM1. The database has 10 categories which correspond to the numerals from zero, one, $\cdots$, to nine. Figure 5 shows sample images of the category "zero". Any image in the database is represented in the binary scale. We randomly chose 20,000 training samples and 4,950 test samples for each category, respectively.

In a preprocessing, we transformed the original images into gray-scale images in low resolution. In this reduction, we divided the original image into 8×8 square regions. We defined the gray-scale value of a region as the number of zero-brightness-value pixels included in the region. A zero-brightness-value pixel corresponds to a pixel which is black. After we obtained the set of the 64-dimensional gray-scale images, we normalized the images in order to set the average and the standard deviation to be zero and one, respectively.

After the preprocessing, we adjusted the parameters of the PCA and the proposed method with the training samples of all categories, and calculated the principal components of the training samples. We defined the average vector of the principal components of a category as the template vector of the category.

Then, we calculated the principal components of the test samples, and classified the principal component vectors with the template vectors subject to the Nearest Neighbor Rule (NNR). NNR classifies a vector into a category to which the nearest template belongs.

In the experiment, we employed the parameters shown in Table 2.

Figure 6 shows the recognition rate with PCA and with the proposed method. In Fig. 6, the horizontal axis indicates the number of the principal components employed for recognition.

The figure shows the effectiveness of the proposed

Fig. 5 Sample images of the category "zero" in an image database of hand-written numerals IPTP-CDROM1. The database has 10 categories which correspond to the numerals from zero, one, $\cdots$, to nine.

Table 2 Parameters of the proposed method.

no. of principal components	10
learning rate	0.0005
parameter T	1.0
no. of hidden neurons	20
no. of training iterations	100,000
initial w	random over [-0.03, 0.03]

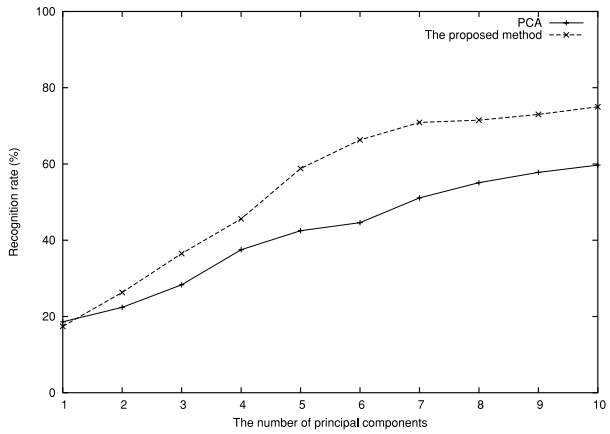

Fig. 6 Recognition rate with Principal Component Analysis and the proposed method. The horizontal axis indicates the number of principal components used to classify a type of characters. Principal components are applied to the classification in the order from the higher to the lower components.

method in pattern recognition. When the number of principal components increases, the recognition rate increases in the both methods.

In this experiment, the recognition is performed for a pattern compressed into low-dimensional space by PCA and by the proposed method. In general, the compression and recognition are based on different assumptions so that the category of a pattern is taken into account in recognition, but is not in compression. Therefore, the performance of a compressor cannot be evaluated only by the performance of its recognition rate. The recognition rate is affected by the relationship between the compressor and the classifier, which in this experiment is the NNR and templates.

4. Discussion

4.1 Complexity and Redundancy

In this paper, we do not describe in detail how to determine the complexity of the extraction and reconstruction functions, and the balance between them.

The complexity of the mapping functions is considered to depend on the user's policy on how to extract the structure of the data-distribution. For example, we can represent all the samples with one principal component corresponding to a curve which connects all samples, if the mapping functions have enough complexity and their adjustment is successful. In this example, however, we cannot extract the summarized structure of the distribution. There exits a tradeoff between the complexity of the mapping functions and the number of the principal components. An external policy out of the proposed method will be required to determine the balance point.

The principal components of the proposed method have redundancy in their scale, since the proposed method has no regularization for the extraction function, such as the normalization of the eigenvectors in PCA.

In the recognition experiment we determined the parameters of PCA and the proposed method in advance of the classification. The class was not taken into account in the adjustment. Therefore, their classification performance depended on the principal components of the pre-determined compressor, whether their scale includes redundancy or not.

How to give the scale of the principal components is application-oriented. A PCA gives a uniform scale, while we can give a variable scale in consideration of the density of the distribution; for example, we can assign a small scale to a high-density region and a large scale to a low-density region. We will try to regularize the scale with some criterion.

4.2 Comparison to Other Methods

Let us assume the ensemble of the m different SMLPs, and the i-th member of the ensemble has i hidden units in the bottleneck layer. In this case, the ensemble of the SMLPs will demonstrate the same performance as the proposed method with respect to reconstruction accuracy.

However, the principal components extracted in each SMLP are not ordered on their contribution of the reconstruction, while the principal components in the proposed method are ordered so that their principal components represent the summarized structure of the data-distribution in the order of significance.

The computational costs of neural-network based methods in an extraction process can be estimated by counting the number of multiplication operations among input vectors and weight values, and by the number of calculations for a sigmoidal function.

We compared computational costs of PCA and neural-network based methods such as a single SMLP with m units in the bottleneck layer, an ensemble of SMLPs each of which has $1, 2, \cdots, m$ units in the bottleneck layer, the Takahashi's Method with m units in the bottleneck layer, and the proposed method which employs a threshold in the units of the hidden layer, where n, m, l corresponded to the dimensionality of input vectors, the dimensionality of principal components, and the number of hidden units, respectively. The numbers of multiplication operations and calculations

Table 3 Computational costs of methods.

Method	Multiplication	Sigmoid
PCA	mn	–
Single SMLP	l(m+n)	*l*
Takahashi's	l(m+n)	*l*
Proposed	lm(n+1)	*ml*
Ens.of SMLPs	lm{(m+1)/2+n}	*ml*

for a sigmoidal function are shown in Table 3.

As shown in Table 3, the proposed method requires higher computational costs than PCA, a single SMLP and Takahashi's method in an extraction process. However, considering the advanced computational ability of the present days, these costs are small enough in a practical application.

As is described in Sect. 1, KPCA has several problems in its practical use with respect to computational costs, dimensionality reduction, and data-reconstruction.

When we perform KPCA for a dataset which contains N samples, we have to solve the eigenvalue problem of the $N \times N$ dot product matrix in the training process. In the test process, we have to evaluate the kernel function N times to extract each principal component. If N is large, both the proposed method and KPCA will be time-consuming in the training process, since the training process depends on N. However, the propose method will be less time-consuming than KPCA in the test process, since the test process in the proposed method does not depends on N.

In KPCA, the dimensionality of the principal components can be larger than the dimensionality of the input space, since the maximum number of the principal components is equal to N. In this case, the dimensionality is not reduced but increased.

The pre-image of a principal component is not known in KPCA, since the data-mapping is unidirectional from the input space into the feature space. For example, when we apply KPCA to a set of facial images, we can not obtain the facial image corresponding to a principal component vector.

These problems have been overcome with the additional optimization to the framework of KPCA [21], [22], however the simplicity of KPCA has been lost.

The crucial disadvantage of the approaches of the autoassociator and the proposed method is the problems of local minima [11], while KPCA, which perform a linear optimization, does not have these problems. However, trial and error is required to determine the proper values of the kernel parameters in KPCA. In the authors' opinion, another problem of KPCA is that KPCA does not explicitly state the way to obtain the kernel parameters in its framework, while the proposed method explicitly optimizes its parameter by training algorithms.

In comparison with non-PCA based methods such as Huffman coding, JPEG and GIF compression, the proposed method and PCA have a feature to construct mapping functions to reduce the dimensionality of the data. Therefore, it is possible to use the proposed method as a pre-compressor for the non-PCA based methods to achieve higher compression rates.

5. Conclusion

In this paper, we applied a proposed method of Nonlinear Principal Component Aanalysis (NLPCA) to the extraction and reconstruction of image data, and we examined the effectiveness of the proposed method in some experiments.

The proposed method does not only extract the characteristics of the distribution of the data, but also preserves the order of principal components. Therefore, the proposed method can be used as an efficient feature extractor for image recognition.

Preserving the order of the principal components will give us some advantages in practical use. The higher principal components extract a more significant summarization of the data-distribution. The user does not need to consider the order of principal components employed. Because of the hierarchy, user can determine the number of principal components after the adjustment of the mapping functions. Moreover, the user can add a new principal component with only an adjustment of an additional mapping function.

In the near future, we will discuss image recognition with other classifiers and pattern generation from perturbed principal components.

Acknowledgments

We sincerely thank Prof. Allinson in University of Manchester Institute of Science and Technology for licensing us to use the UMIST Face Database, and thank the Postal Research Institute for licensing us to use the hand-written numeral database IPTP-CDROM1. We appreciate the many brilliant comments we received from reviewers of this paper.

This work was supported by the 21st Century Center of Excellence Program, "The innovative research on symbiosis technologies for human and robots in the elderly dominated society" in Waseda University, and "Waseda University grant for special research projects, No. 2004B-882."

References

[1] H. Hotelling, "Analysis of complex statistical variables into principal components," Journal of Educational Psychology, vol.24, pp.417–441, pp.498–520, 1933.
[2] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed., Springer-Verlag, 2000.
[3] T. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," Neural Netw., vol.2, pp.459–473, 1989.
[4] K. Diamantaras and S. Kung, Principal Component Neural Networks Theory and Applications, John Wiley & Sons, 1996.
[5] K. Watanabe, H. Ito, H. Matsuda, and T. Oohori, "Multi-tandem perceptron for K-L transformation," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J75-D-II, no.11, pp.1925–1932, Nov. 1992.
[6] K. Watanabe, T. Oohori, and T. Shimozawa, "A theoretical study on the convergibility of unit perceptron," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J75-D-II, no.11, pp.1933–1939, Nov. 1992.
[7] H. Masuda, T. Oohori, and K. Watanabe, "A three-layered neural network for K-L transformation," IEICE Trans. Inf. & Syst.

(Japanese Edition), vol.J77-D-II, no.2, pp.397–404, Feb. 1994.
[8] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, vol.3, no.1, pp.71–86, 1991.
[9] H. Murase and S. Nayar, "3D object recognition from appearance—Parametric eigenspace method," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J77-D-II, no.11, pp.2179–2187, Nov. 1994.
[10] R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations, John Wiley & Sons, 1977.
[11] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput., vol.10, no.5, pp.1299–1319, 1998.
[12] J. Karhunen and J. Joutsensalo, "Generalization of principal component analysis, optimization problems, and neural network," Neural Netw., vol.8, no.4, pp.549–562, 1995.
[13] T. Hastie and W. Stuetzle, "Principal curves," Journal of the American Statistical Association, vol.84, no.406, pp.502–516, 1989.
[14] B. Irie and M. Kawato, "Acquisition of internal representation by multi-layered perceptron," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J73-D-II, no.8, pp.1173–1178, Aug. 1990.
[15] D. DeMers and G. Cottrell, "Nonlinear dimensionality reduction," in Advances in Neural Information Processing Systems 5, pp.580–587, Morgan Kaufmann, 1993.
[16] T. Takahashi, R. Tokunaga, and Y. Hirai, "On supervised learning algorithm of three-layer linear perceptron—An extension of Baldi-Hrnik's theorem," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J80-D-II, no.5, pp.1267–1275, May 1997.
[17] R. Seagusa, H. Sakano, and S. Hashimoto, "Nonlinear principal component analysis to preserve the order of principal components," Neurocomputing, no.61, pp.57–70, 2004.
[18] R. Saegusa and S. Hashimoto, "On the evaluation of a nonlinear principal component analysis," Proc. 2nd Int. Conf. on Neural Networks and Computational Intelligence, pp.66–72, 2004.
[19] D.B. Graham and N.M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences, vol.163, ed. H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman-Soulie, and T.S. Huang, pp.446–456, Springer Verlag, 1998. The UMIST Face Database, http://images.ee.umist.ac.uk/danny/database.html
[20] T. Suenaga, A. Sato, and H. Sakano, "Cluster discriminant analysis for feature space visualization," Proc. IEE Int. Conf. on KES, pp.146–150, 2002.
[21] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in Advances in Neural Information Processing Systems 11, pp.536–542, MIT Press, 1999.
[22] M. Tipping, "Sparse kernel principal component analysis," in Advances in Neural Information Processing Systems 13, pp.633–639, MIT Press, 2001.

Ryo Saegusa received his B. Eng., M. Eng. and Dr. Eng. degrees in Applied Physics from Waseda University, Tokyo, Japan, in 1999, 2001 and 2005, respectively. He has been a research associate at the Department of Applied Physics of Waseda University since 2004. His research interests include neural computing, data analysis and image processing.

Hitoshi Sakano received his B.S. degree in Physics from Chuo University, Tokyo, Japan, in 1988 and his M.S. degree in Physics from Saitama University, Saitama, Japan, in 1990. Since 1990, he has been at NTT Data Co., Tokyo, Japan. His research interests include pattern recognition such as character recognition and biometrics. Since 2000, he has been a visiting researcher at the Advanced Research Institute for Science and Engineering, Waseda University, Tokyo, Japan.

Shuji Hashimoto received his B.S., M.S. and Dr. Eng. degrees in Applied Physics from Waseda University, Tokyo, Japan, in 1970, 1973 and 1977, respectively. He is currently a Professor in the Department of Applied Physics, School of Science and Engineering, Waseda University. Since 2000, he has been a director of the Humanoid Robotics Institute, Waseda University. From 1979 to 1991, he was with the Faculty of Science, Toho University, Chiba, Japan. His research interests are in human communication and Kansei information processing, including image processing, music systems, neural computing and humanoid robotics.