社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2005－ＡＶＭ－51（６）
2005／12／12

Facial Expression Recognition by Supervised ICA with Selective Prior

Fan CHEN† and Kazunori KOTANI†

† School of Information Science,
Japan Advanced Institute of Science and Technology,
1-1, Asahidai, Nomi, Ishikawa 923-1292 Japan
E-mail: {chen-fan, ikko}@jaist.ac.jp

Abstract Feature selection is required when using the Independent Component Analysis (ICA) in feature extraction for pattern classification. Selection during ICA might provide a better candidate set of features. We propose a supervised ICA with a selective prior for the de-mixing coefficients so that those features with higher significance in discrimination could emerge easier during the learning. We formulate the learning rule for the supervised ICA in a form of the natural gradient approach and develop the algorithm of supervised ICA in facial expression analysis. The efficiency of the proposed algorithm has been investigated by numerical experiments.

Key words Facial Expression Recognition, ICA, Feature Selection, Sparse Coding

1. Introduction

In an appearance-based method for facial expression recognition, features for classification are extracted by projecting the image into a subspace that is most significant in discriminating samples into a predefined number of clusters. Direct optimization of some specified criteria of separability, as the linear or nonlinear discrimination analysis does, might cause heavy overfitting because the dimension of image is usually much larger than the size of training set. Primary Component Analysis (PCA) and ICA focus on extracting some statistical properties of samples, which to some extend improve the generality of extracted features. In facial expression recognition, some facts suggest that ICA might be more effective than PCA in feature extraction. Facial expression consists of those features standing for minor, non-rigid, local variations of faces, which are usually less significant in PCA bases than those for lighting, head pose, and personal difference. [1] Further, the phase spectrum, related to higher-order statistics, contains more structural information in images that drives human perception than the power spectrum. [2] The importance of higher-order statistics in natural images to the response properties of cortical cells has been explored in Refs. [3] [4] [5], and the extraction of higher-order statistics by means of ICA was discussed in Ref. [6]. ICA has been applied to the face recognition in Ref. [2] and to the facial expression analysis in Ref. [7], where the efficiency of ICA was verified.

In the classical ICA, the derived independent components are fully exchangeable in order, i.e., permutation ambiguity, where the original order provides no information on the significance of components in discrimination. A feature selection is necessary to be performed along with the feature extraction. The selection can be applied before, after or during ICA. In Ref. [2], Best Individual Feature (BIF) selection was adopted where features were chosen according to some defined criteria individually. Methods by means of Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) were also proposed. [8] Since the selection is performed after ICA, the features are limited to be chosen from the set of the obtained independent components. To create a candidate set with enough representative features in discrimination, a large number of independent components should be learned, which may be computationally expensive. It is meaningful to search for a way to affect the selection of features before or during ICA. The Generalized Eigenspace Method based on Class-features (GEMC) [9] makes a selection before ICA by heuristically replacing PCA with a discriminant analysis as the pre-processing to ICA, which still lacks a mathematical explanation. ICA in a local facial residue space is also proposed for face recognition, which uses the pre-specified residue space to limit the selection of independent components before applying ICA. [10]

In the present paper, we consider an approach to implement the feature selection during the search of independent components. A constraint ICA has been proposed for the analysis of electroencephalogram (EEG) signals, where all component should be sparse and close to a supplied reference signal by including a correlation term. [11] In our case, we try to design a method to let those components with

higher degree of separation emerge easier than others. Exactly, from the aspect of information maximization, the classical ICA is computed under the scheme of Maximum Log-Likelihood(MLL) estimation. [12] Instead of using the uniform prior for demixing coefficients in MLL, we take the Maximum a Posteriori(MAP) estimation. A prior defined on the degree of separation is introduced on the demixing coefficients, which in turn increases the probability of the corresponding component to be significant in classification.

In Section 2, we will formulate the supervised ICA and give the algorithm for facial expression recognition. In Section 3, numerical experiments are made and the performance of our proposed algorithm is investigated by making comparison with the classical ICA. We also discuss on the influence of the introduced selective prior. Finally, we summarize the present paper and explain our future work.

2. Supervised Independent Component Analysis

We first formulate the supervised ICA. Let $Y = [y^{(ki)}|k \in \{1,\cdots,K\}, i \in \{1,\cdots,N_k\}]$ be the matrix of N observed samples from K classes with N_k samples in the k-th class and satisfy $N = \sum_{k=1}^{K} N_k$. The i-th sample of class k, $y^{(ki)} = [y_1^{(ki)},\cdots,y_D^{(ki)}]^T$, is a D-dimensional vector. Provided Y as the training data set, the classical ICA assumes that these samples are generated from Q statistically independent sources. $S = [s^{(ki)}|k \in \{1,\cdots,K\}, i \in \{1,\cdots,N_k\}]$ represents the signals generated by those sources, where $s^{(ki)} = [s_1^{(ki)},\cdots,s_Q^{(ki)}]^T$ corresponds to $y^{(ki)}$. Those signals from different sources are linearly mixed, i.e., $Y = VS$, where the D-row Q-column matrix V is for the mixing coefficients. The purpose of ICA is to search for the coefficients V that makes the sources as statistically independent as possible. If we let $W = V^{-1}$ be the inverse (or pseudo-inverse) of V, W is the demixing matrix and satisfies $S = WY$. For any sample y, the extracted feature in ICA will be $s = Wy$. Note that we consider the noiseless case in the present paper.

Bell and Sejnowski have proposed an algorithm from the viewpoint of information maximization, where V is learned from samples by maximizing the log-likelihood criterion [12], i.e.,

$$V_{\mathrm{ICA}} = \arg\max_{V} \log P(Y|V). \tag{1}$$

Motivated by the reasons described in the introduction, we search for a way to make a selection of features during ICA so that those independent components with higher degree of separation are easier to emerge than others, which is achieved by introducing a prior distribution for the coefficients. We derive the learning rules by means of the MAP estimation, where V is obtained by maximizing the following criterion, i.e.,

$$\begin{aligned} V_{\mathrm{sICA}} &= \arg\max_{V} \log P(V|Y) \\ &= \arg\max_{V} [\log P(Y|V) + \log P(V)]. \end{aligned} \tag{2}$$

As in the classical ICA, $\log P(Y|V)$ is derived as [12]

$$\begin{aligned} &\log P(Y|V) \\ &= \log \int P(Y|V,S)P(S)dS \\ &= \log \int \prod_d \prod_k \prod_i \delta\{y_d^{(ki)} - \sum_q (V_{dq} s_q^{(ki)})\} \\ &\quad \times \prod_k \prod_i \prod_q P_q\{s_q^{(ki)}\} dS \\ &= -N \log|V| + \sum_k \sum_i \sum_q \log P_q\{\sum_d [V^{-1}]_{qd} y_d^{(ki)}\} \end{aligned} \tag{3}$$

with $\delta\{x\}$ being the Dirac delta function. Without special explanations, k varies from 1 to K while i varies from 1 to N_k for the suffixes of summation here and hereafter. We define the prior as follows:

$$P(V) = P(W) = \prod_q P_w(w_q), \tag{4}$$

$$P_w(w) = \frac{1}{Z_w} \exp\{\lambda w [M_{bc}(Y) - M_{wc}(Y)] w^T\}, \tag{5}$$

where $w_q = [w_{q1},\cdots,w_{qD}], W = [w_1^T,\cdots,w_Q^T]^T$. Z_w is the partition function while $M_{bc}(Y)$ and $M_{wc}(Y)$ are the between-class scatter matrix and within-class scatter matrix, defined by Eqs.(6) and (7).

$$M_{bc}(Y) = \frac{1}{N} \sum_k N_k ||\overline{y}^{(k)} - \overline{y}||^2, \tag{6}$$

$$M_{wc}(Y) = \frac{1}{N} \sum_k \sum_i ||y^{(ki)} - \overline{y}^{(k)}||^2. \tag{7}$$

We define $M_s(Y) = M_{bc}(Y) - M_{wc}(Y)$ for short. $\overline{y}^{(k)}$ represents the mean vector for samples in class k and $\overline{y}$ is the mean value for all samples. λ is a hyper-parameter introduced to control the influence of the prior. For $\lambda > 0$, an independent component whose demixing coefficients are of larger degree of separation will have a higher prior probability. Exactly, there are several choices for the prior, which should also coincide with the following process of classification. In out case, we take the form of subtraction between the between-class scatter matrix and within-class scatter matrix so that the prior could be controlled to avoid possible singularity. We maximize the MAP criterion

$$\begin{aligned} &\log P(V|Y) = \log P(W|Y) \\ &= N \log|W| + \sum_k \sum_i \sum_q \log P_q\{\sum_d w_{qd} y_d^{(ki)}\} \\ &\quad + \lambda w M_s(Y) w^T + Const \end{aligned} \tag{8}$$

under the constraints of $||w_q|| = 1$ for all $q \in \{1, \cdots, Q\}$, by differentiating the criterion with respect to w_{qd} according to the following rule, i.e.,

$$\frac{\partial}{\partial w_{qd}} \log|W| = [W^{-1}]_{dq} = V_{dq}. \quad (9)$$

The differential reads

$$\frac{\partial \log P(Y,W)}{\partial w_{qd}} = NV_{dq} + \sum_k \sum_i \frac{P_q'(\sum_d w_{qd} y_d^{(ki)})}{P_q(\sum_d w_{qd} y_d^{(ki)})} y_d^{(ki)} + 2\lambda \sum_l w_{ql}[M_s(Y)]_{ld}. \quad (10)$$

We take $P_q(x) \propto \cosh(x)$ and rewrite these differential equations for all w_{qd} into a compact form as one matrix differential which is defined as component-wise differentiation, i.e.,

$$\frac{\partial \log P(Y,W)}{\partial W} = N\{V^T - \frac{1}{N}\tanh[S]Y^T + \frac{2\lambda}{N} W M_s(Y)\}, \quad (11)$$

where $\tanh[S]$ means the calculation of tanh over all elements in matrix S. Due to the existence of inverse matrix V which is computationally expensive in the iterative learning, we adopt the natural gradient approach, proposed by Amari [13], to derive the learning rule:

$$\Delta W = \eta \left\{ \frac{\partial \log P(Y,W)}{\partial W} \right\} W^T W = N\eta\{I - \frac{1}{N}\tanh[S]S^T + \frac{2\lambda}{N} W M_s(Y) W^T\} W, \quad (12)$$

where η is the learning rate. Comparing with the classical ICA, our supervised ICA holds a prior term for demixing coefficients.

When applied to facial expression recognition, the supervised ICA is performed on the PCA coefficients instead of directly on the image data X, i.e., $Y = W_{\rm PCA}X$. $W_{\rm PCA}$ is the matrix of PCA eigenvectors. Since all eigen-vectors that correspond to nonzero eigen-values in PCA are adopted, there is no information lost during this preprocessing. The updating rule is finally derived as follows:

$$W^{(t+1)} = W^{(t)} + N\eta\{I - \frac{1}{N}\tanh[S^{(t)}][S^{(t)}]^T + \frac{2\lambda}{N} W^{(t)} M_s(Y)[W^{(t)}]^T\} W^{(t)}. \quad (13)$$

The algorithm for the supervised ICA is summarized in Table 1 and the final bases for extracting features are computed as $W_{\rm F} = WW_{\rm PCA}$. Instead of using a Langrange multiplier, we simply implement the constraint $||w_q|| = 1$ in Step (b) of Table 1. Exactly, the scale of w_q should not affect the sparseness of derived components in the classical ICA, i.e., scale ambiguity. As a fix-point learning algorithm, the behavior of convergence is still not fully predictable. The introduction of Step (b) requires a different learning rate and a different convergence threshold. Therefore, it is difficult to make a

Table 1 The learning algorithm of the supervised ICA

a) Initialize W and calculate $M_{bc}(Y) - M_{wc}(Y)$;
b) Normalize W by rows so that $||w_q|| = 1$;
c) Calculate S from $S = WY$;
d) Calculate ΔW;
e) Update W by $W \leftarrow W + \Delta W$;
f) Calculate $\log P(Y,W)$. If the difference between two iterations is less than a threshold, exit. If not, repeat (b) to (f).

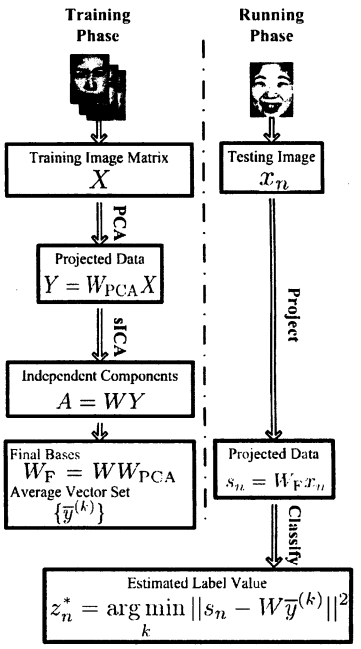


Fig. 1 A block diagram for the processing flows in both the learning phase and the running phase of facial expression recognition. All input image data will be normalized in face position and histogram-equalized as preprocessings.

Nor Hap Ang Fea Dis Sad Sur

Fig. 2 Some samples in our numerical experiments from the JAFFE database. (Nor: Normal Hap: Happiness Ang: Anger Fea: Fear Dis: Disgust Sad: Sadness Sur: Surprise)

precise analysis on the influence of Step (b). From numerical experiments, whose data are not given in the present paper, we have found no significant differences in the recognition rate caused by applying Step (b) to the classical ICA. For the supervised ICA, the constraint $||w_q|| = 1$ is required to stabilize the influence of the prior term, which also helps improve the convergence behavior of the algorithm.

Let $\widehat{X} = [x_n | n \in \{1, \cdots, \widehat{N}\}]$ be the matrix by putting all testing images into different columns and $\widehat{N}$ be the number of samples in the testing set. We define $Z = \{z_n \in \{1, \cdots, K\} | n \in \{1, \cdots, \widehat{N}\}\}$ to represent the true classified labels for observed data, and define a recognition rate as

$$r_c = \frac{1}{\widehat{N}} \sum_{n=1}^{\widehat{N}} \delta(z_n, z_n^*), \tag{14}$$

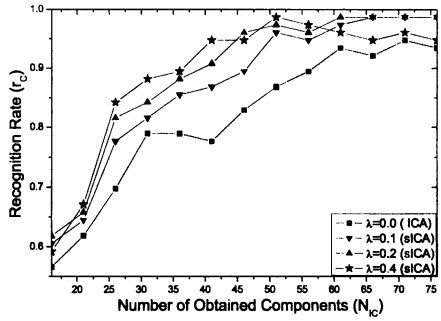
where $\delta(x, y)$ is the Kronecker delta. z_n^* is the estimated label value which is estimated according to the following criterion

$$z_n^* = \arg\min_k ||s_n - W\overline{y}^{(k)}||^2 \tag{15}$$

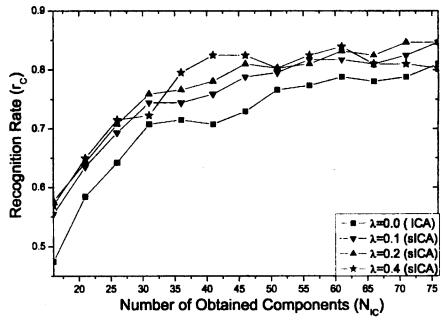
and $s_n = W_F x_n$. A block diagram for the whole process is given in Fig. 1.

3. Experiments and Discussions

Different from the compact coding in PCA, ICA performs sparse coding, i.e., each pattern could be represented only by a small set of the bases. The maximum number of bases is unbounded and no information of discrimination from the training dataset is explicitly included. As a result, there is no guarantee on whether the features we have extracted are really significant in discrimination or not, because we could never obtain the whole set of indepedent components. With a worse candidate set, even the post selection might fail to provide satisfiable results. A considerable way to make improvements is to apply the selection during the search of independent components so that a better candidate set could be found, which is the major idea of our proposed algorithm. In the following numerical experiments, we therefore will focus on the comparison between the supervised ICA and the classical ICA under the same conditions to investigate the effect by introducing the prior term and by changing the hyperparameter λ. The database we are using is the Japanese Female Facial Expression (JAFFE) Database [14], which includes 213 images in total. We normalize these images and pick up 76 images to form the training set. The obtained bases are tested both on the training set and a testing set consisting of the remaining 137 images. Some normalized samples are given in Fig. 2. All images are resized to 32×40 pixels. Thus the dimension of x_n is 1280 and the dimension of Y is $D = 76$, which is the number of all nonzero eigen-values in PCA on the training set.



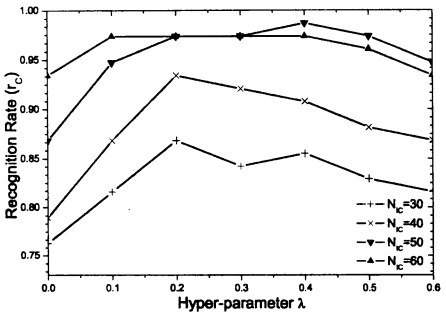
(a)



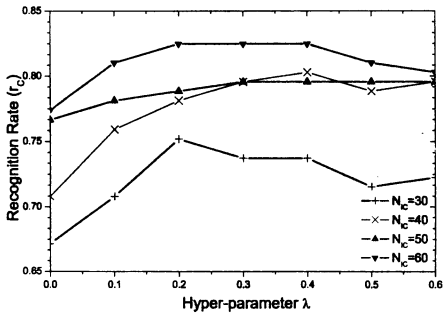
(b)

Fig. 3 Recognition rate r_c is plotted as a function of the number of independent components N_{IC} under different λ for (a) the training set of samples and (b) the testing set. The supervised ICA (sICA) outperforms the classical ICA, especially for a median N_{IC}.

We first plot the recognition rate r_c as a function of the number of independent components (N_{IC}), which is equal to Q in Section 2, for the training set and the testing set under different λ values in Fig. 3 (a) and (b), respectively. The learning rate η is set to be 0.00001 for all cases. We find that higher recognition rates have been achieved by including the selective prior for almost all N_{IC} values, which suggest that a better set of candidate features can be found by the supervised ICA. Different from those discrimination analysises which make direct optimizations on some criteria of separability, the supervised ICA includes the selective prior only to affect the selection of independent components without deteriorating the sparseness. In Fig. 4 (a) and (b), we further investigate the transition of recognition rate as a function of λ at different N_{IC}. For both the cases of training set and testing set, the recognition rates first ascend with the increase of λ and then descend when λ gets too large and causes a heavy bias on the sparseness of the obtained



(a)



(b)

Fig. 4 Recognition rate r_c is plotted as a function of λ at different $N_{\rm IC}$ for (a) the training set of samples and (b) the testing set. The learning rate η is set to be 0.00001 for all the cases. A properly selected λ helps improve the performance.

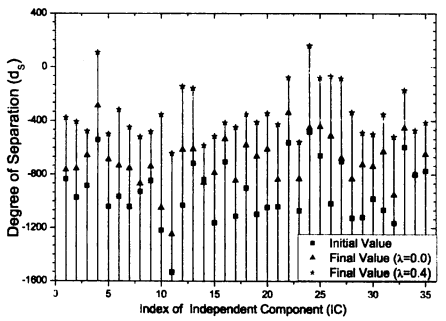


Fig. 5 Degree of separation $d_s(w)$ is plotted for the case of $N_{\rm IC} = 35$. Each vertical bar stands for one independent component. Starting from the same initial values denoted by the rectangles, the final values of $d_s(w)$ for both the cases $\lambda = 0.0$ and $\lambda = 0.4$ are given by the triangles and the stars, respectively.

independent components. Therefore, a tradeoff between the sparseness and the discrimination degree should be taken to achieve the best results.

We define the degree of separation $d_s(w)$ as $d_s(w) = wM_s(Y)w^T$, which is plotted in Fig. 5 for the case of $N_{\rm IC} = 35$. Each vertical bar represents one independent component. Starting from the same initial values denoted by the rectangles, the final values of $d_s(w)$ for both the cases $\lambda = 0.0$ and $\lambda = 0.4$ are given by the triangles and the stars, respectively. For all components, the ratio of separation increases with a large λ, which proves the effect of the prior term. As a result, those components with higher degree of separation improve the recognition rate. In Fig. 6, we visualize the base from the 10th, the 11th, and the 27th components from ICA, sICA with $\lambda = 0.3$ and sICA with $\lambda = 0.6$. With the increasing of λ, the area emphasized by the components, such as the eyebrow and nose, gets broader, which might be the reason for improved discrimination. Due to the use of natural gradient approach, the local optimal of sICA is near to those of ICA for a given initialization. A selection over a wider range requires an estimating method that could jump between different local optimals, e.g. sampling.

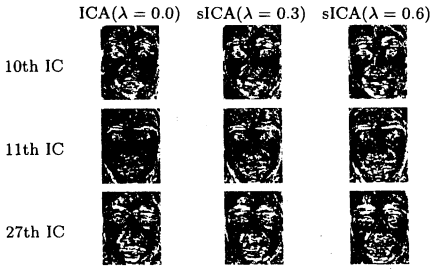
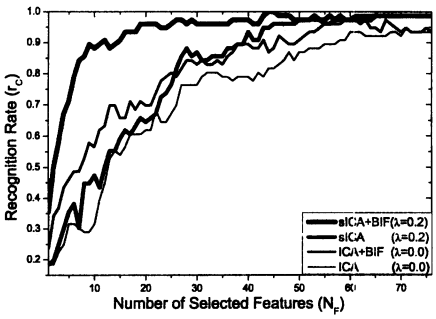


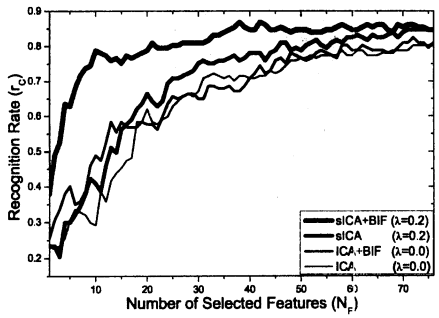
Fig. 6 Some basis images for the case of $N_{\rm IC} = 35$.

Although several methods for the feature selection after the learning of ICA were proposed in Ref. [8], they are also applicable to our approach. The supervised ICA intends to search for a candidate set of features with higher degree of separation than the classical ICA, by selecting features from which even better recognition rate can be achieved. In Fig. 7(a) and (b), we make a comparison between four methods, i.e., the supervised ICA with Best Individual Feature(BIF) selection, the supervised ICA without BIF, the classical ICA with BIF, and ICA without BIF. In the BIF selection of the present paper, all features are sorted in descendant order of their degree of separation $d_s(w)$, and then the first $N_{\rm F}$ features are selected for classification. We note that the supervised ICA with BIF gives the best performance, which verifies the capacity of the supervised ICA in learning a better candidate set of features. We also find that the supervised ICA without BIF still outperforms the classical ICA with BIF, which confirms the robustness of the supervised

ICA in recognition rate by learning those independent components with higher degree of separation from samples when a median N_{IC} is used. On the other hand, although BIF improves the robustness of the performance over the whole range of N_F, the best recognition rate does not change much only by means of BIF selection for the same learning algorithm, as depicted in Figs. 7 (a) and (b). This result suggests that learning a candidate set of features with higher degree of separation might be more important than performing a post selection, which is the point where the supervised ICA outperforms the classical ICA.



(a)



(b)

Fig. 7 Recognition rate r_c is plotted as a function of N_F, which is the number of features we select, for sICA with BIF, sICA without BIF, the classical ICA with BIF and ICA without BIF. (a) Result for the training set. (b) Result for the testing set. We can find that since the supervised ICA provides a better set of candidate features, the supervised ICA with BIF selection has the best performance.

4. Conclusion and Future Work

Utilization of discrimination information from the given training dataset is essential to a successful recognition. In the classical ICA, an unsupervised method of feature extraction, no classification information of training set is included explicitly. We have proposed a supervised ICA for facial expression recognition in the present paper. The major purpose is to improve the significance of obtained features in discrimination. A selective prior has been introduced to the classical ICA and the learning rule is derived under the MAP scheme. We made numerical experiments to investigate the influence of new prior term and make comparison with the classical ICA. Our method shows better performance than the classical ICA, especially in increasing the recognition rate under a median number of independent components. There are still some problems left for us to study, such as the decision of optimal λ and the design of a better learning algorithm for faster and more robust convergence. Investigation on various priors is also a part of our future work.

References

[1] C. Nastar, "Face recognition using deformable matching," in Face Recognition: from Theory to Applications, Wechsler, H., et al., eds., pp.206-229, Springer-Verlag, New York.

[2] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," IEEE Trans. on Neural Networks, vol.13 pp.1450-1464, 2002.

[3] D. J. Field, "What is the goal of sensory coding?" Neural Computation, vol.5, pp.559-601, 1994.

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

[5] E. Simoncelli, and B. Olshausen, "Natural image statistics and neural representation," Annual Review of Neuroscience, vol.24 pp.1193-1216, 2001.

[6] Y. Karklin, and M. S. Lewicki, "Learning higher-order structures in natural images," Network: Comput. Neural Syst., vol.14 pp.483-499, 2003.

[7] M. S. Bartlett, G. L. Donato, J. R. Movellan, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Image representations for facial expression coding," Advances in Neural Information Processing Systems, vol.12, pp.886-892, 2000.

[8] H. K. Ekenel, and B. Sankur, "Feature selection in the independent component subspace for face recognition," Pattern Recognition Letters, vol.25, pp.1377-1388, 2004.

[9] I. Eguchi, and K. Kotani, "Facial expression analysis by generalized eigen-space method based on class-features(GEMC)," Proc. 2005 IEEE Int'l Conf. on Image Processing, vol.1, MonAmPO3-6, 2005.

[10] T. K. Kim, H. Kim, W. Hwang, and J. Kittler, "Independent component analysis in a local facial residue space for face recognition," Pattern Recognition, vol.37, pp.1873-1885, 2004.

[11] C. J. James, and O. Gibson, "Electromagnetic brain signal analysis using constrained ICA," Proc. 2nd European Medical and Biological Engineering Conference, vol.1, pp.426-427, 2002.

[12] A. J. Bell, and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, vol.7, pp.1129-1159, 1995.

[13] S. Amari, "Natural gradient works efficiently in learning," Neural Computation, vol.10, pp.251-276, 1998.

[14] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," Proc. 3rd IEEE Int'l Conf. on Automatic Face and Gesture Recognition, vol.1, pp.200-205, 1998.