LH-004

Facial Expression Recognition by Fast Supervised Independent Component Analysis

Fan CHEN * Kazunori KOTANI †

1 Introduction

In Principle Component Analysis(PCA), local variations caused by facial expressions are less significant than global variations, e.g., lighting conditions, personal differences, etc. [1] Independent Component Analysis(ICA) is argued to be more effective than PCA in feature extraction for facial expression recognition, due to its ability of encoding local variations by performing sparse coding.[2]

A problem in ICA is permutation ambiguity, i.e., the derived independent components are fully exchangeable in order, where the original order provides no information on the significance of components in discrimination. As a result, the derived independent components may not be most distinctive for the classification task, especially when only a small subset of components is derived. One solution is to include a process of feature selection into the feature extraction of ICA. Selection after performing ICA, as the Best Individual Feature (BIF) selection in Ref [3], limits the universe set for choosing features. Selection before performing ICA, as GEMC [4] that replaces PCA with MDA as preprocessing to ICA does, still lacks a mathematical understanding. A natural way is to incorporate feature selection into ICA. Especially, we try to design a method to let those components with higher degree of class separation emerge easier than others. The classical ICA in Ref.[5] was shown to be derivable under the scheme of Maximum Log-Likelihood (MLL) estimation. [6] Instead of using the uniform prior for de-mixing coefficients in MLL, we take the Maximum *a* Posteriori (MAP) estimation. A prior defined on the degree of class separation is introduced to the de-mixing coefficients, which in turn increases the probability of the corresponding independent component to be significant in classification.

In the present paper, we include classification information into ICA to propose a supervised ICA(sICA) and derive a fixed-point learning algorithm for facial expression recognition. Numerical experiments show that our method significantly improved the robustness of recognition rate under a median number of independent components, which is meaningful in speeding up the extraction of distinctive features.

2 Facial Expression Recognition by sICA

2.1 Generative Model of sICA

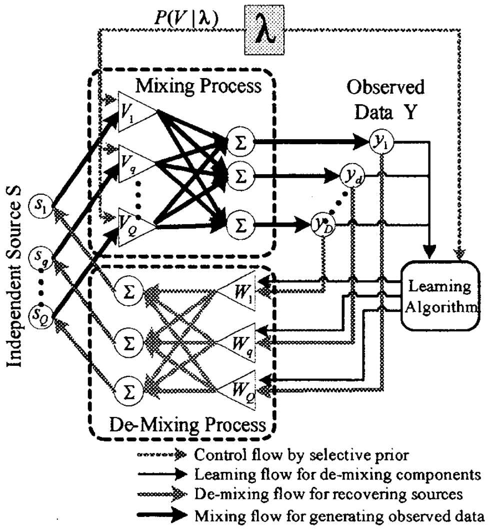


Figure 1: Generative model of sICA and the corresponding de-mixing process.

Fig. 1 illustrates the generative model of sICA. ICA assumes that observed samples Y are linear mixed signals from Q statistically independent sources S, i.e., $Y = VS$. V is the mixing matrix and $W = V^{-1}$ is the corresponding de-mixing matrix that satisfies $S = WY$. ICA tries to estimate W to recover S from Y. Different from maximizing the log-likelihood criterion which assumes an uniform prior for V, i.e., $V_{\text{ICA}} = \arg\max_V \log P(Y|V)$[6], we search for a way to make a selection of most distinctive features during ICA, by introducing a priori for the coefficients and maximizing the following MAP criterion,

$$V_{\text{sICA}} = \arg\max_V [\log P(Y|V) + \log P(V)]. \quad (1)$$

Note that we use $P(V|Y) = \frac{P(V,Y)}{P(Y)} \propto P(V,Y) = P(Y|V)P(V)$. $P(V) = P(W)$ is defined on the degree of separation:

$$P(W) = \prod_q P_w(w_q) = \prod_q \frac{\exp\{\lambda w M_s(Y) w^T\}}{Z_w}, \quad (2)$$

where $w_q = [w_{q1}, \cdots, w_{qD}], W = [w_1^T, \cdots, w_Q^T]^T$. Z_w is the partition function while $M_s(Y)$ is defined as

$$M_s(Y) = \frac{1}{N}\{\sum_k N_k ||\overline{y}^{(k)} - \overline{y}||^2 - \sum_{ki} ||y^{(ki)} - \overline{y}^{(k)}||^2\}. \quad (3)$$

*School of Information Science, Japan Advanced Institute of Science and Technology, Ashahi-dai 1-8, Nomi, Ishikawa 923-1292, Japan. E-mail: chen-fan@jaist.ac.jp

†School of Information Science, Japan Advanced Institute of Science and Technology,Ashahi-dai 1-8, Nomi, Ishikawa 923-1292, Japan. E-mail: ikko@jaist.ac.jp

$\overline{y}^{(k)}$ 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 de-mixing coefficients are of larger degree of class separation will have a higher prior probability.

2.2 Fixed-point sICA

Different from natural gradient-based ICA with fixed learning rate[7], fixed-point ICA provides improved robustness and speed of convergence by taking adaptive learning rate. Without loss of generality, we assume all sample data $\{y^{(ki)}\}$ are mean-centered. A pre-whitening preprocessing is first applied to the sample data Y to reduce the complexity of ICA by de-correlating sample data. We solve the eigen-decomposition on the covariance matrix $E\{yy^T\}$ as $E\{yy^T\} = U^T\Lambda U$, where U is the matrix of eigenvectors and Λ is the diagonal matrix of all eigenvalues. Pre-whitened sample matrix Y^{pw} is computed as $Y^{pw} = W^{pw}Y$ where $W^{pw} = U^T\Lambda^{1/2}U$ is the whitening matrix. Independent components are then estimated by maximizing the criterion in Eq.(1) on pre-whitened samples Y^{pw}. We further diagonalize the scatter matrix $M_s(Y^{pw})$ as $M_s(Y^{pw}) = A^T\Lambda_s(\hat{Y})A$, where A is an $D \times D$ orthogonal matrix which satisfies $A^TA = I$.

If we let $\hat{w} = wA^T$ which is a $1 \times D$ vector, we rewrite the prior in Eq.(2) as

$$P(W^{pw}) = \prod_q P_w(w_q^{pw}) = \prod_q P_{\hat{w}}(\hat{w}_q), \qquad (4)$$

$$P_{\hat{w}}(\hat{w}) = \frac{1}{Z_{\hat{w}}}\exp\{\lambda\hat{w}\Lambda_s(\hat{Y})\hat{w}^T\}. \qquad (5)$$

Accordingly, the posteriori is given as

$$\log P(V^{pw}|Y^{pw}) = logP(\hat{V}|\hat{Y}) = N\log|\hat{W}|$$
$$+ \sum_{k,i,q}\log P_q\{\sum_d \hat{w}_{qd}\hat{y}_d^{(ki)}\} + \lambda\sum_q \hat{w}_q\Lambda_s(\hat{Y})\hat{w}_q^T + C.$$

Since $[y^{pw}]^{(ki)}$ has been pre-whitened, we have uncorrelatedness and unit variance of the $\hat{S}^{(ki)}$. Accordingly, the following relation, i.e.,

$$|I| = |E\{\hat{s}\hat{s}^T\}| = |\hat{W}||A||E\{y^{pw}[y^{pw}]^T\}||A||\hat{W}^T|, \quad (6)$$

means that $|\hat{W}|$ is constant. In order to optimize the likelihood under the constraint $||w_q|| = 1$, we define the criterion by using the Lagrange multiplier,

$$L = \log P(\hat{W}|\hat{Y}) + \sum_q \alpha_q(\hat{w}_q\hat{w}_q^T - 1) \qquad (7)$$

with its first-order differential with respect to $\hat{w}_{qd}$ being

$$\frac{\partial L}{\partial\hat{w}_{qd}} = \frac{\partial\log P(\hat{W}|\hat{Y})}{\partial\hat{w}_{qd}} + 2\alpha_q\hat{w}_{qd}$$
$$= -NE[g(\hat{w}_q\hat{y})\hat{y}_d] + 2\lambda\hat{w}_q[\Lambda_s(\hat{Y})]_d + 2\alpha_q\hat{w}_{qd} \qquad (8)$$

with $g(x) = P_q^{'}(x)/P_q(x)$. In a vector form of differential, we derive

$$\frac{\partial L}{\partial\hat{w}_q} = -NE[g(\hat{w}_q\hat{y})\hat{y}^T] + 2\lambda\hat{w}_q[\Lambda_s(\hat{Y})] + 2\alpha_q\hat{w}_q. \quad (9)$$

To maximize the criterion, we set the first-order differential to zero, i.e.,$\partial L/\partial\hat{w}_q = 0$, under the constraint $||w_q|| = 1$ and obtain α_q, i.e.,

$$0 = 2\alpha_q\hat{w}_q\hat{w}_q^T - NE[g(\hat{w}_q\hat{y})\hat{y}^T]\hat{w}_q^T + 2\lambda\hat{w}_q[\Lambda_s(\hat{Y})]\hat{w}_q^T,$$
$$\alpha_q = \frac{N}{2}E[g(\hat{w}_q\hat{y})\hat{w}_q\hat{y}] - \lambda\hat{w}_q[\Lambda_s(\hat{Y})]\hat{w}_q^T. \qquad (10)$$

The second-order gradient is

$$\frac{\partial^2 L}{\partial\hat{w}_{qd}^2} = -NE[g^{'}(\hat{w}_q\hat{y})\hat{y}_d^2] + 2\lambda[\Lambda_s(\hat{Y})]_{dd} + 2\alpha_q$$
$$\approx -NE[g^{'}(\hat{w}_q\hat{y})] + 2\lambda[\Lambda_s(\hat{Y})]_{dd} + 2\alpha_q. \quad (11)$$

Finally, the fixed-point update rule for $\hat{w}_{qd}$ reads

$$\hat{w}_{qd} \leftarrow \hat{w}_{qd} - \frac{\partial L}{\partial\hat{w}_{qd}}/\frac{\partial^2 L}{\partial\hat{w}_{qd}^2}$$
$$= \frac{NE[g(\hat{w}_q\hat{y})\hat{y}_d] - N\hat{w}_{qd}E[g^{'}(\hat{w}_q\hat{y})]}{-NE[g^{'}(\hat{w}_q\hat{y})] + 2\alpha_q + 2\lambda[\Lambda_s(\hat{Y})]_{dd}} \qquad (12)$$

By taking an adaptive learning rate, fixed-point algorithm may provide faster and more robust estimation than the natural gradient method which uses a fixed learning rate. If we let $g(\hat{S})$ and $g^{'}(\hat{S})$ to denote the element-wise calculation of functions $g(x)$ and $g^{'}(x)$ on sources $\hat{S} = \hat{W}\hat{Y}$, we have $NE[g^{'}(\hat{w}_q\hat{y})] = [g^{'}(\hat{S})\mathbf{1}]_q$. We further simplify the update rule as

$$\hat{W} \leftarrow \Phi \circ [g(\hat{S})\hat{Y}^T - G^{'}\hat{W}], \qquad (13)$$

by defining $G^{'} = \text{diag}\{NE[g^{'}(\hat{w}_q\hat{y})]|q = 1,\cdots,Q\}$ and a $Q \times D$ matrix Φ with $\Phi_{qd} = [-G_q^{'} + 2\alpha_q + 2\lambda[\Lambda_s(\hat{Y})]_{dd}]^{-1}$. The symbol $\circ$ defines the component-wise multiplication of two matrices. The algorithm is given in Table 1. Finally, we have de-mixed source as

$$S = \hat{S} = \hat{W}\hat{Y} = \hat{W}A^TW^{pw}Y = WY \qquad (14)$$

and de-mixing bases as $W = \hat{W}A^TW^{pw}$.

2.3 Apply sICA to Expression Recognition

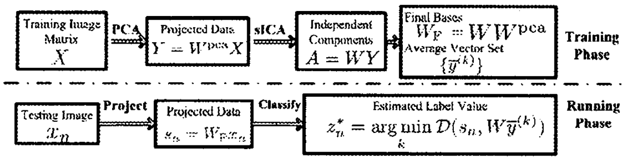

Training Image Matrix X
PCA
Projected Data $Y = W^{pca}X$
sICA
Independent Components $A = WY$
Final Bases $W_F = WW^{pca}$
Average Vector Set $\{\overline{y}^{(k)}\}$
Training Phase
Testing Image x_n
Project
Projected Data $s_n = W_F x_n$
Classify
Estimated Label Value $z_n^* = \arg\min_k \mathcal{D}(s_n, W\overline{y}^{(k)})$
Running Phase

Figure 2: 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.

We perform sICA on the PCA coefficients Y instead of directly on the image data X where $Y = W^{pca}X$.

Accordingly, the final bases for extracting features are computed as $W^{\mathrm{F}} = WW^{\mathrm{pca}}$. Let $\widetilde{X} = [x_n | n \in \{1, \cdots, \widetilde{N}\}]$ be the matrix by putting all testing images into different columns and $\widetilde{N}$ be the number of samples in the testing set. We define $Z = \{z_n \in \{1, \cdots, K\} | n \in \{1, \cdots, \widetilde{N}\}\}$ as the true classified labels for observed data, and define a recognition rate as $r_c = \frac{1}{\widetilde{N}} \sum_{n=1}^{\widetilde{N}} \delta(z_n, z_n^*)$, where $\delta(x, y)$ is the Kronecker delta. z_n^* is the estimated label value which is the label of the class whose center has the minimum distance to the current feature vector, i.e.,

$$z_n^* = \arg\min_{k} \mathcal{D}(s_n, W\bar{y}^{(k)})$$

and $s_n = W_{\mathrm{F}} x_n$. $\mathcal{D}(x_a, x_b)$ calculates the distance between two feature vectors, which is defined as $\mathcal{D}(x_a, x_b) = 1.0 - (x_a^T x_b)/(|x_a||x_b|)$. A block diagram for the whole process is given in Fig. 2.

3 Experiments and Discussions

We will focus on the comparison between sICA and ICA under same conditions to investigate the effect by introducing the prior and by changing hyper-parameter λ for facial expression analysis.

3.1 Experimental Conditions

We use the Japanese Female Facial Expression (JAFFE) Database [8], which includes 213 images in total. These images are aligned in face position and histogram-equalized. Some samples are given in Fig. 3. All images are resized to 64×80 pixels. The goal of recognition is to classify them into neutral face or one of six elemental facial expressions suggested by Ekman et al.[9], i.e., happiness, anger, fear, disgust, sadness and surprise. Numerical experiments have been performed on 4 randomly selected training sets with size N being $70, 80, 90$ and 100, respectively. For each N, a pair of two mutually exclusive sets was created, one with N images for training, and another with $213 - N$ images for testing. Average results over five-time computations from different random initializations are compared in the present paper.

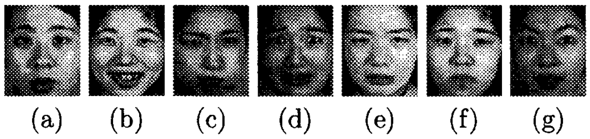


Figure 3: Several normalized samples in the JAFFE database. (a) Neutral (b) Happiness (c) Anger (d) Fear (e) Disgust (f) Sadness and (g) Surprise.

3.2 Recognition Rates in sICA

Performances of ICA and sICA in facial expression recognition are compared in Fig.4 for classifying all testing datasets. For comparison, performances of PCA and MDA (a multi-class extension of linear discriminant analysis) are also compared in our experiments. Recognition rate r_c is plotted as a function of Q with hyper-parameter λ being set to 0.35 ad hoc. In our experiments of facial expression recognition, ICA

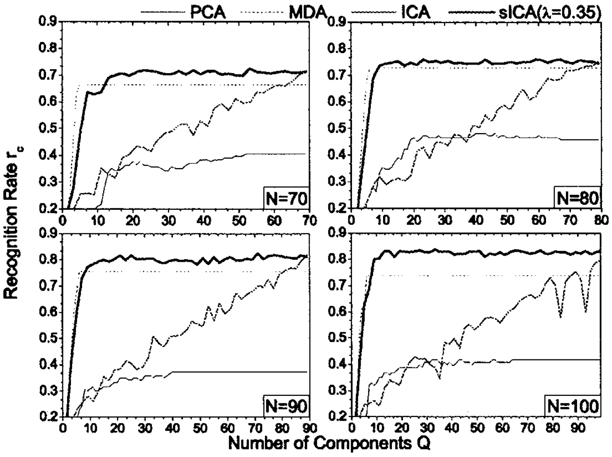


Figure 4: Recognition rates r_c by using PCA, MDA, ICA and sICA are compared on the four testing datasets with training sample size N being 70,80,90 and 100 in four graphs, respectively. The hyper-parameter λ is heuristically set to 0.35.

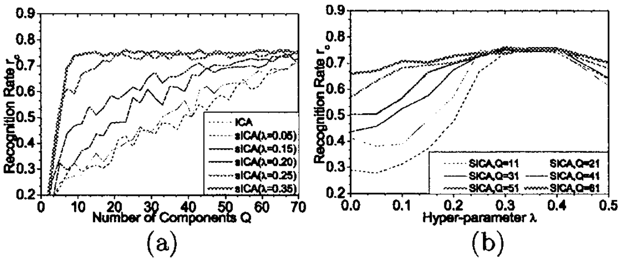
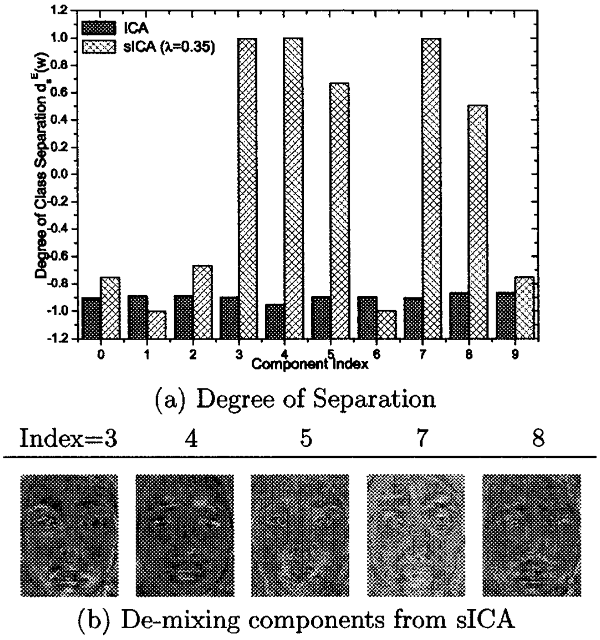


Figure 5: Recognition rate r_c is plotted as a function of number of components Q at different hyper-parameter λ in (a) for the testing dataset with training sample size $N = 80$. The transition of r_c w.r.t. λ is shown in (b). A properly selected λ helps improve the performance.



(a) Degree of Separation

Index=3 4 5 7 8

(b) De-mixing components from sICA

Figure 6: Degrees of class separation $d_s(w)$ are compared on components from sICA and classical ICA. Number of components Q is set to 10. Each bar represents one component. Five components with highest $d_s(w)$ appear in (b).

outperforms PCA in all cases. The reason is thought to be that facial expression is mainly caused by local variations which are less significant in PCA bases than those global variations, e.g., pose variation, lighting, and personal difference, etc.

Our sICA significantly improves the robustness of performance under a much smaller number of components. With this characteristic of sICA, it is possible to derive a small subset of independent components that are distinctive in classification, which will significantly reduce the training time when dealing with a large training dataset. A smaller set of independent components also speeds up the running phase of recognition. Compared to MDA, sICA increases the recognition rate by 5 points in average for all testing datasets.

3.3 Characteristics of sICA

a) Behavior under Different Hyper-parameter λ

We plot recognition rate r_c as a function of λ at different number of components Q for $N = 80$ in Fig.5(a). With the increasing of λ, the robustness of recognition rate r_c against small number of components Q could be improved. The transition of recognition rate r_c w.r.t. hyper-parameter λ appear in Fig.5(b), which shows that too large λ causes a heavy bias on the sparseness of the obtained independent components and then deteriorates the result. A tradeoff between the sparseness and the discrimination degree should be taken to achieve the best results.

b) Degree of Separation in sICA

We define the degree of class separation $d_s(w) = wM_s(Y)w^T$ for de-mixing component w. In Fig.6, all components derived by sICA are compared to those derived by ICA in $d_s(w)$. sICA achieves higher degrees of class separation than ICA for almost all components. A common feature between components of high $d_s(w)$ is that they both emphasize on some facial parts, e.g., eyebrows, cheeks, mouth corners, etc., that are thought to be essential in understanding facial expressions.

4 Conclusion and Future Work

We have proposed a supervised ICA for facial expression recognition by performing the feature selection along with the learning of ICA. We include a selective prior defined on the scatter matrix into ICA and derive the learning rule in a fixed-point algorithm. Numerical experiments show that our method outperforms ICA, especially in increasing the recognition rate under a median number of independent components. Our future works include the decision of optimal λ and the investigation on various priors.

Table 1: sICA by fixed-point algorithm

a) Set the number Q of independent components and Y be the matrix of sample data;

b) Produce pre-whitened data Y^{pw} from Y by computing $Y^{pw} = W^{pw}Y$;

c) Calculate and diagonalize $M_s(Y) = A^T\Lambda_s(\hat{Y})A$ where $\hat{Y} = A^TY^{pw}$ represents the sample data in the rotated subspace by A;

d) Initialize randomly $\hat{W}^{(0)}$;

e) Set $t = 0$ and convergence threshold ϵ.

f) Calculate de-mixed source $\hat{S}^{(t)} = \hat{W}^{(t)}\hat{Y}$;.

g) Calculate $g(\hat{S}^{(t)})$, $g^{'}(\hat{S}^{(t)})$, $\alpha_q^{(t)}$ and $\Phi^{(t)}$;

h) Update de-mixing matrix $\hat{W}$:
$\hat{W}^{(t+1)} \leftarrow \Phi^{(t)} \circ [g(\hat{S}^{(t)})\hat{Y}^T - [G^{'}]^{(t)}\hat{W}^{(t)}]$;

i) Do a symmetric orthogonalization on $\hat{W}$ as
$\hat{W}^{(t+1)} \leftarrow (\hat{W}^{(t+1)}[\hat{W}^{(t+1)}]^T)^{-1/2}\hat{W}^{(t+1)}$;

j) Check the convergence. If $\mathrm{Tr}|\hat{W}^{(t+1)}[\hat{W}^{(t)}]^T - I| < \epsilon$, go to (k). If not, repeat (f) to (j);

k) Compute final sICA bases as $W = \hat{W}A^TW^{pw}$.

References

[1] Nastar, C.: Face recognition using deformable matching. Face Recognition: from Theory to Applications(Wechsler, H., et al., Eds.), Springer-Verlag, New York, (1998) 206–229

[2] Bartlett, M. S., Donato, G. L., Movellan, J. R., Hager, J. C., Ekman, P., and Sejnowski, T. J.: Image representations for facial expression coding. Advances in Neural Information Processing Systems **12** (2000) 886–892

[3] Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J.: Face recognition by independent component analysis. IEEE Trans. on Neural Networks **13** (2002) 1450–1464

[4] Eguchi, I., and Kotani, K.: Facial expression analysis by generalized eigen-space method based on class-features (GEMC). ICIP'2005 **1** (2005) MonAmPO3–6

[5] Bell, A. J., and Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation **7** (1995) 1129–1159

[6] MacKay, D. J. C.: Maximum likelihood and covariant algorithms for independent component analysis. Technical report, University of Cambridge (1996)

[7] Amari, S.: Natural gradient works efficiently in learning. Neural Computation **10** (1998) 251–276

[8] Lyons, M. J., Akamatsu, S., Kamachi, M., and Gyoba, J.: Coding facial expressions with Gabor wavelets. Proc. 3rd IEEE Int'l Conf. on Automatic Face and Gesture Recognition **1** (1998) 200–205

[9] Ekman, P., Friesen, W.V., and Ellsworth, P., :Emotion in the human face: Guidelines for research and an integration of findings, Pergamon Press, New York, (1972).