INFORMATION FUSION BASED ON KERNEL ENTROPY COMPONENT ANALYSIS IN DISCRIMINATIVE CANONICAL CORRELATION SPACE WITH APPLICATION TO AUDIO EMOTION RECOGNITION

Lei Gao[1,2] *Lin Qi*[1], *Ling Guan*[2]

1. School of Information Engineering, Zhengzhou University
2.Department of Electrical and Computer Engineering, Ryerson University
iegaolei@gmail.com; ielqi@zzu.edu.cn; lguan@ee.ryerson.ca

ABSTRACT

As an information fusion tool, Kernel Entropy Component Analysis (KECA) is realized by using descriptor of information entropy and optimized by entropy estimation. However, as an unsupervised method, it merely puts the information or features from different channels together without considering their intrinsic structures and relations. In this paper, we introduce an enhanced version of KECA for information fusion, KECA in Discriminative Canonical Correlation Space (DCCS). Not only the intrinsic structures and discriminative representations are considered, but also the natural representations of input data are revealed by entropy estimation, leading to improved recognition accuracy. The effectiveness of the proposed solution is evaluated through experiments on two audio emotion databases. Experimental results show that the proposed solution outperforms the existing methods based on similar principles.

Index Terms— Information fusion, emotion recognition, kernel entropy component analysis, discriminative canonical correlation space

1. INTRODUCTION

Information Fusion is referred as a process which integrates a set of multiple information sources, associated features, and intermediate decisions to achieve more reliable recognition performance [1]. The effective utilization and integration of the content across multiple distinct yet complementary sources of information is becoming an increasingly important research topic in many applications. Since multimodal data contains more information, the combination of multimodal data may potentially provide a more complete and discriminatory description of the intrinsic characteristics of the patterns, and produce improved system performance [2].

However, the benefits usually come with certain challenges, with the main obstacles lying in the identification of a more complete and discriminative representations among multiple modalities[3]. In order to address the aforementioned concerns, investigations have been carried out into the performances of different information fusion techniques.

The existing information fusion methods are usually classified into four levels including data level, feature level, score level, and decision level [4]. In recent years, intelligent feature level fusion has drawn significant attentions from the research communities of multimedia and biometrics due to its capacity of information preservation and has made impressive progress [5-6]. Among these methods, the popular linear strategies are linear discriminant analysis (LDA), principal component analysis (PCA), canonical correlation analysis (CCA), etc. Recently, there has been extensive interest in non-linear feature transform. Instead of assuming linear relationship, kernel methods have been proposed to obtain non-linear correlation among the original data, which leads to kernel LDA, kernel PCA and kernel CCA.

Nevertheless, the theoretical foundation of these methods largely depends on the second order statistics, such as variance, correlation, mean square error and so on. Since the second order statistics are only optimal for Gaussian-like distribution, a poor estimator is usually obtained if the underlining distribution greatly differs from Gaussian, failing to reveal the nature of input data. In order to address this issue, kernel entropy component analysis (KECA) was proposed, which utilizes descriptor of information entropy and achieves better performance than the traditional kernel based methods [7-8]. However, as an unsupervised method, KECA merely put the information or features from different channels together without considering the intrinsic structures and relations. In this paper, we propose an enhanced KECA for information fusion, KECA in Discriminative Canonical Correlation Space (DCCS). Not only the intrinsic structures and discriminative representations are considered, but also the generic representations of input data are revealed, leading to improved recognition accuracy effectively.

The remainder of this paper is organized as follows: Section 2 introduces the KECA in DCCS (KECA+DCCS) and the proposed fusion procedure. In Section 3, implementation of the KECA+DCCS for audio emotion recognition is presented. The experimental results and analysis are given in Section 4. Conclusions are drawn in Section 5.

This work is supported by the National Natural Science Foundation of China (NSFC, No.61071211), the State Key Program of NSFC (No. 61331201), the Key International Collaboration Program of NSFC (No. 61210005) and the Discovery Grant of Natural Science and Engineering Council of Canada (No. 238813/2010).

2. KECA IN DISCRIMINATIVE CANONICAL CORRELATION SPACE

One of the major concerns for information fusion is to identify the complete and discriminative representations from different information sources. In this section, we introduce KECA+DCCS to address this problem. In the following, we first briefly describe the fundamentals of DCCS and KECA, and then formulate KECA+DCCS.

2.1. Discriminative Canonical Correlation Space

The DCCS for information fusion rests on the following facts: 1) the correlation among the features in different channels is taken as the metric of the similarity; 2) the within-class similarity and the between-class dissimilarity are considered jointly by DCCS.
Given two sets of zero-mean random features $x_1 \in R^{m_1 \times n}, x_2 \in R^{m_2 \times n}$ for *c* classes containing *n* samples and $Q = m_1 + m_2$. Concretely, the discriminative model aims to seek the projection vectors $\omega = [{\omega_1}^T, {\omega_2}^T]^T$ for fused features extraction so that the within-class similarity is maximized and the between-class dissimilarity is minimized. Therefore, it is formulated as the following optimization problem:

$$\arg\max_{\omega_1,\omega_2} \rho = {\omega_1}^T \overset{\sim}{C_{x_1x_2}} \omega_2 \tag{1}$$

Subject to

$${\omega_1}^T C_{x_1x_1} \omega_1 = {\omega_2}^T C_{x_2x_2} \omega_2 = 1 \tag{2}$$

where $\overset{\sim}{C_{x_1x_2}} = C_w - \delta C_b (\delta > 0), C_{x_kx_k} = x_k {x_k}^T (k = 1, 2)$. C_w and C_b denote the within-class and between-class matrixes of two sets, respectively. Then, C_w and C_b can be written in the form of $C_w = x_1 A {x_2}^T$, $C_b = -x_1 A {x_2}^T$ as shown in [9].
where

$$A = \left[\begin{pmatrix} H_{n_{i1} \times n_{i1}} & \cdots & 0 \\ \vdots & H_{n_{il} \times n_{il}} & \vdots \\ 0 & \cdots & H_{n_{ic} \times n_{ic}} \end{pmatrix} \right] \in R^{n \times n} \tag{3}$$

with n_{il} being the number of samples in the *l*th class of x_i set and $H_{n_{i1} \times n_{i1}}$ in the form of $n_{i1} \times n_{i1}$ with unit values for all the elements in $H_{n_{i1} \times n_{i1}} (i = 1, 2)$.
Substituting C_w and C_b into (1) yields:

$$(C - D)\omega = \rho D \omega \tag{4}$$

where

$$C = \left[\begin{pmatrix} x_1 {x_1}^T & x_1 A {x_2}^T \\ x_2 A {x_1}^T & x_2 {x_2}^T \end{pmatrix} \right],$$

$$D = \left[\begin{pmatrix} x_1 {x_1}^T & 0 \\ 0 & x_2 {x_2}^T \end{pmatrix} \right].$$

Eq. (4) is solved as the generalized eigenvalue(GEV) problem.

2.2. Kernel Entropy Component Analysis

Kernel Entropy Component Analysis (KECA) was proposed to improve Kernel Principal Component Analysis (KPCA). When selecting the appropriate eigenvectors for the best projection, the chosen eigenvectors have to contribute to the entropy estimate of the input data in KECA. The mathematical explanation of calculating the entropy estimate is summarized in this part:
Renyi quadratic entropy is given in Eq. (5) where probability density function of dataset is *p*(x):

$$H(p) = -\log \int p^2(x) dx \tag{5}$$

Eq. (6) is used instead of Eq. (5) owing to the monotonic nature of logarithmic functions:

$$V(p) = \int p^2(x) dx \tag{6}$$

Then *V*(p) is estimated in Eq. (7)

$$V(p) = \frac{1}{N} \sum\nolimits_{x \in D} \frac{1}{N} \sum\nolimits_{x_t \in D} k_\sigma(x, x_t) = \frac{1}{N^2} 1^T K 1 \tag{7}$$

where *N* is the number of samples.
Rewriting Eq. (7) leads to a different expression for *V(p)*

$$V(p) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\sqrt{\lambda_i} {\alpha^T}_i 1 \right)^2 \tag{8}$$

where λ_i and α_i are corresponding to eigenvalues and eigenvectors of *K*. The projection of KECA onto the *i*th principal axis in the kernel feature space is defined as

$$\Phi_{eca} = {D_i}^{\frac{1}{2}} {E_i}^T \tag{9}$$

where D_i consists of the top *i* eigenvalues of *K* and E_i stores the corresponding eigenvectors as columns. For more information, please refer to [10].

2.3. KECA+DCCS

In this section, we present the KECA+DCCS on two sets of zero-mean random features $x_1 \in R^{m_1 \times n}, x_2 \in R^{m_2 \times n}$. Let X_1, X_2 denote the projections of the two discriminative canonical vectors, respectively, i.e.

$$X_1 = {\omega_1}^T x_1;\ X_2 = {\omega_2}^T x_2 \tag{10}$$

which can be rewritten as

$$X = \begin{pmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \omega_{dccs}^T x \tag{11}$$

where

$$\omega_{dccs} = \begin{pmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{pmatrix} \tag{12}$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{13}$$

Based on the definition of KECA, KECA+DCCS is mathematically presented in the following equation:

$$V(p) = \frac{1}{n}\sum\nolimits_{X_1 \in R} \frac{1}{n}\sum\nolimits_{X_2 \in R} k_\sigma(X_1, X_2) = \frac{1}{n^2} 1^T K^{'} 1 \tag{14}$$

The projection of KECA+DCCS onto the *i*th principal axis in the kernel feature space is defined as

$$\Phi_{eca+dccs} = (D'_i)^{\frac{1}{2}} E'^T \tag{15}$$

where D'_i consists of the top *i* eigenvalues of *K'* and E'_i stores the corresponding eigenvectors as columns.

Compared with the solutions in [7-8], KECA+DCCS first transforms the original features into discriminative canonical correlation space. DCCS can be seen as a way of guiding discriminative feature selection toward the underlying semantics to find basis vectors for two sets of variables. It reveals discriminative representations among different groups of variables. Moreover, based on the definition of DCCS, as the transformed sets of linear combinations are those with the largest correlation subject to the condition that they are orthogonal to the former canonical variates, it can eliminate redundant information and integrate complementary data effectively before KECA is implemented. Therefore, it can potentially improve the recognition accuracy.

3. AUDIO EMOTION RECOGNITION BASED ON KECA+DCCS

In what follows, we examine the performance of KECA+DCCS in audio emotion recognition problem. Emotion recognition plays an important role in our daily social interactions and activities. It reflects an individual's state of the mind in response to the internal and external stimuli. In this work, we use the set of six principal emotions: happiness, sadness, anger, fear, surprise and disgust, proposed by Ekman and his colleagues [13]. The emotional state of an individual can be inferred from different sources such as voice, facial expressions, body language, ECG, and EEG. Among various modalities, audio is one most natural and noninvasive type of trait. Moreover, it can be easily captured by low-cost sensing devices, suitable for potential deployment in a wide range of applications.

3.1. Audio Features Extraction

To build an emotion recognition system, the extraction of features that can truly represent the representative characteristics of the intended emotion should be first properly addressed. Our goal is to simulate human perception of emotion, and identify possible features that can convey the underlying emotions in speech regardless of the language, identity, and context. As prosody and MFCC have been shown to be the two primary indicators of a speaker's emotional states [10], we investigate the fusion of two types of features which are extracted as follows:

(1) 25 prosodic features as used in [10];

(2) 65 MFCC features: the mean, median, standard deviation, max, and range of the first 13 MFCC coefficients;

3.2. Recognition Algorithm

In terms of emotion recognition algorithm, we use the method proposed in [11], which can be summarized as follows:

Given two sets of features, represented by feature matrices

$$X^1 = [{x^1}_1, {x^1}_2, {x^1}_3, ...{x^1}_d] \tag{16}$$

and

$$X^2 = [{x^2}_1, {x^2}_2, {x^2}_3, ...{x^2}_d] \tag{17}$$

$dist[X^1 X^2]$ is defined as

$$dist[X^1 X^2] = \sum_{j=1}^{d} \left\| {x^1}_j - {x^2}_j \right\|_2 \tag{18}$$

where $\|a - b\|_2$ denotes the Euclidean distance between the two vectors *a* and *b*.

Let the feature matrices of the *N* training samples as $F_1, F_2, ...F_N$ and each sample belongs to some class C_i $(i = 1, 2...c)$, then for a given test sample *I*, if

$$dist[I, F_l] = \min_j dist[I, F_j] \tag{19}$$

and

$$F_l = C_i \tag{20}$$

the resulting decision is $I = C_i$.

4. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of the proposed solution, experiments have been conducted on RML emotion database and eNTERFACE (eNT) emotion database [12], both with the 6 basic emotions (happiness, sadness, anger, fear, surprise and disgust). In RML database, the samples were recorded at a sampling rate of 22050 Hz, and a frame rate of 30 frames per second. For eNTERFACE database, all the samples are with a sampling rate of 48000 Hz and a frame rate of 25 frames per second. In the experiments 456 audio samples of eight subjects from RML database and 456 audio samples of ten subjects from eNTERFACE (eNT) database are selected, respectively. We divide both of the audio samples into training and testing subsets containing 360 and 96 samples each randomly. As a benchmark, the performances of using prosodic and MFCC in emotion recognition are first evaluated, which are shown as Table 1. The recognition accuracy is calculated as the ratio of the number of correctly classified samples over the total number of testing samples.

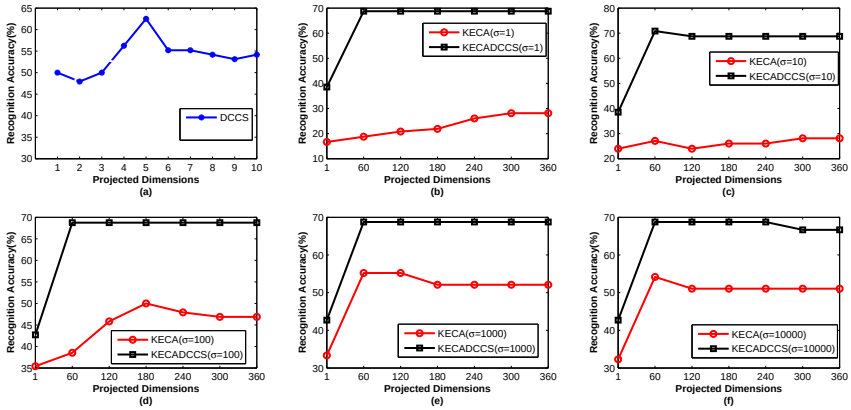

Fig.1 (a) Experimental results of RML emotion database based on DCCS; (b)-(f)Experimental results of RML emotion database based on KECA and KECA+DCCS (σ =1, 10, 100, 1000, 10000)

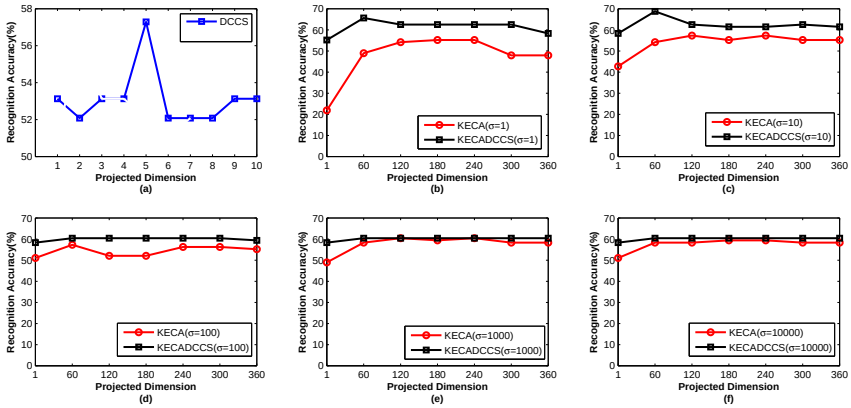

Fig.2 (a) Experimental results of eNT emotion database based on DCCS; (b)-(f)Experimental results of eNT emotion database based on KECA and KECA+DCCS (σ =1, 10, 100, 1000, 10000)

Since the performance of kernel based algorithms may be significantly affected by the selected kernel functions and the corresponding parameters, we have conducted extensive experiments using Gaussian kernels, and the reported results are based on these kernels with σ =1, 10, 100, 1000, 10000. Then, we compare the performance of KECA+DCCS with those by DCCS and KECA. During the comparison, the number of projected dimensions for DCCS is reduced to 10 and it achieves the best performance when the projected dimension is equal to 5, showing the discriminative power of DCCS. The overall recognition accuracies are shown in Fig. 1 and Fig. 2. From these figures, it can be seen that KECA+DCCS outperforms DCCS and KECA, in terms of recognition accuracy, especially on the RML database.

Table 1. Results of Emotion Recognition with Single Feature

Single Feature	Recognition Accuracy
Prosodic(RML)	51.04%
MFCC(RML)	37.50%
Prosodic(eNT)	50.00%
MFCC(eNT)	39.58%

5. CONCLUSIONS

In this paper, a new information theoretic tool, kernel entropy component analysis in discriminative canonical correlation space (KECA+DCCS), is introduced for information fusion and applied to audio emotion recognition. After processed by the proposed fusion method, most of useful information is properly preserved and better recognition accuracy is achieved. The proposed solution outperforms the existing methods based on similar principles. Experimental results demonstrate the feasibility of the proposed strategy. Although we focus on the information fusion between two sets of variables in this paper, information fusion with multiple sets of variables will be our future research tasks.

6. REFERENCES

[1] L. Guan, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki, and M. T. Ibrahim, "Multimodal information fusion for selected multimedia applications," *International Journal of Multimedia Intelligence and Security*, vol. 1, no. 1, pp. 5-32, 2010.

[2] T. Joshi, S. Dey, and D. Samanta, "Multimodal biometrics: state of the art in fusion techniques," *International Journal of Biometrics*, vol. 1, no. 4, pp. 393-417, 2009.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp.345-379, 2010.

[4] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human computer interfaces and intelligent environments: A survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692-1715, 2010.

[5] B. Khaleghi, K. Alaa, O. K. Fakhreddine and N. R. Saiedeh, "Multisensor data fusion: A review of the state-of-the-art." *Information Fusion*, vol. 14, no. 1, pp. 28-44, 2013.

[6] R. C. Luo and C. C. Chang. "Multisensor fusion and integration: A review on approaches and its applications in mechatronics." *IEEE Transactions on Industrial Informatics,* vol. 8, no. 1, pp. 49-60, 2012.

[7] Z. Xie, and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools." *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2013.

[8] Z. Xie, and L. Guan, "Multimodal information fusion of audio emotion recognition based on kernel entropy component analysis", *2012 IEEE International Symposium on Multimedia (ISM)*, pp.1-8, 2012.

[9] L. Gao, L. Qi, E. Chen, and L. Guan. "Discriminative multiple canonical correlation analysis for multi-feature information fusion." *2012 IEEE International Symposium on Multimedia (ISM)*, pp.36-43, 2012.

[10] R. Jenssen, Kernel entropy component analysis, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 5, pp. 847-860, 2010.

[11] B. H. Shekar, M. S. Kumari, L. M. Mestetskiy and N. F. Dyshkant,"Face recognition using kernel entropy component analysis," *Neurocomputing*, vol. 74, no. 6, pp.1053-1057, 2011.

[12] Y. Wang, L. Guan and A.N. Venetsanopoulos, "Kernel based fusion with application to audiovisual emotion recognition." *IEEE Trans. on Multimedia*, vol. 14, no. 3, pp. 597-607, Jun 2012.

[13] P. Ekman and W. Friesen, "Constants across cultures in the Face and Emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124-129, 1971.