International Journal of Semantic Computing
Vol. 7, No. 1 (2013) 25–42

DOI: 10.1142/S1793351X13400023

MULTIMODAL INFORMATION FUSION OF AUDIO EMOTION RECOGNITION BASED ON KERNEL ENTROPY COMPONENT ANALYSIS

ZHIBING XIE* and LING GUAN†

Ryerson Multimedia Research Lab
Ryerson University, 245 Church Street
Toronto, Ontario, Canada
http://www.rml.ryerson.ca
**zhibing.xie@ryerson.ca*
†*lguan@ee.ryerson.ca*

This paper focuses on the application of novel information theoretic tools in the area of information fusion. Feature transformation and fusion is critical for the performance of information fusion, however, the majority of the existing works depend on second order statistics, which is only optimal for Gaussian-like distribution. In this paper, the integration of information fusion techniques and kernel entropy component analysis provides a new information theoretic tool. The fusion of features is realized using descriptor of information entropy and is optimized by entropy estimation. A novel multimodal information fusion strategy of audio emotion recognition based on kernel entropy component analysis (KECA) has been presented. The effectiveness of the proposed solution is evaluated through experimentation on two audiovisual emotion databases. Experimental results show that the proposed solution outperforms the existing methods, especially when the dimension of feature space is substantially reduced. The proposed method offers general theoretical analysis which gives us an approach to implement information theory into multimedia research.

Keywords: Multimodal information fusion; emotion recognition; kernel entropy component analysis.

1. Introduction

Multimodal information fusion is described as a process which integrates a set of multiple information sources, extracted features, and intermediate decisions to achieve more reliable analysis results [1, 2]. The growing interest in the research of multiple modalities is due to its potential capabilities for eliminating certain critical limitations of single modality. The traditional human computer interfaces such as mouse and keyboard are considered too restrictive for a natural interaction. Numerous efforts have been focused on various non-intrusive sensors so that humans can conduct their activities in a more natural way without feeling the presence of these sensors. This necessitates the employment of multimodality, especially for multimedia systems. Hence the objective of multiple information fusion is to utilize complementary information and improve the accuracy of the overall recognition

performance. Since different sensors may carry redundant, complementary, or even conflict information, utilizing all data leads to a system which is able to understand the users more thoroughly than using a single modality.

2. Critical Issues of Information Fusion

2.1. *Challenges of information fusion*

The advantages of information fusion come with certain obstacles in the analysis process [3, 4]. The major reason for this is due to the dissimilar characteristics of the multiple modalities involved, since different modalities are basically captured in various formats at various rates. Therefore, in order to accomplish a task better, the fusion process needs to address several critical issues, for instance, synchronization of multiple modalities. Hence the selection of the techniques and algorithms used for fusion is the key factor to the performance and accuracy of a multimodal system.

Another important issue associated with multimodal techniques is the variation of input data. Multiple modalities usually have different levels of confidence and reliability in reaching a decision. Moreover, what needs more attention is that different modalities may be independent or correlated, since both independence and correlation equally provide valuable insight under different scenarios. In order to address the concerns mentioned above, investigations have been carried out into the performance of information fusion techniques.

2.2. *Levels of fusion*

The schemes of the existing information fusion are usually classified into five modules: sensor level fusion, feature level fusion, score level fusion, decision level fusion and hybrid fusion [4, 5].

2.2.1. *Sensor level fusion*

Sensor level fusion, or data level fusion, refers to the integration of raw information from two or more sensors. Although fusion at sensor level is expected to enhance the recognition accuracy, it usually cannot be used in multimodal fusion because of the incompatibility of data from different modalities.

2.2.2. *Feature level fusion*

Feature level fusion is a kind of solution which combines different feature vectors, obtained either with different modalities or by applying different feature extraction algorithms to the same modality. The fusion at feature level is also called as early fusion. Although the feature level contains richer information about the raw data, it is difficult to achieve time synchronization and same format, and hard to obtain correlation between heterogeneous features. Furthermore concatenation of feature vectors may result in a feature vector with high dimensionality, which largely increases the computational load.

2.2.3. *Score level fusion*

Score level fusion, which is known as intermediate level fusion, integrates matching scores provided by the several modalities. Its advantages include simple implementation and scalability.

2.2.4. *Decision level fusion*

Decision level fusion is also called as late fusion which refers to the combination of likelihood values or probability scores achieved from separate single modality to make a combined decision. The fusion at this level allows extensive flexibility in choosing individual classifiers. However, the fusion at decision level has disadvantages, including loss of correlation at feature level and tedious learning process.

2.2.5. *Hybrid fusion*

Hybrid fusion is a method based on the strategies mentioned above which has been developed by many researchers, since an approach of hybrid fusion can combine the benefits of both early and late fusion.

3. Limitations of Existing Methods

Because of heterogeneous measurements from different features or modalities, the extracted data are often imprecise and incomplete. Before classification, removing the redundant data, integrating complementary information and processing feature vectors are essential. Therefore feature transformation and fusion methods are used to create a subset of new features by a combination of the original features.

Linear strategies are typically used to discard redundant components and reduce high dimensionality of the data. The objective is to obtain higher discrimination of low dimensional data from the original high dimensional data. The widely used linear approaches include linear discriminant analysis (LDA), principal component analysis (PCA), canonical correlation analysis (CCA), cross-modal factor analysis (CFA) and so on [6]. One of the widely used methods is canonical correlation analysis (CCA). CCA is a statistical approach which realizes linear dimensionality reduction and feature fusion by calculating maximally correlated linear projections [7]. Unlike CCA, cross-modal factor analysis (CFA) is a novel method for cross-model association. Besides noise removal, it provides a selection capability of reduced feature [8].

These methods transform high dimensional data to low dimensional features mostly depending on second order statistics, such as variance, correlation, mean square error, etc. The feature extraction is usually based on top eigenvalues and corresponding eigenvectors of certain matrices. For example, PCA uses the variance as the metric. However the foundation of the existing methods, second order statistics, is only optimal for Gaussian-like distribution. Therefore if a distribution differs greatly from Gaussian, the second order statistical tool is a poor estimator.

Moreover, most of the existing methods assume that there always exists linear relationship among the original data. However, in many situations non-linear feature extraction is necessary. In order to achieve non-linear transformation, kernel method is proposed as non-linear implementation, which leads to kernel PCA [9], kernel CCA [10] and kernel CFA [11]. However these methods still choose top eigenvalues and eigenvectors of the kernel matrix, and it does not reveal the nature of input data set.

In order to overcome this problem, we are motivated to apply kernel entropy component analysis (KECA) as an alternative [12]. Unlike the existing methods which depend on the second order statistics of the data set, kernel entropy component analysis (KECA) is based on information theory and preserves the maximum Renyi entropy of the input data with the smallest number of extracted features. From the aspect of classification, KECA enables nonlinear data analysis and captures the higher order statistics of the data. KECA does not correspond to the top eigenvalues and eigenvectors of the kernel matrix. On the other hand, the feature transformation and fusion based on KECA is achieved by largest contribution of entropy estimation.

In the following part, some information theoretical tools are introduced and an information-theoretically optimal method, kernel entropy component analysis, is described. Experimental results demonstrate that the performance of the proposed strategy is better than the existing methods, especially when the feature space has substantial dimension reduction.

4. Information Entropy and Information Theory

4.1. *Shannon entropy*

In physics, entropy is a way of estimation which correlates with the quantity of kinematic randomness. However in information theory, entropy is no longer a physical concept in thermodynamics.

The concept of information entropy was first introduced by Claude Shannon as a measure of statistical uncertainty which is widely used in communication theory [13]. It provided a mathematical framework to quantify and formulate the nature of information beyond physical laws and it was quickly accepted by science and engineering communities. Shannon states that a measure of the information amount contained in a series of events can be expressed by Shannon entropy. Shannon entropy is shown as

$$H_s(X) = -\sum_k p(x_k) \log p(x_k) \tag{1}$$

or

$$H_s(X) = -\int f_x(x) \log f_x(x) \tag{2}$$

where $p(x_k)$ and $f_x(x)$ are the discrete and continuous probability density function of data set respectively, and k is the total number of data set in the discrete case.

Shannon entropy is believed to be an effective metric to measure the uncertainty of random quantities in terms of probabilistic behavior of an information source. A fundamental property of entropy is its single scalar which measures the uncertainty in a form of probability density. Its robust and elegant properties make it more suitable to capture the characteristics of information. Moreover it can be interpreted as a means of quantifying information content. It can also be extended to measure dissimilarity between data.

4.2. *Renyi entropy*

Renyi entropy is one of the widely used generalizations of information entropy [14]. Renyi entropy of order α of a random variable X is expressed as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_1^N p_k^\alpha \right) \tag{3}$$

or

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\int f_x^\alpha(x) dx \right) \tag{4}$$

where $\alpha \geq 1$. Shannon entropy is a special case of Renyi entropy when α converges to one.

In this paper, Renyi quadratic entropy is employed, because it is more flexible than Shannon entropy. Renyi quadratic entropy is expressed as

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

where $p(x)$ is probability density function generated by the data set $\boldsymbol{D} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

The main reason for choosing $\alpha = 2$ is that the entropy value can be elegantly estimated by pdf $p(x)$ (probability density function) generated from the data set D and a computationally efficient entropy estimator can be realized. The entropy estimator is achieved by replacing the pdf with non-parametric density estimator called Parzen window density estimator [15].

4.3. *Kernel method*

Kernel method is widely used in the nonlinear problem of data analysis. The fundamental principle of kernel method is mapping the original data onto a feature space by a non-linear transformation and employing linear algorithms in the new space.

If the input space consists of $x_i \in \mathbf{R}_d$ in the set X, the non-linear mapping is expressed as follows.

$$\phi : \mathbf{R}_d \rightarrow F \tag{6}$$

$$x \to \phi(x) \tag{7}$$

where $F \in \mathbf{R}_l, l \geq d$.

Kernel function is a function k satisfying the following condition which is known as kernel trick.

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, \quad x_i, x_j \in X. \tag{8}$$

An explicit expression of the non-linear mapping ϕ is difficult to determinate. However kernel trick allows us obtain inner products in a feature space of possibly infinite dimensionality directly without calculating the explicit mapping ϕ. This lets us solve nonlinear problems by operating in a high dimensional feature space by linear machine learning algorithm expressed via inner products. However, the kernel function must satisfy the Mercers condition, i.e. positive semi-definite.

Some widely used kernel functions include linear kernel $k(x_i, x_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, polynomial kernel $k(x_i, x_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + 1)^d$, exponential kernel $k(x_i, x_j) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{2\sigma^2})$ and Gaussian kernel. The Gaussian kernel is defined as follows:

$$k_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right) \tag{9}$$

where σ is known as the kernel size.

4.4. *Parzen window density estimator*

Non-parametric density estimator is a strategy that estimates the probability distribution of a given data set without any assumptions of shapes or parameters. One of the non-parametric estimators is Parzen window which can be viewed as a kernel function and creates a close connection between information theory and kernel method [16]. Parzen window density estimator is given by

$$\tilde{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{10}$$

where $K(\cdot)$ is the kernel and h is a smoothing parameter called width. A non-parametric density estimator is achieved by replacing the actual pdf by its Parzen window estimator. Therefore, by using Parzen window method which is known to be consistent and efficient, non-parametric estimator for entropy does not require an explicit estimation of pdf.

5. Feature Transformation Based on Kernel Entropy Component Analysis

In kernel entropy component analysis, the entropy measure is Renyi quadratic entropy and the density estimation is realized by Parzen window density estimator [12].

The continuous Renyi quadratic entropy is expressed as

$$H(p) = -\log\left(\int p^2(x)dx\right) = -\log V(p) \tag{11}$$

where $V(p) = \int p^2(x)dx = E\{p(x)\}$. $V(p)$ can be viewed as expectation w.r.t. the density $p(x)$.

Because of the monotonicity properties of logarithmic function, we only need to consider the quantity $V(p) = \int p^2(x)dx$. In order to estimate $p(x)$ or $V(p)$, Parzen window density estimator is applied. Parzen window density estimator can be viewed as a kind of realization of the kernel trick, and it can be considered as a sum of inner products computed in the feature space. Then Parzen window density estimator based on the kernel notation can be rewritten as follows.

$$\tilde{p}(x) = \frac{1}{N\sigma}\sum_{x_i \in D} K\left(\frac{x - x_i}{\sigma}\right) = \frac{1}{N}\sum_{x_i \in D} k_\sigma(x, x_i) \tag{12}$$

where $k_\sigma(x, x_i)$ is the kernel of Parzen window density estimator centered at x_i and σ stands for the kernel size.

We assume a positive semi-definite Parzen kernel with Gaussian function $k_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2})$. The convolution theorem for Gaussian function states that the convolution of two Gaussian functions is another Gaussian function, with $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. Since $V(p)$ can be considered as $E\{p(x)\}$, we can get the following estimation [17].

$$\tilde{V}(p) = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_\sigma(x_i, x_j) = \frac{1}{N^2}\mathbf{1}^T\mathbf{K}\mathbf{1} \tag{13}$$

where element (i, j) of the $N \times N$ kernel matrix $\boldsymbol{K}$ is equal to $k(x_i, x_j)$, and $\mathbf{1}$ is a $N \times 1$ vector containing all ones. Therefore Renyi quadratic entropy is compactly expressed in terms of the kernel matrix.

Furthermore we can express Renyi entropy estimation in terms of eigenvalues and eigenvectors of the kernel matrix through eigen-decomposition. The eigen-decomposition of $\boldsymbol{K}$ can be shown as $\mathbf{K} = \mathbf{E}\mathbf{D}\mathbf{E}^T$, where $\boldsymbol{D}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \ldots, \lambda_N$ and $\boldsymbol{E}$ is a matrix with the corresponding eigenvectors $\boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_N$ as columns. Hence the empirical Renyi entropy estimation equals to the elements of the corresponding kernel matrix. The above expression can be rewritten to obtain the following results [12].

$$\tilde{V}(p) = \frac{1}{N^2}\mathbf{1}^T\mathbf{K}\mathbf{1} = \frac{1}{N^2}\mathbf{1}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{1} = \frac{1}{N^2}\sum_{i=1}^{N}(\sqrt{\lambda_i}\alpha_i^T\mathbf{1})^2 \tag{14}$$

where λ_i and $\boldsymbol{\alpha}_i$ are the i-th eigenvalue and eigenvector of Parzen window kernel matrix $\boldsymbol{K}$, and $\mathbf{1}$ is a $N \times 1$ vector of ones.

This expression is so-called entropy-values [18]. From the above expression, since each term $\sqrt{\lambda_i}\alpha_i^T$ contributes to the total entropy estimation, it is easily observed

that both eigenvalues and eigenvectors contribute to the entropy estimation. On the other hand, certain eigenvalues and the corresponding eigenvectors make more contribution to the total entropy estimation. The eigenvalues and eigenvectors are selected based on the largest entropy estimation instead of largest eigenvalues, which leads to different results from the existing methods, like kernel PCA, kernel CCA, etc. Therefore, kernel entropy component analysis is a feature transformation technique projecting original space onto a feature subspace spanned by the kernel principal axes corresponding to the largest contribution of Renyi entropy.

The algorithm procedure of kernel entropy component analysis is summarized as follows [19, 20]. First of all, the input data is the feature vector $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\}$, which needs feature transformation and fusion. As is mentioned above, we choose Gaussian function as kernel function and obtain the kernel matrix $\boldsymbol{K}$ with elements $K_{ij} = k(x_i, x_j)$. Then the eigen-decomposition of $\boldsymbol{K}$ is conducted. The next step is to choose the first n largest entropy estimation based on $\sqrt{\lambda_i}\alpha_i^T$; After that the kernel feature space data set $\phi_{eca} = \mathbf{D}^{\frac{1}{2}}\mathbf{E}^T$ can be computed, since we can conclude that $\phi_{eca}^T\phi_{eca} = (\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T)^T\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T = \mathbf{E}\mathbf{D}\mathbf{E}^T = \mathbf{K}$; Last we can complete feature transformation by ϕ_{eca}.

From the information-theoretic point of view, KECA preserves as much as possible of Renyi entropy between input space and kernel feature subspace with the smallest number of features, hence the information contents of input and transformed data are maximally similar. On the other hand, from the viewpoint of information fusion, KECA helps realize a semi-supervised fusion method which can reduce the dimensionality of the features by eliminating data redundancy and utilizing data complementarity in the form of entropy measures.

6. Audio Emotion Recognition Based on Kernel Entropy Component Analysis

6.1. *Audio emotion recognition*

Since the speech information is the most natural method of interaction between human and machine, the study of the emotion recognition based on audio information has drawn increasing attention recently. Audio emotion recognition is defined as extraction of the emotional state from a speaker's speech information. Many solutions have been proposed to process a spoken utterance and identify the emotional information. However, this task requires sufficient artificial intelligence to recognize the signal of human voices. Despite the considerable progress in speech processing, we are still far from building a successful natural speech emotion interaction system between man and machine [21].

Emotion recognition based on audio information is useful for applications of natural and friendly communication between humans and computers, such as intelligent human computer interaction, security and surveillance, online entertainment and education, etc. Some recent specific applications include intelligent

household robot for natural and friendly interaction with human beings [22] and fear type emotion recognition system dedicated to visual-audio surveillance [23].

Emotion recognition based on speech information is challenging due to the following aspects [24]. First of all, it is not very clear which speech feature is most suitable in efficiently characterizing different emotional states without depending on the lexical content or the speaker. A proper and efficient feature selection is believed to affect the recognition performance significantly. The second problem is added by acoustic variability, which is influenced by speaking styles and speaking rates of different speakers. The commonly used speech features such as pitch and energy are vulnerable to these shortcomings.

6.2. *Audio features for emotion recognition*

Since speech data are not always stationary, speech signal is commonly divided into frames in speech processing. The signal is considered to be stationary within every frame. The widely used speech features include continuous speech features and spectral speech features. Continuous speech features (or prosodic speech features) like pitch and energy are called local features as well and they are extracted from each frame of audio signal, while spectral speech features are called global features as well and calculated as statistics of all speech features [24].

Continuous speech feature has been widely used in emotion recognition based on audio signal. It is believed that continuous features like pitch and energy are the primary indicator of a speaker's emotion state and convey much of the emotional information. Because of temporal information present in speech information, continuous speech features are superior in terms of classification accuracy. By using the sufficient number of continuous feature vectors, complex classifiers such as support vector machine (SVM) and hidden Markov model (HMM) can be trained reliably and the model parameters can be accurately estimated [25].

Besides time-dependent continuous features, spectral features are chosen as a different representation for speech information. It has been shown that the spectral features based on cepstral analysis like MFCC (Mel-frequency cepstral coefficient) and LPCC (linear predictive cepstral coefficient) clearly outperform the linear based features like LPC (linear predictor coefficient). Spectral features have different representations of the nature of the information. Moreover, the number of spectral speech features is less, hence feature selection algorithms are executed faster, and the efficiency of classifier training is relatively high [26].

There has been no agreement on which features are more efficient and suitable for emotion recognition, but many researchers have claimed that local and global features are efficient in distinguishing different states of emotion. Since both continuous speech features and spectral speech features have their own advantages and limitations, it is believed that the integration of continuous and spectral features conveys more information about the human emotional state. If one modality fails to detect an emotion, the other modality is able to help to improve the performance. Therefore

the complementary relationship of these modalities leads to higher classification accuracy than what is achieved by a single feature [27].

6.3. *Information fusion of audio features based on KECA*

There is a wide investigation on emotion states which reveals that at least six basic emotions are universal, therefore this paper focuses on six principal emotions including happiness, surprise, sadness, fear, disgust and anger. Many solutions have been introduced for emotion recognition based on audio information. Some detailed review of the cutting-edge works can be found in [27−29].

The performance of emotion recognition based on audio fusion has been demonstrated by some works. However, it is far from an ultimate solution due to unsatisfactory accuracy and efficiency of the proposed solutions. Moreover, most works treat the audio features as independent information and have not built up a close relationship between them. In this paper, a new bimodal fusion solution for audio signal at feature level has been investigated. Figure 1 describes a block diagram of the proposed strategy.

First of all, a procedure of feature selection and fusion should be conducted to extract the significant features and reduce the dimensionality of the feature space. In order to reduce the noise, a wavelet coefficient threshold method is realized at the stage of pre-processing. Since leading and trailing edges do not have useful messages, they are then eliminated. Since audio analysis is reliable when the signal is stationary, a short time analysis can be performed within a short time interval of articulatory stability. Therefore the audio signal should be windowed into a succession of sequences which are also called frames. Before feature extraction, a Hamming window of size 512 points with 50% overlap between adjacent windows is multiplied on each speech frames.

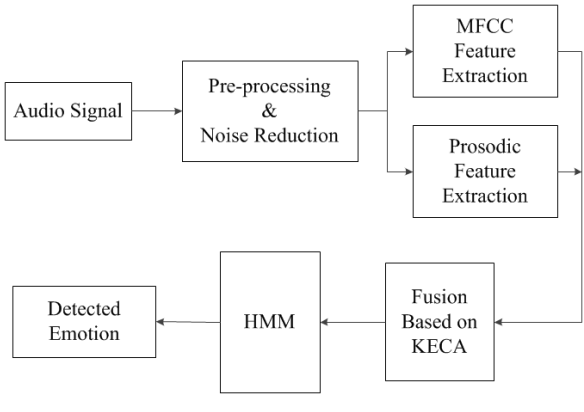

Fig. 1. Block diagram of feature level fusion for audio emotion recognition based on KECA.

Prosodic features are related to the rhythmic content of the audio signal. They are normally represented by the statistics of intensity, pitch, fundamental frequencies, formant frequencies, etc. In this paper, the statistics and variations of pitch, energy, pause length, speaking rate and formant frequencies are extracted as continuous features. Regarding spectral features, MFCC (Mel-frequency cepstral coefficient) is one of the most effective analytical tools in the area of speech recognition. It is widely used because it is physically related to the behavior of human ears by ceptral analysis. In this paper, the MFCC features are calculated based on speech frames. The first thirteen coefficients are used as useful features, because most of the information energy is stored in the first few coefficients. The statistics of each coefficient, such as median, mean, standard deviation, maximum value and minimum value, are combined to form a feature vector. After continuous and spectral analysis, the audio feature vector is obtained through the concatenated features of successive frames [30].

Before classification, dimensionality reduction and feature fusion should be performed, since the performance of classification largely relies on the discriminant ability of the transformed features. A large feature vector contains rich information about modalities, but it usually suffers from curse of dimensionality. In this issue, data points of a large feature vector become sparse, such that the finite set of sampling data may not adequately represent the underlining distributions for classification. In addition to this limitation, the complex classifiers do not work well if the input features have an extremely high dimensionality.

In order to alleviate these problems, the extracted audio features are analyzed using the proposed fusion approach based on kernel entropy component analysis (KECA) in this paper. The feature fusion strategy achieves two kinds of improvement. The first one is to integrate all audio features with maximally preserving the content of information and select a subset features which retain the original feature characteristics. The second is to transform the original space into a transformed space. Dimensionality reduction is achieved by generating a new feature vector based in transformed domain.

During the stage of classification, the resulting features are viewed as an input to a HMM (hidden Markov model) with a mixture of Gaussian densities. HMM is trained as the classifier because of its outstanding performance of modeling the temporal characteristics of audio signal. The output of HMM is the likelihood of a given data sequence which can be considered as the state of the detected emotion [31].

7. Experimental Results

In order to evaluate the effectiveness of the proposed strategy, extensive experiments have been conducted on RML emotion database [32] and eNTERFACE emotion database [33]. The RML emotion database contains 720 audiovisual emotional expression samples. There are eight human subjects speaking six languages. Six basic human emotions are expressed: anger, disgust, fear, sadness, surprise, and happiness.

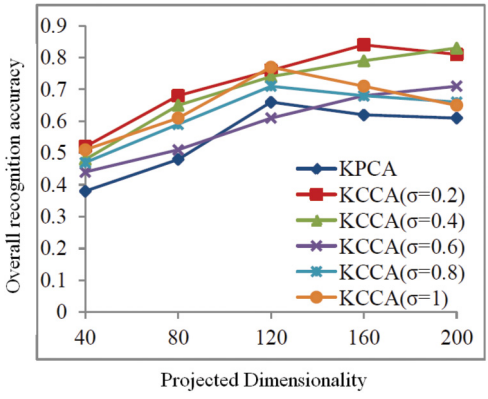

Fig. 2. Experimental results of RML emotion database based on KPCA and KCCA.

The samples are recorded at a sampling rate of 22,050 Hz using a single channel 16-bit digitization with a frame rate of 30 fps. The eNTERFACE emotion database contains 44 subjects coming from 14 different nationalities and expressing 6 basic emotions, with a sampling rate of 48,000 Hz and a frame rate of 25 fps. In the experiments, each audio sample has truncated to 2 second long and divided into 10 segments. The dimensionality of audio features is empirically set to 240. The evaluation procedure is based on cross-validation. For training, 75% of samples are selected randomly while the rest are for testing [32].

Figure 2 describes the comparison of overall recognition accuracy between KPCA and KCCA on RML database and Fig. 3 shows the experimental results of RML database based on KECA. Figure 4 shows the overall performance of KPCA and KCCA on eNTERFACE database and Fig. 5 displays the performance of KECA on eNTERFACE database. σ stands for kernel size in all figures. From these figures, it can be seen that KECA has better accuracy than KPCA and KCCA, and it

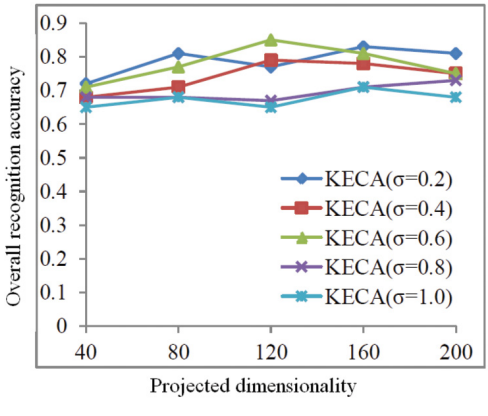

Fig. 3. Experimental results of RML emotion database based on KECA.

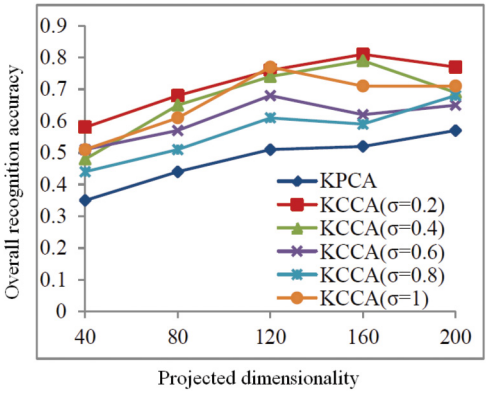

Fig. 4. Experimental results of eNTERFACE emotion database based on KPCA and KCCA.

outperforms KPCA and KCCA in dimensionality reduction and accuracy performance. The overall recognition accuracy of the proposed method based on KECA levels off even if the projected dimensionality is low. This shows that the extracted features at high dimension may contain redundant or noisy data. After processed by fusion method based on KECA, most of useful information is preserved and stable accuracy is achieved. On the other hand, the performance of KCCA and KPCA is largely degraded if the projected dimensionality decreases, which means that the fusion method is not properly selected and the degraded results are generated. These results demonstrate the noticeable improvement of dimensionality reduction and the ability of preserving useful information of the proposed solution.

Figure 6 is the confusion matrix of average performance based on KECA. Figure 7 shows the confusion matrix of average performance based on KPCA. Figure 8 demonstrates the result of the confusion matrix of average performance based on KCCA. All these three figures display the confusion matrix of average performance

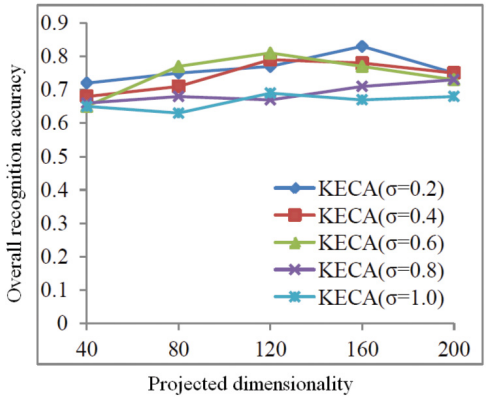

Fig. 5. Experimental results of eNTERFACE emotion database based on KECA.

Desired Emotion	Detected Emotion					
	Happiness	Disgust	Fear	Angry	Surprise	Sadness
Happiness	**79.58**	4.10	4.55	3.78	5.75	2.24
Disgust	6.41	**69.38**	5.89	6.42	5.57	6.33
Fear	6.64	5.14	**71.85**	4.78	6.83	4.76
Angry	4.78	2.54	2.56	**83.45**	4.32	2.35
Surprise	6.34	3.58	3.95	6.13	**75.14**	4.86
Sadness	7.69	6.39	5.29	6.53	5.43	**68.67**

Fig. 6. Confusion matrix of average performance on two databases based on KECA.

Desired Emotion	Detected Emotion					
	Happiness	Disgust	Fear	Angry	Surprise	Sadness
Happiness	**71.34**	8.53	2.60	5.35	4.85	7.33
Disgust	10.67	**57.74**	5.66	9.54	11.34	5.05
Fear	8.44	7.17	**61.91**	7.93	11.34	3.21
Angry	3.18	8.04	7.51	**72.94**	4.69	3.64
Surprise	5.34	4.05	5.81	4.75	**68.63**	11.42
Sadness	13.42	2.21	4.20	11.42	9.31	**59.44**

Fig. 7. Confusion matrix of average performance on two databases based on KPCA.

Desired Emotion	Detected Emotion					
	Happiness	Disgust	Fear	Angry	Surprise	Sadness
Happiness	**72.52**	4.34	6.77	3.91	3.25	9.21
Disgust	10.29	**60.85**	4.12	8.98	6.42	9.34
Fear	5.89	6.32	**63.54**	7.31	11.84	5.10
Angry	6.10	9.01	3.98	**69.11**	7.21	4.59
Surprise	11.85	2.67	3.12	10.53	**69.23**	2.60
Sadness	10.60	5.53	7.43	7.02	8.29	**61.13**

Fig. 8. Confusion matrix of average performance on two databases based on KCCA.

on RML emotion database and eNTERFACE emotion database. From these figures, it is obviously noted that the results are reasonably satisfactory and encouraging. Compared with the methods based on KPCA and KCCA, the strategy based on KECA has better accuracy and stability.

Figures 9 and 10 show the comparison of average recognition accuracy between fusion and non-fusion methods on RML emotion database and eNTERFACE emotion database. Figures 11 and 12 display the confusion matrix of fusion and non-fusion

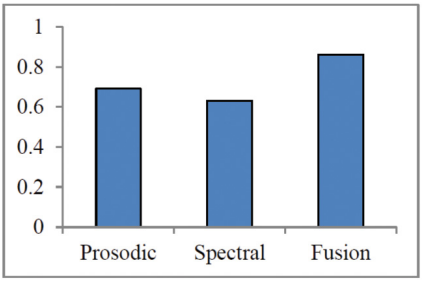

Fig. 9. Comparison between fusion result and non-fusion results on RML emotion database.

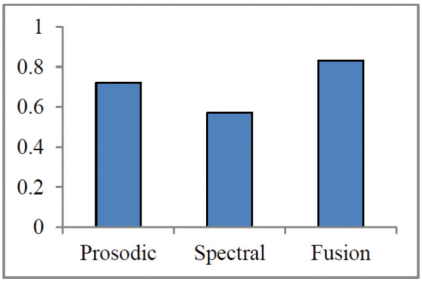

Fig. 10. Comparison between fusion result and non-fusion results on eNTERFACE emotion database.

Desired Emotion	Detected Emotion					
	Happiness	Disgust	Fear	Angry	Surprise	Sadness
Happiness	**82.41**	1.23	5.82	2.33	3.01	5.20
Disgust	6.21	**73.89**	5.09	4.77	5.63	4.41
Fear	7.32	4.74	**74.52**	3.41	6.99	3.02
Angry	2.13	2.38	2.88	**85.89**	3.40	3.32
Surprise	5.74	4.22	6.12	4.21	**76.01**	3.70
Sadness	6.40	6.91	3.28	5.90	3.17	**74.34**

Fig. 11. Confusion matrix of fusion result based on two emotion databases.

Desired Emotion	Detected Emotion					
	Happiness	Disgust	Fear	Angry	Surprise	Sadness
Happiness	**73.12**	0.98	7.06	6.74	9.54	2.56
Disgust	10.71	**70.78**	3.67	9.03	2.40	3.41
Fear	9.23	3.97	**65.21**	3.51	9.21	8.87
Angry	4.31	1.29	7.22	**73.90**	4.81	8.47
Surprise	3.90	11.95	5.73	4.83	**67.45**	6.14
Sadness	7.41	4.67	8.02	9.20	8.21	**62.49**

Fig. 12. Confusion matrix of non-fusion result based on two emotion databases.

results. It is observed that the fusion result based on KECA is superior to that obtained by non-fusion methods with single modality involved. This confirms the effectiveness of the fusion solution based on kernel entropy component analysis.

In summary, from the above experimental results, it is concluded that the overall performance of KECA outperforms the existing methods, like KPCA and KCCA. In addition, it provides best recognition accuracy and more efficient dimensionality reduction, and more stable recognition results.

8. Conclusion

This paper aims at multimodal information fusion issues in multimedia application. Several critical issues including, the challenges of fusion, the fusion levels, and the drawbacks of the existing fusion methods are presented. A novel information theoretic tool, kernel entropy component analysis (KECA), has been described. In this paper, KECA is applied into the application of audio emotion recognition. A new solution of feature level fusion for emotion recognition based on kernel entropy component analysis (KECA) has been introduced. The method presented in this paper shows good ability of dimensionality reduction without influencing much accuracy performance in multimedia application. Extensive experimental results demonstrate the feasibility of the proposed strategy.

References

[1] L. Guan, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki and M. Ibrahim, Multimodal information fusion for selected multimedia applications, *Int. J. Multimedia Intelligence and Security* **1**(1) (2010) 5−32.

[2] A. Ross and A. Jain, Information fusion in biometrics, *Pattern Recognition Letters, Special Issue on Multimodal Biometrics* **24**(3) (2003) 2115−2125.

[3] T. Joshi, S. Dey and D. Samanta, Multimodal biometrics: State of the art in fusion techniques, *Int. J. Biometrics* **1**(4) (2009) 393−417.

[4] P. Atrey, M. Hossain, A. El Saddik and M. Kankanhalli, Multimodal fusion for multimedia analysis: A survey, *Multimedia Systems* **16**(6) (2010) 345−379.

[5] S. Shivappa, M. Trivedi and B. Rao, Audiovisual information fusion in human computer interfaces and intelligent environments: A survey, *Proceedings of the IEEE* **98**(10) (2010) 1692−1715.
[6] M. Lazaridis, A. Axenopoulos, D. Rafailidis and P. Daras, Multimedia search and retrieval using multimodal annotation propagation and indexing techniques, *Signal Processing: Image Communication* **28**(4) (2013) 351−367.
[7] Y. Shin and C. Park, Analysis of correlation based dimension reduction methods, *International Journal of Applied Mathematics and Computer Science* **21**(3) (2011) 549−558.
[8] Y. Wang, L. Guan and A. Venetsanopoulos, Audiovisual emotion recognition via cross-modal association in kernel space, in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, 2011, pp. 1−6.
[9] B. Schölkopf, A. Smola and K. Müller, Kernel principal component analysis, *Artificial Neural Networks (CANN'97)*, 1997, pp. 583−588.
[10] X. Xu and Z. Mu, Feature fusion method based on kcca for ear and profile face based multimodal recognition, in *Proc. Int. Conf. on Automation and Logistics*, 2007, pp. 620−623.
[11] Y. Wang, L. Guan and A. Venetsanopoulos, Kernel crossmodal factor analysis for multimodal information fusion, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2384−2387.
[12] R. Jenssen, Kernel entropy component analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **32**(5) (2010) 847−860.
[13] C. Shannon, A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**(1) (2001) 3−55.
[14] A. Rrnyi, On measures of entropy and information, *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547−561.
[15] R. Jenssen, Information theoretic learning and kernel methods, *Information Theory and Statistical Learning*, 2009, pp. 209−230.
[16] E. Parzen, On estimation of a probability density function and mode, *The Annals of Mathematical Statistics* **33**(3) (1962) 1065−1076.
[17] R. Jenssen, T. Eltoft, M. Girolami and D. Erdogmus, Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm, *Advances in Neural Information Processing Systems* **19** (2007) 633−640.
[18] R. Jenssen, Kernel entropy component analysis: New theory and semi-supervised learning, in *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2011, pp. 1−6.
[19] L. Gómez-Chova, R. Jenssen and G. Camps-Valls, Kernel entropy component analysis in remote sensing data clustering, *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2011, pp. 3728−3731.
[20] F. He, M. Li and J. Yang, Adaptive clustering of production state based on kernel entropy component analysis, *International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1−8.
[21] L. Guan, Y. Wang and Y. Tie, Toward natural and efficient human computer interaction, in *Proc. Int. Conf. on Multimedia and Expo (ICME)*, 2009, pp. 1560−1561.
[22] X. Huahu, Y. Jian and G. Jue, Application of speech emotion recognition in intelligent household robot, in *Proc. Int. Conf. on Artificial Intelligence and Computational Intelligence (AICI)*, Vol. 1, 2010, pp. 537−541.
[23] C. Clavel, I. Vasilescu, L. Devillers, G. Richard and T. Ehrette, Fear-type emotion recognition for future audiobased surveillance systems, *Speech Communication* **50**(6) (2008) 487−503.

[24] M. El Ayadi, M. Kamel and F. Karray, Survey on speech emotion recognition: Features, classification schemes and databases, *Pattern Recognition* **44**(3) (2011) 572−587.
[25] Y. Lin and G. Wei, Speech emotion recognition based on hmm and svm, in *Proc. Int. Conf. on Machine Learning and Cybernetics*, Vol. 8, 2005, pp. 4898−4901.
[26] T. Nwe, S. Foo and L. De Silva, Speech emotion recognition using hidden markov models, *Speech Communication* **41**(4) (2003) 603−623.
[27] Z. Zeng, M. Pantic, G. Roisman and T. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **31**(1) (2009) 39−58.
[28] R. Cowie, E. Douglas-Cowie and N. Tsapatsoulis, Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine* **18**(1) (2011) 32−80.
[29] G. Potamianos, C. Neti, G. Gravier, A. Garg and A. Senior, Recent advances in the automatic recognition of audiovisual speech, *Proceedings of the IEEE* **91**(9) (2003) 1306−1326.
[30] Y. Wang and L. Guan, Recognizing human emotional state from audiovisual signals, *IEEE Trans. on Multimedia* **10**(5) (2008) 936−946.
[31] Y. Wang, R. Zhang, L. Guan and A. Venetsanopoulos, Kernel fusion of audio and visual information for emotion recognition, *Image Analysis and Recognition*, 2011, pp. 140−150.
[32] Y. Wang, L. Guan and A. Venetsanopoulos, Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition, *IEEE Trans. on Multimedia* **14**(3) (2012) 597−607.
[33] O. Martin, I. Kotsia, B. Macq and I. Pitas, The enterface'05 audio-visual emotion database, in *Proc. 22nd Int. Conf. on Data Engineering Workshops*, 2006, pp. 8.