

FF-SKPCCA: Kernel probabilistic canonical correlation analysis

Reza Rohani Sarvestani¹ · Reza Boostani¹

Published online: 5 September 2016
© Springer Science+Business Media New York 2016

Abstract Several information fusion methods are developed for increasing the recognition accuracy in multimodal systems. Canonical correlation analysis (CCA), cross-modal factor analysis (CFA) and their kernel versions are known as successful fusion techniques but they cannot digest the data variability. Probabilistic CCA (PCCA) is suggested as a linear fusion method to capture input variability. A new kernel PCCA (KPCCA) is proposed here to capture both the nonlinear correlations of sources and input variability. The functionality of KPCCA decreases when the number of samples, which determines the size of kernel matrix increases. In the conventional fusion methods the latent variables of different modalities are concatenated; consequently, a large-scale covariance matrix with just limited number of samples must be estimated. To overcome this drawback, a sparse KPCCA (SKPCCA) is introduced which sacrifices the covariance matrix elements at the cost of decreasing its rank. In the final stage of the gradual evolution of KPCCA, a new feature fusion manner is proposed for SKPCCA (FF-SKPCCA) as a second stage fusion. This proposed method unifies the latent variables of two modalities into a feature vector with an acceptable size. Audio-visual databases like M2VTS (for speech recognition) eNTERFACE and RML (for emotion recognition) are applied to assess FF-SKPCCA compared to

state-of-the-art fusion methods. The comparative results indicate the superiority of the proposed method in most cases.

Keywords Feature fusion · Canonical correlation analysis (CCA) · Probabilistic CCA · Kernel CCA

1 Introduction

Single modal recognition systems are not always promising due to partial observability and data variability [3]. To compensate for this drawback, high performance recognition systems in a simultaneous manner collect data from different sources (e.g. audio and video) since none of them can solely characterize all input states [2, 5]. The main issue in multimodal systems is to develop an efficient fusion method to fuse the features from different modalities in order to enrich a discriminative feature set [1, 2, 4]. Fusion techniques are applied to several bimodal recognition systems like automatic audio-visual speech recognition [5], human computer interaction [6, 49], biometric systems [7, 51], video indexing [8], object tracking [9] and emotion recognition [10–12, 48].

Information fusion can take place at the decision [18] or feature levels [28]. At the decision level, elicited features of each modality are applied to a specific classifier and the final result is obtained through classifiers' decision fusion. The well-known methods at this level are single winner, majority voting, Bayesian and Dempster-Shafer [29]. Nevertheless, decision fusion cannot take the correlation of modalities into account which reduces the performance of this approach compared to the feature fusion approach in a significant manner.

✉ Reza Boostani
boostani@shirazu.ac.ir
Reza Rohani Sarvestani
rrohani@cse.shirazu.ac.ir

¹ CSE & IT Department, Electrical and Computer Engineering Faculty, Shiraz University, Shiraz, Iran

Though feature fusion methods are a bit complicated, they have the advantage of unveiling the implicit linear/nonlinear correlation among the sources. The feature fusion procedure is similar to that of the human brain, since data from different sensors (e.g. ears, eyes) are fused in the brain to draw an accurate decision. Since the neurons' function is nonlinear, the brain fusion takes the benefit of nonlinear mapping. Observing the ability of the brain fusion encourages the research teams to develop mathematical methods for data fusion of different sources. In early feature fusion methods, features of different modalities were arranged in a high dimensional feature vector. In fact, no conceptual fusion was made except putting features beside each other and constructing a long vector. After feature fusion, these high dimensional vectors were fed to a classifier in order to assign a label to each vector.

The number of estimated parameters in some classification schemes like neural networks or Bayesian based classifiers depends on the feature size, that is, there exist, a direct relation between an increase in the number of features, and their computational complexity. In parametric classifiers the number of which depend on the input size have a serious problem in handling the datasets which contain low number of high dimensional samples. This problem is named small sample size (SSS) problem, and to solve it, a great volume of training data is necessary for a valid learning. One of the controversial issues in the multimodal recognition systems is the existence of statistical correlation among the features of different modalities. In one sense independency between modalities can diminish the redundancy and improve the recognition performance, and in the other, this correlation should not be observed as a negative factor, because it sometimes leads to a better data analysis, denoising and enhancement [12, 21, 24]. Besides the discussion of pros and cons of this correlation, it is better to estimate the statistical correlation among the features which are elicited from different sources. If this correlation is in a measured accurate manner the redundant features can be estimated; otherwise, even strong learners cannot consider all interactions among great number of features, hence a possible decrease in classifiers' performance. For instance, during a certain activity in audio-visual recognition systems facial expression and speech signal affect each other, which lead to the availability of a degree of correlation [2, 5, 12, 13, 50].

Canonical correlation analysis (CCA) [14] is a well-known fusion method that elicits common features (latent variables) between two sets of feature vectors. In CCA, these sets are projected into a new domain, named correlation space. By maximizing the sets cross-correlation in this space, latent variables of two modalities are extracted to be applied to a classifier.

Cross-modal factor analysis (CFA) [15] is another linear statistical method for extracting latent features from two sources. CFA adopts a criterion which minimizes the Frobenius norm between two data streams in the transformed domain.

CCA and CFA are adopted in practical recognition systems such as face detection during talking [16], biometric [17, 18], audio-visual speaker detection [19] medical imaging [7, 47] and audio-visual synchronization [20].

To enable conventional fusion methods capture linear/nonlinear dependencies, the nonlinear kernels are applied to project the features into the kernel space and then fuse them together. For this purpose, the kernel CCA (KCCA) [21] and kernel CFA (KCFA) [12] are introduced as the nonlinear versions of CCA and CFA respectively. They have been and are being applied in various data fusion applications like specific radar emitter identification [22], audiovisual biometric aliveness checking [23] and audiovisual emotion recognition [12]. Real applications involved in a high degree of variability cannot obtain promising results through the deterministic fusion methods, since during audio/video data recording some artificial and natural variations are inevitable. For instance, the authors in [12] applied KCCA to elicit the latent variables of audio and video modalities but their findings were not convincing.

To overcome the data variability, the authors in [24] proposed the probabilistic CCA (PCCA) method in order to digest this variability. Since PCCA is a linear method, it finds a linear projection, to map the data in the correlation space on which the variance is maximized [25–27].

Attempt is made here to propose the kernel version of PCCA (KPCCA) to capture both the dependencies and data variability in nonlinear correlation space. Although applying a nonlinear kernel increases the ability of fusion methods, the dimension of kernel matrix directly depends on the number of samples, that is provided that the number samples are low, the estimating of covariance matrix will be encountered by SSS problem; on the contrary provided that the number of samples are high, the dimension of kernel matrix increase.

The gradual evolution of KPCCA is assessed here by developing a sparse version of KPCCA (SKPCCA) which scarifies the elements of covariance matrix in order to avoid SSS problem. Here, SKPCCA solves the KPCCA problem at the cost of diminishing the rank of covariance matrix by a well-known sparse technique. The computational burden of SKPCCA is remarkably decreased compared to that of KPCCA.

In the final stage, a feature fusion method upon the framework of SKPCCA is proposed which seeks to fuse the latent variables of different modalities instead of just concatenate them together in a long feature vector in a conceptual sense.

In this proposed method, the latent variables of the two modalities are unified into one vector with a size equal to the size of latent variable of the corresponding modality with numerous samples.

The remainder of this paper is organized as follows: the methods of identifying latent variables in the correlation space are explained in Section 2; the main proposed method with its major continuance component are introduced in Section 3; the applied datasets and the feature extraction techniques are discussed in Section 4; the empirical results and their comparisons with the state-of-the-art methods are presented in Section 5 and the article is concluded in Section 6.

2 Background methods

In a multimodal recognition system, instead of analyzing the data of each modality alone, the main concern is estimating a joint correlation subspace where the features (latent variables) can be extracted. There exist several fusion techniques for constructing this subspace; in this study only a few will be of concern.

2.1 CCA and KCCA methods

CCA [14] is known as the most famous statistical fusion method which finds a linear map for projecting features of two modalities in the correlation space in a manner that their cross-correlation is maximized. Given two feature sets of \mathbf{x} and \mathbf{y} , CCA estimates two linear projections \mathbf{W}_x and \mathbf{W}_y , respectively. These two projecting are named canonical correlation matrices which are determined by maximizing their cross-correlation in a manner that \mathbf{W}_x and \mathbf{W}_y become diagonal (compact) in the projected space.

To capture nonlinear correlations between two sources, CCA is equipped with kernel (KCCA) [21]. The optimization criterion of which is described as follows:

$$\max_{\alpha, \beta} \frac{\alpha^T \mathbf{K}_x \mathbf{K}_y \beta}{\sqrt{[\alpha^T \mathbf{K}_x \mathbf{K}_x \alpha] [\beta^T \mathbf{K}_y \mathbf{K}_y \beta]}} \tag{1}$$

where \mathbf{K}_x and \mathbf{K}_y are the kernel matrices of x and y , respectively.

The KCCA criterion is optimized through the generalized eigen-value decomposition method. When kernel functions are non-invertible, conventional regularization techniques are applied to convert (1) into (2) [30]:

$$\max_{\alpha, \beta} \frac{\alpha^T \mathbf{K}_x \mathbf{K}_y \beta}{\sqrt{[\alpha^T (\mathbf{K}_x^2 + \tau \mathbf{K}_x) \alpha] [\beta^T (\mathbf{K}_y^2 + \tau \mathbf{K}_y) \beta]}} \tag{2}$$

where $0 \leq \tau \leq 1$. The computation details of CCA and KCCA are expressed in Appendix A.

2.2 CFA and KCFA methods

The cross-modal factor analysis (CFA) technique is proposed by [15]. In this technique, features are extracted from two modalities followed by determining two separate linear maps for projecting their features into the cross-modal space. The main difference between CFA and CCA is in their objective functions. CFA minimizes the Frobenius norm between the projected features while CCA maximizes their correlation.

CFA is capable of detecting linear relations between two feature sets while it suffers from the lack of capturing non-linear correlations between them. Similar to KCCA, the kernelized version of CFA is introduced as KCFA by [12] in order to enhance its detection ability. In a similar manner, before applying CFA, the feature sets of two modalities are implicitly projected to the high dimensional kernel space, where the standard CFA is applied. To explain the KCFA routine the projected feature vectors in the cross-modal associated domain are computed as:

$$\begin{aligned} & \frac{1}{\sqrt{\alpha_j^T \mathbf{K}_x \alpha_j}} \alpha_j^T \begin{bmatrix} K(x', x_1) \\ K(x', x_2) \\ \dots \\ K(x', x_n) \end{bmatrix} \\ & \frac{1}{\sqrt{\beta_i^T \mathbf{K}_y \beta_i}} \beta_i^T \begin{bmatrix} K(y', y_1) \\ K(y', y_2) \\ \dots \\ K(y', y_n) \end{bmatrix} \end{aligned} \tag{3}$$

where α_j and β_i are the eigenvectors of $\mathbf{K}_y \mathbf{K}_x$ and $\mathbf{K}_x \mathbf{K}_y$, respectively. Further details are expressed in Appendix A

The main drawback of CCA, KCCA, CFA and KCFA is the absence of considering the uncertainty of input observations of \mathbf{x} and \mathbf{y} since no deterministic correlation space can be found for two sets of uncertain features.

2.3 Probabilistic CCA

To deal with the uncertainty problem of CCA, a probabilistic version of CCA (PCCA) is proposed by [24] where the projected latent variables provide maximum variance in the joint correlation space (Fig. 1). To insert this uncertainty parameter, they assigned a specific Gaussian function for each single source of data, which is expressed as:

$$\begin{aligned} z & \sim \mathcal{N}(0, I_d) \quad 1 \leq d \leq \min(p, q) \\ \mathbf{x}|z & = \mathcal{N}(z \mathbf{W}_x^T + \mu_x, \varphi_x) \\ \mathbf{y}|z & = \mathcal{N}(z \mathbf{W}_y^T + \mu_y, \varphi_y) \end{aligned} \tag{4}$$

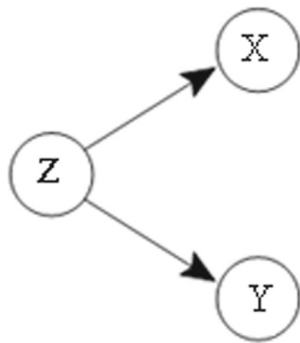


Fig. 1 Graphical model for showing the relations of two modalities and its latent variable z [24]

where, z is the latent variable of the two modalities of \mathbf{x} and \mathbf{y} and μ and φ are the mean and covariance of each modality, respectively. They revealed that the posterior expectation of z given \mathbf{x} and \mathbf{y} is determined as follows [24]:

$$\begin{aligned} \varphi_x &= C_{xx} - \mathbf{W}_x \mathbf{W}_x^T \\ \varphi_y &= C_{yy} - \mathbf{W}_y \mathbf{W}_y^T \\ E(z|\mathbf{x}) &= \mathbf{x} \mathbf{W}_x M_x^{-1}, M_x = I + \mathbf{W}_x^T \varphi_x^{-1} \mathbf{W}_x \\ E(z|\mathbf{y}) &= \mathbf{y} \mathbf{W}_y M_y^{-1}, M_y = I + \mathbf{W}_y^T \varphi_y^{-1} \mathbf{W}_y \end{aligned} \tag{5}$$

where \mathbf{W}_x and \mathbf{W}_y are the early d canonical directions of \mathbf{x} and \mathbf{y} and C_{xx} and C_{yy} are the covariance matrices of \mathbf{x} and \mathbf{y} respectively.

Another solution for (4) can be reached by applying the expectation maximization (EM) scheme [24]. This method provides a general solution for PCCA scheme which yields the following updated described in (6):

$$\begin{aligned} \mathbf{W}_{t+1} &= C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} (M_t^{-1} + M_t^{-1} \mathbf{W}_t^T \varphi_t^{-1} C \varphi_t^{-1} \mathbf{W}_t M_t^{-1})^{-1} \\ \varphi_{t+1} &= \begin{pmatrix} (C - C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \mathbf{W}_t^T)_{11} & 0 \\ 0 & (C - C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \mathbf{W}_t^T)_{22} \end{pmatrix} \end{aligned} \tag{6}$$

where $M_t = I + \mathbf{W}_t^T \varphi_t^{-1} \mathbf{W}_t$.

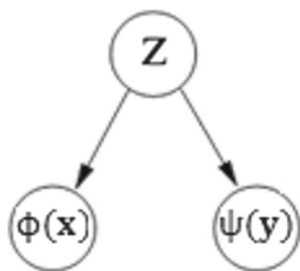


Fig. 2 Graphical representation of the proposed method

3 Proposed method

The proposed method and its constituent components are presented in detail following a gradual evolution inflicted on PCCA.

3.1 The kernel PCCA (KPCCA)

Although PCCA performs well in finding linear correlations between two sets of \mathbf{x} and \mathbf{y} contaminated by the variability factors, its performance decreases when their correlation is nonlinear. To solve this problem, first the features are passed through nonlinear kernels of $(\phi(\mathbf{x})$ and $\psi(\mathbf{y}))$, next, PCCA is applied to them in order to elicit the latent variables [31] (Fig. 2).

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \\ \phi(\mathbf{x})|z &= \mathcal{N}(z \mathbf{W}_x^T + \mu_x, \varphi_x) \\ \psi(\mathbf{y})|z &= \mathcal{N}(z \mathbf{W}_y^T + \mu_y, \varphi_y) \end{aligned} \tag{7}$$

In KPCCA, similar to KCCA, \mathbf{W}_x and \mathbf{W}_y are determined through $\mathbf{W}_x = \phi(\mathbf{x})^T \alpha$ and $\mathbf{W}_y = \psi(\mathbf{y})^T \beta$ followed by implementing the procedure in Bach and Jordan method [4] to obtain the parameters. By inserting \mathbf{W}_x and \mathbf{W}_y into (5), the posterior expectation of parameters is estimated as:

$$\begin{aligned} \varphi_x &= \phi(\mathbf{x})^T (I - \alpha \alpha^T) \phi(\mathbf{x}) = \phi(\mathbf{x})^T \Omega_x \phi(\mathbf{x}) \\ \varphi_y &= \psi(\mathbf{y})^T (I - \beta \beta^T) \psi(\mathbf{y}) = \psi(\mathbf{y})^T \Omega_y \psi(\mathbf{y}) \\ E(z|\mathbf{x}) &= \phi(\mathbf{x}) \cdot [\phi(\mathbf{x})^T \alpha] M_x^{-1} = \mathbf{K}_x \alpha M_x^{-1}, \\ M_x &= I + \alpha^T (I - \alpha \alpha^T)^{-1} \alpha = I + \alpha^T (\Omega_x)^{-1} \alpha \\ E(z|\mathbf{y}) &= \psi(\mathbf{y}) \cdot [\psi(\mathbf{y})^T \beta] M_y^{-1} = \mathbf{K}_y \beta M_y^{-1}, \\ M_y &= I + \beta^T (I - \beta \beta^T)^{-1} \beta = I + \beta^T (\Omega_y)^{-1} \beta \end{aligned} \tag{8}$$

where $\Omega_x = I - \alpha \alpha^T$ and $\Omega_y = I - \beta \beta^T$.

The general solution to estimate these transformation matrices is achieved by applying EM algorithm in (6). By inserting $\mathbf{W}_x = \phi(\mathbf{x})^T \alpha$ and $\mathbf{W}_y = \psi(\mathbf{y})^T \beta$ into (6), and by defining $\varpi(\mathbf{O}) = \begin{bmatrix} \phi(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \psi(\mathbf{y}) \end{bmatrix}$ $\gamma = [\alpha \beta]^T$ and $\varphi = \varpi(\mathbf{O})^T \Omega \varpi(\mathbf{O})$ (where $\Omega = \begin{bmatrix} \Omega_x & \mathbf{0} \\ \mathbf{0} & \Omega_y \end{bmatrix}$) and by assuming that \mathbf{K}_x and \mathbf{K}_y are invertible, (6) is yield:

$$\begin{aligned} \varpi(\mathbf{O})^T \gamma_{t+1} &= \varpi(\mathbf{O})^T [\Omega_t^{-1} \gamma_t M_t^{-1} (M_t^{-1} + M_t^{-1} \gamma_t^T \Omega_t^{-1} \Omega_t^{-1} \gamma_t M_t^{-1})^{-1}] \\ \varphi_{t+1} &= \varpi(\mathbf{O})^T \Omega_{t+1} \varpi(\mathbf{O}) \\ &= \varpi(\mathbf{O})^T \begin{pmatrix} (I - \Omega_t^{-1} \gamma_t M_t^{-1} \gamma_t^T)_{11} & 0 \\ 0 & (I - \Omega_t^{-1} \gamma_t M_t^{-1} \gamma_t^T)_{22} \end{pmatrix} \\ &\quad \times \varpi(\mathbf{O}) \end{aligned} \tag{9}$$

consequently, (9) can be modified into (10) as follow:

$$\gamma_{t+1} = \Omega_t^{-1} \gamma_t M_t^{-1} \left(M_t^{-1} + M_t^{-1} \gamma_t^T \Omega_t^{-1} \Omega_t^{-1} \gamma_t M_t^{-1} \right)^{-1}$$

$$\Omega_{t+1} = \begin{pmatrix} \left(I - \Omega_t^{-1} \gamma_t M_t^{-1} \gamma_{t+1}^T \right)_{11} & 0 \\ 0 & \left(I - \Omega_t^{-1} \gamma_t M_t^{-1} \gamma_{t+1}^T \right)_{22} \end{pmatrix} \quad (10)$$

The above equations illustrate the learning procedure of transformation matrices (α and β) applied for obtaining latent variables of both the modalities. It is obvious that if \mathbf{K}_x and \mathbf{K}_y are not invertible and the above equations

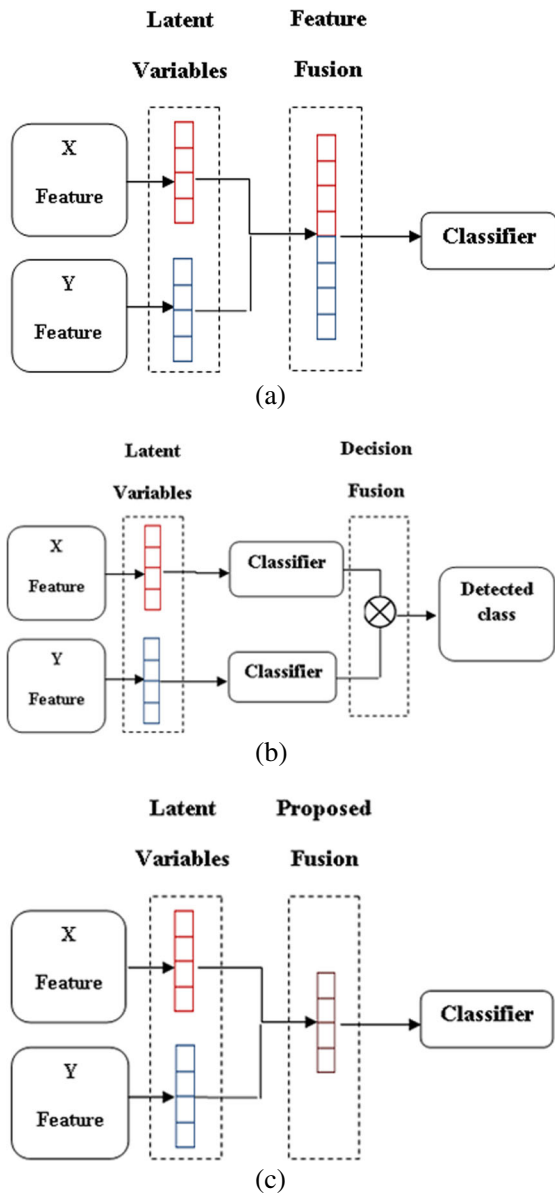


Fig. 3 Fusion level approaches (a) Feature fusion (b) Decision fusion (c) Proposed fusion method

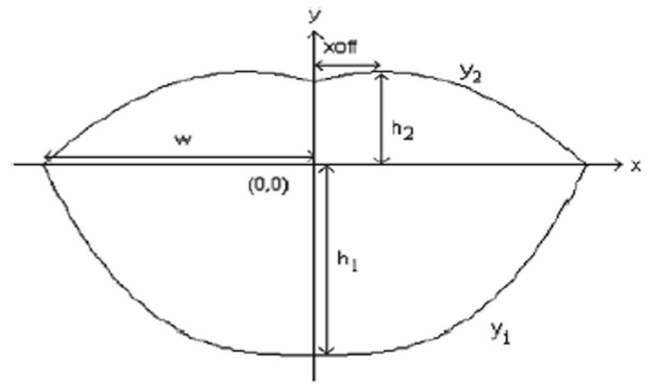


Fig. 4 Geometric lip model

would not be solved. An alternative solution is the regularization approach similar to KCCA [30]. In methods where a prior knowledge on $\mathbf{W} \sim N(0, r^{-1}I_s)$ is considered, the regularization parameter r is determined by applying EM algorithm [27, 32]. Applying the described regularization method yield the following iterative solution for γ :

$$\gamma_{t+1} = \Omega_t^{-1} \gamma_t M_t^{-1} \left(M_t^{-1} + M_t^{-1} \gamma_t^T \Omega_t^{-1} \Omega_t^{-1} \gamma_t M_t^{-1} + r \lambda_\phi \lambda_K \right)^{-1} \quad (11)$$

where $\lambda_\phi = \text{trac}(\phi)$ and $\lambda_K = \text{trac}(\mathbf{K}_x) + \text{trac}(\mathbf{K}_y)$.

3.2 The sparse KPCCA (SKPCCA)

In kernel based methods, the dimension of projected features in the kernel space (\mathbf{K}_x and \mathbf{K}_y) is related to the number of data samples. When the dimension size is low, the full-covariance matrices of Ω_x and Ω_y can be estimated properly; otherwise, the covariance matrices cannot be estimated accurately due to limited number of samples. The innovative manner to solve this problem, where additional latent variables z_x and z_y are avoided in constructing high dimensional inputs are introduced by [25, 26]:

$$z_x z_y z \sim \mathcal{N}(0, I_d)$$

$$\phi(\mathbf{x}) | z z_x = \mathcal{N}(z \mathbf{W}_x^T + z_x \mathbf{V}_x^T + \mu_x, \sigma_x^2 I)$$

$$\psi(\mathbf{y}) | z z_y = \mathcal{N}(z \mathbf{W}_y^T + z_y \mathbf{V}_y^T + \mu_y, \sigma_y^2 I) \quad (12)$$

where, σ_i^2 is the variance parameter and \mathbf{V}_x and \mathbf{V}_y are the two extra projection matrices for individual sources

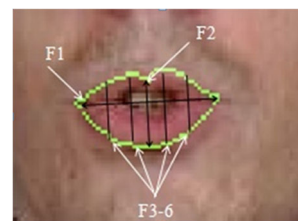


Fig. 5 Lip feature extraction

of \mathbf{x} and \mathbf{y} respectively. Here, by marginalizing z_x and z_y , a model similar to (11) is developed which generates the following full-rank covariance matrix: $\Omega_i = v_i v_i^T + \sigma_i^2 I (\mathbf{V} = \phi(i)^T v)$. This model is able to solve the SSS problem [26] and for calculating the covariance matrices of Ω_x and Ω_y , it leads to a better estimation of the transform matrix γ . The pseudo code of SKPCCA is illustrated in Algorithm 1.

Algorithm 1 Sparse Kernel PCCA(SKPCCA)

- 1) Assume that v_x and v_y are fixed and marginalized over z_x and z_y to get $\Omega_x = v_x v_x^T + \sigma_x^2 I$ and $\Omega_y = v_y v_y^T + \sigma_y^2 I$
 - Update the parameters $\gamma = [\alpha \ \beta]$ according to (11)
- 2) Marginalize over z to get $\Omega_x = \alpha \alpha^T + \sigma_x^2 I$
 - Update the parameter v_x according to (11) and by replacing γ with v_x
- 3) Update σ_x^2 using

$$\sigma_x^2 = \frac{1}{d_x} \text{trace}(I - A_x^T v_x^T - \alpha \alpha^T)$$
 where d_x is the dimension of x , $A_x = M_x^{-1} v_x^T \Omega_x^{-1}$ and v_x is the new updated value.
 - Repeat the above two sub-steps for the parameters related to y , replacing all subscripts x with y and α to β .

3.3 The new feature fusion method

Conventional feature fusion methods concatenate latent variables of two sources (Fig. 3a) and construct high dimensional feature vectors in the correlation space. These large-scale vectors are applied to a classifier. This procedure has two drawbacks: availability of redundancy between the modalities and not coping with large-scale feature vectors.

To solve this dimensionality problem, a new feature fusion method is proposed which mathematically fuses latent variables of both the modalities into a unified low-dimensional feature vector (Fig. 3c).

To explain this method in a mathematically manner, conditional expectation of unified latent variables \mathbf{x} and \mathbf{y} is optimized according to [24]. By considering $\mathbf{W}_x = \phi(\mathbf{x})^T \alpha$, and $\mathbf{W}_y = \psi(\mathbf{y})^T \beta$, (13) is obtained:

$$E(z|\mathbf{x}, \mathbf{y}) = [\mathbf{K}_x \alpha \ \mathbf{K}_y \beta] \begin{bmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{bmatrix} \times \begin{bmatrix} M_x^{-1} \\ M_y^{-1} \end{bmatrix} \tag{13}$$

where, α and β are the two matrices applied in high-dimensional concatenated features of \mathbf{x} and \mathbf{y} , in order to unify them at z , where, $P_d = M_x^{-1} * (M_y^{-1})^T$. A low-dimensional feature vector is obtained as the input classifier. The schematic of the above mentioned gradual evolution is illustrated in (Fig. 3c).

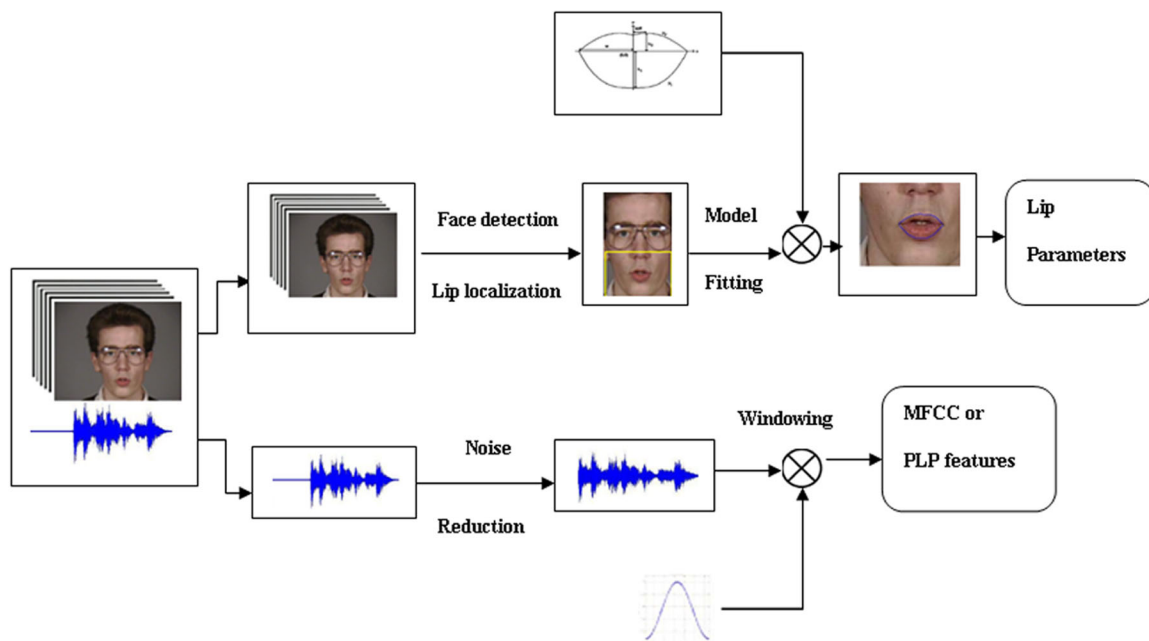


Fig. 6 Audio-Video feature extraction process for speech recognition

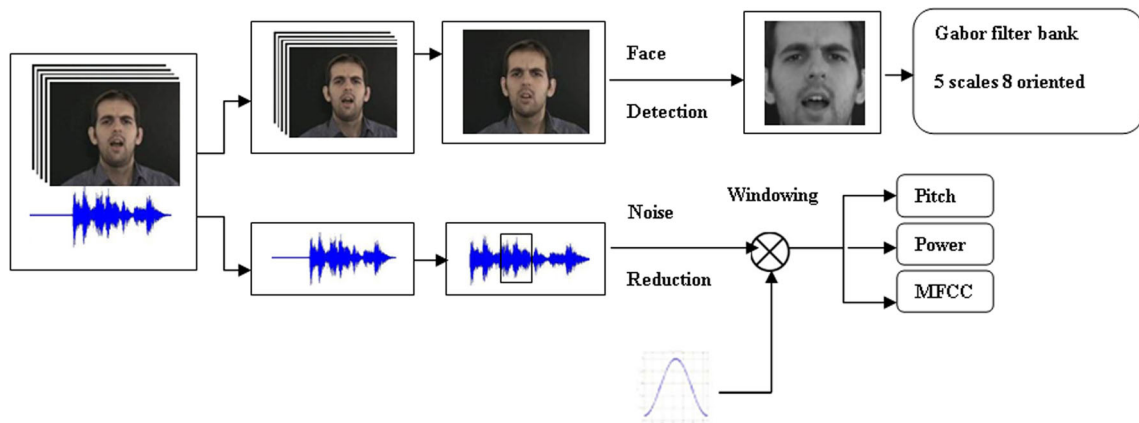


Fig. 7 Audio-Video feature extraction process for emotion recognition

4 Real applications

Research findings in this field indicate that the fusion of information from acoustic and facial expression modalities

improve the performance of both speech and emotion recognition systems [4, 5, 12]. Acoustic signals or video frames cannot provide a promising result on their own since each contains partial information on the subject during speak-

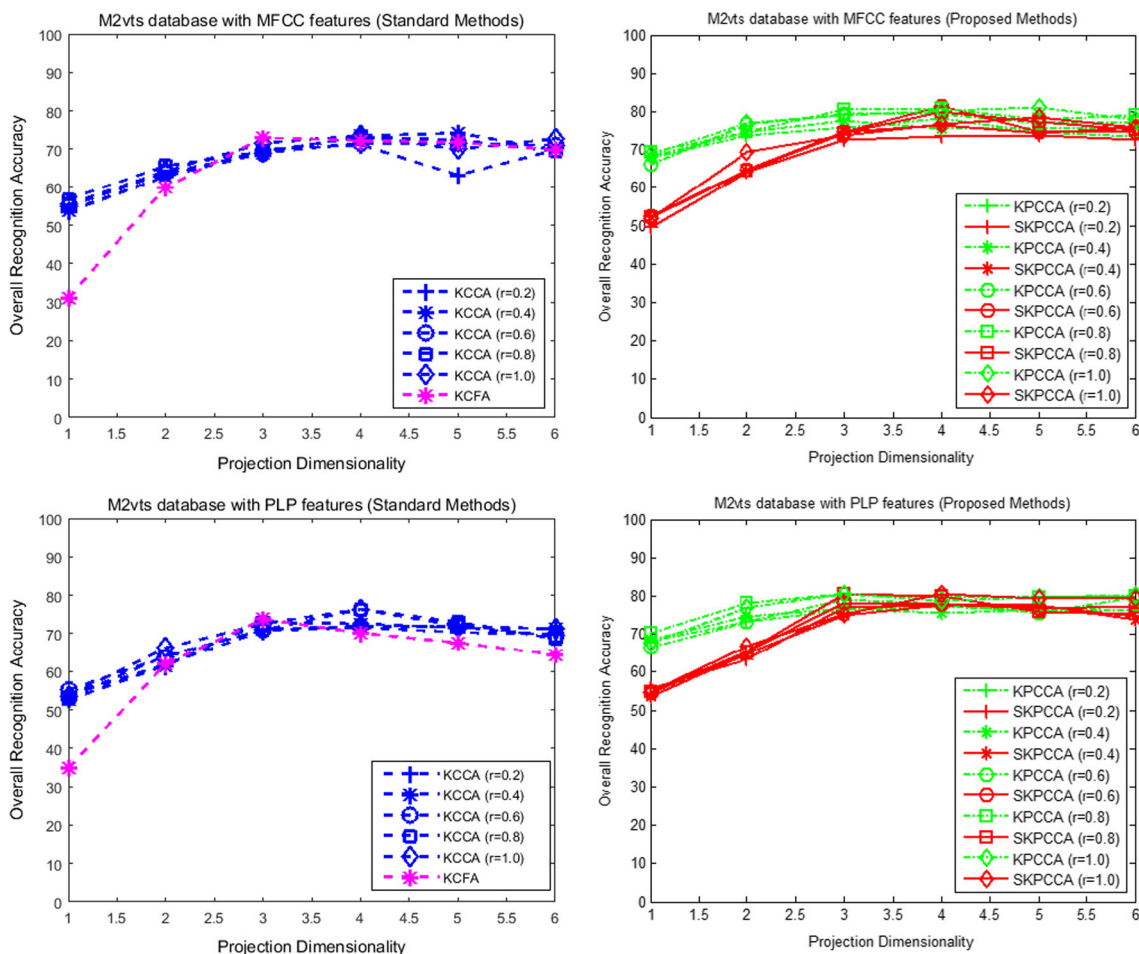


Fig. 8 Experimental results of Feature Fusion method. Top Left: M2vts database with MFCC features on KCCA and KCFA methods. Top Right: M2vts database with MFCC features on KPCCA and

SKPCCA methods. Bottom Left: M2vts database with PLP features on KCCA and KCFA methods. Bottom Right: M2vts database with PLP features on KPCCA and SKPCCA methods

ing. The variability factors affect the recording quality (e.g. noise, head pose etc.) reduce the recognition accuracy.

4.1 Speech recognition

The best descriptive audio signal feature is Cepstral coefficients [33], categorized as nonlinear features, elicited in short windows and represent the vocal tract state. Variants of the Cepstrum features include the popular Mel-Warped Cepstrum or Mel-frequency Cepstral Coefficients (MFCCs) [28] and the perceptual linear predictor (PLP) [34]. To extract the audio features, the additive background noise is eliminated [35] and the clean signals are divided into successive windows with the lengths are named as time frames. A windowing function, like Hamming, is usually applied to ongoing speech signals and the first 12 MFCC coefficients are extracted from each windowed signal.

In extracting informative features from lip motion of a subject, a proper detection method should be adopted in

order to estimate lip contour in successive video frames. To detect lip contour, [33] proposed a method which partitions a given face image (colored image) into lip and non-lip regions based on the intensity and color features. By applying spatial fuzzy C-mean clustering to these features and simulating a simple geometric lip model, the lip contour reveals. The geometric lip model described by (14) is presented in Fig. 4.

$$y_1 = h_1 \left(\left(\frac{x - sy_1}{w} \right)^2 \right)^{1+\delta^2} - h_1,$$

$$y_2 = \frac{-h_2}{(w - x_{off})^2} (|x - sy_2| - x_{off})^2 + h_2 \tag{14}$$

where $x \in [-w, w]$ at $(0,0)$.

After fitting the lip model on each image, a lip contour is characterized with six parameters (features), which are applied in successive frames, Fig 5.

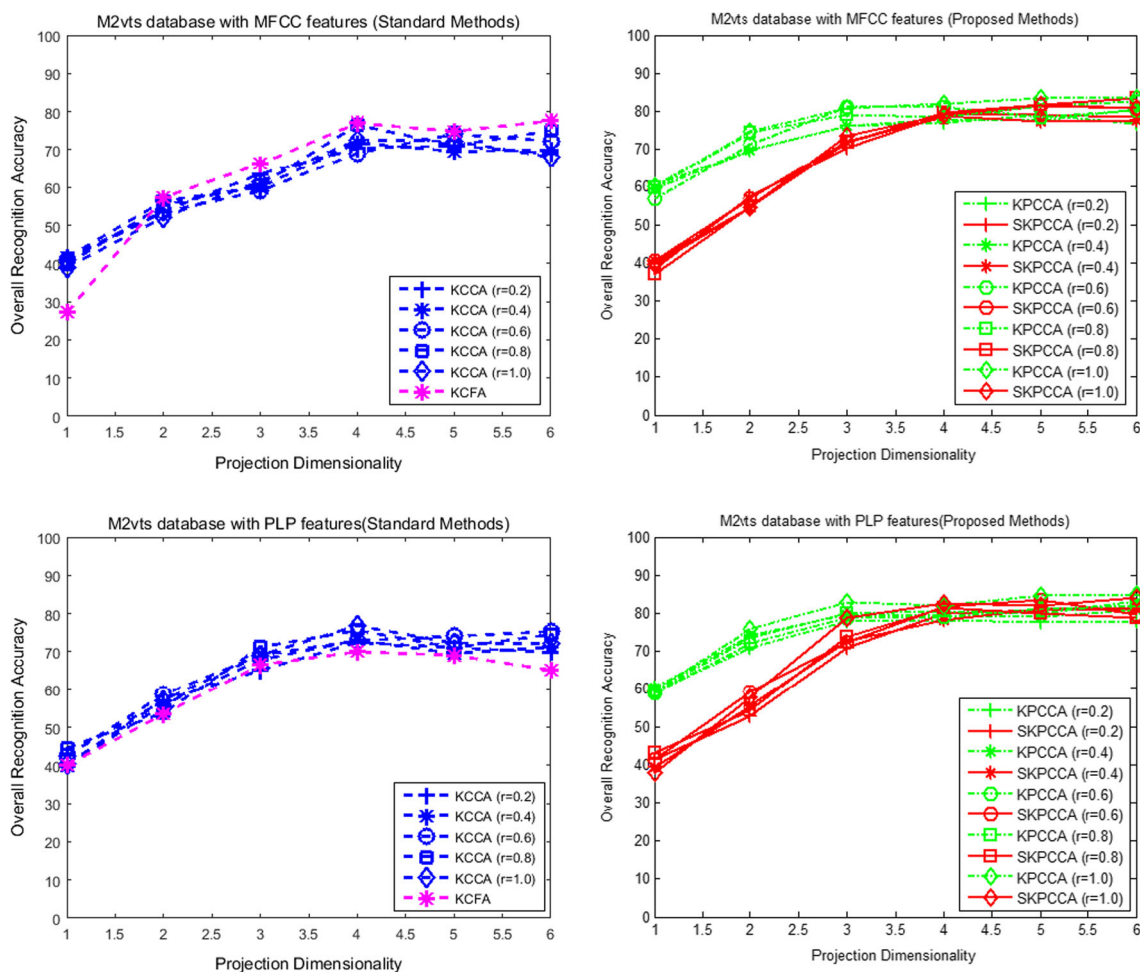


Fig. 9 Experimental results of Decision Fusion method. Top Left: M2vts database with MFCC features on KCCA and KCFA methods. Top Right: M2vts database with MFCC features on KPCCA and

SKPCCA methods. Bottom Left: M2vts database with PLP features on KCCA and KCFA methods. Bottom Right: M2vts database with PLP features on KPCCA and SKPCCA methods

Our audio-visual feature extraction process is drawn in Fig. 6.

4.2 Emotion recognition

To extract audio features, first the Hamming window is applied to the short time frame (512 samples) in order to preserve the stationary property of the signal. Successive windows have 50 % overlap. Next, the background noise is reduced by applying a threshold to the energy of wavelet coefficients in different scales and the ones with energy less than the threshold are removed and the signal is reconstructed again [35]. To obtain appropriate features, the p_i^{th} period and the energy of the signal in each window [36] and spectral features (first 13 Mel-Frequency Cepstral Coefficients (MFCC)) [12, 37] are extracted. Finally, all extracted acoustic features on each time frame are arranged as a feature vector.

In each video frame, the features of facial expression are extracted from the face region. Since successive windowed signals have 50 % overlap something that the video frames,

are without; the visual features are elicited from the nearest audio time. One challenge of feature extraction from each frame is to determine the accuracy rate of face region detection algorithm. In this article, the Haar cascade technique [38] is applied for the face detection.

Images of successive frames are all normalized into a frame size of 64×64 pixels. Gabor wavelet is very effective for describing spatial frequencies in images [39]. Accordingly, a Gabor filter bank with 5 scales and 8 orientations is used to extract the facial expression features [12, 40] which construct a very high dimensional feature vector. Thus, each sub-band of the sample is reduced to a size of 32×32 pixels followed by applying Principle Component Analysis (PCA) in them in order to reduce the number of features and elicit a rich feature set, as shown in Fig. 7.

5 Experimental results

To evaluate the performance of the proposed algorithms, the following databases are applied: M2vts database [41]

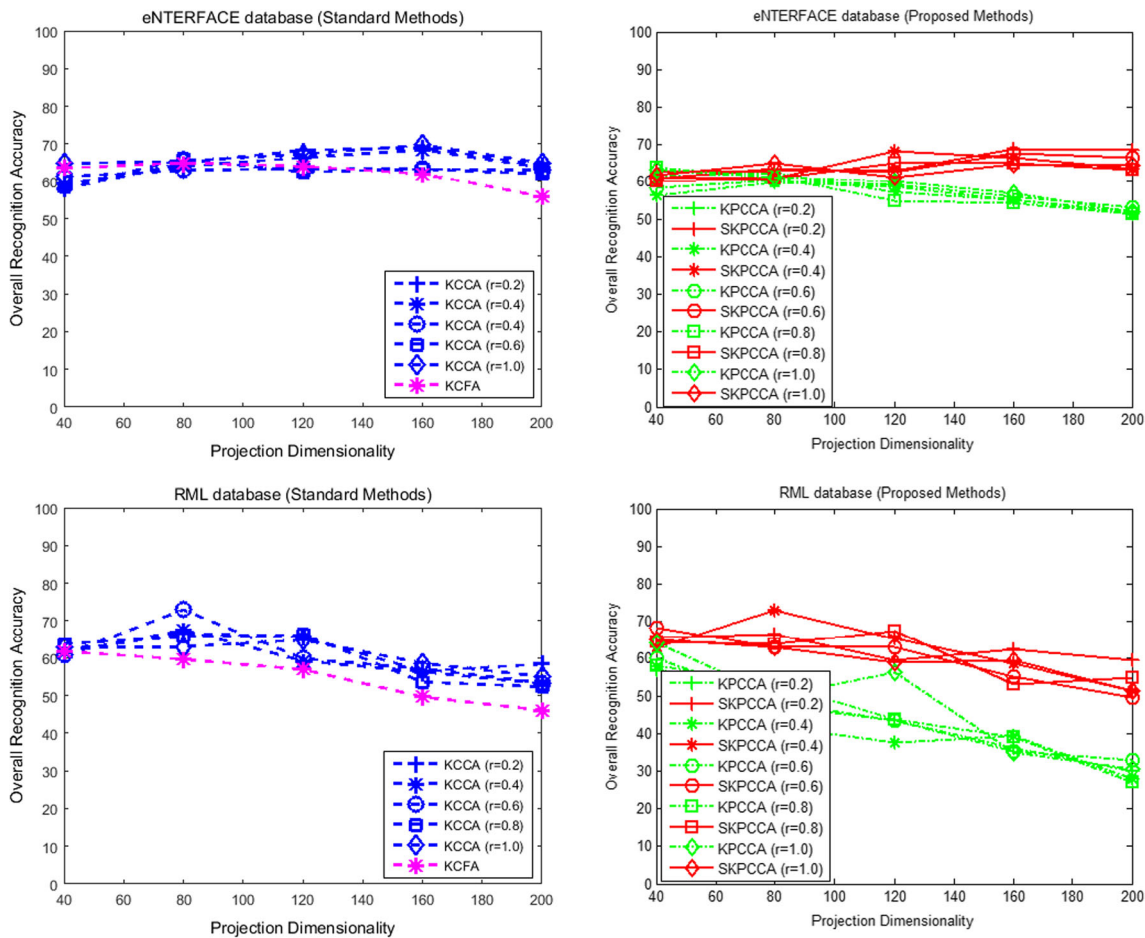


Fig. 10 Experimental results of Feature Fusion method. Top Left: eINTERFACE on KCCA and KCFA methods. Top Right: eINTERFACE on KPCCA and SKPCCA methods. Bottom Left: RML on KCCA and KCFA methods. Bottom Right: RML on KPCCA and SKPCCA methods

for audio visual speech recognition and eNTERFACE [42] and Ryerson (RML) databases [11] for audio visual emotion recognition.

The parameters of HMM classifiers are estimated in the cross-validation phase. The number of states is changed from 1 to 6 and the number of Gaussian components is changed from 1 to 5 within each state. The best HMM result for the speech datasets is achieved through 3 hidden states and 1 Gaussian mixture per state. Similarly, the best HMM results for the emotion recognition dataset are achieved through 6 hidden states and 3 Gaussian mixtures within each state. To find the variance parameter of Gaussian kernel, different values ($\sigma = 2, 6, 10, 14,$ and 20) are evaluated where $\sigma = 14$ for both applications has led to the best solution. The experiments are conducted by considering different values of the regularization parameter r ranged from 0.2 to 1.

5.1 Results on speech data

The audio-visual database (M2VTS) [41], contains 185 recordings of 37 subjects (12 females and 25 males) and provides 5 shots per person. During each shot, the subjects are asked to count from ‘0’ to ‘9’ and their audio and video data is recorded by the sampling rate of 48 KHz and frame rate of 25Hz, respectively. In this database, the speaker dependent recognition rate is applied to deal with high complexity and memory requirements of kernel methods.

The experimental results for audio-visual speech recognition accuracy using MFCC and PLP acoustic features for feature fusion and decision fusion, are shown in Figs. 8 and 9, respectively. In this proposed algorithm, the dimension sizes 1, 2 and 3 are of concern for ν space. The results obtained from the tests indicate that for $\nu = 3$, the best result is achieved. It is observed that in lower dimensions,

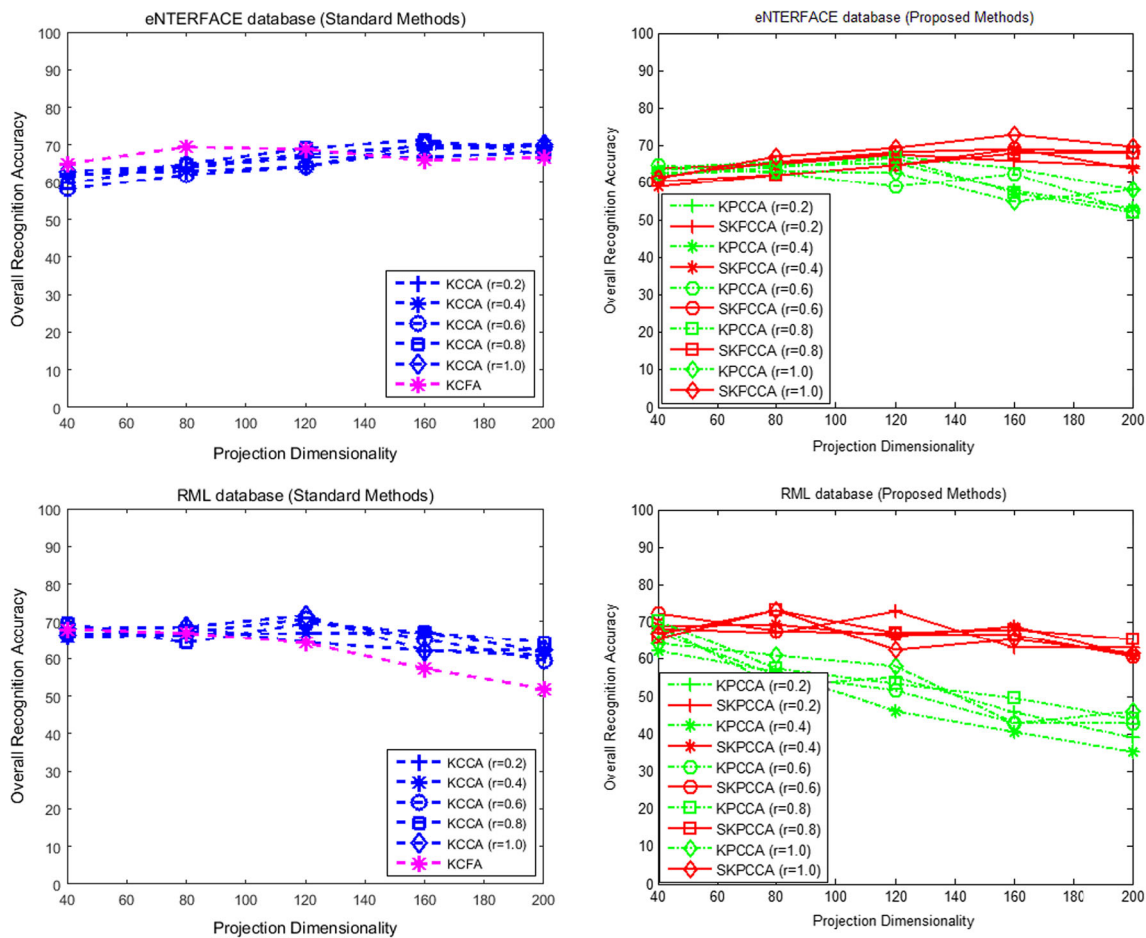


Fig. 11 Experimental results of Decision Fusion method. Top Left: eNTERFACE on KCCA and KCFA methods. Top Right: eNTERFACE on KPCCA and SKPCCA methods. Bottom Left: RML on KCCA and KCFA methods. Bottom Right: RML on KPCCA and SKPCCA methods

KPCCA provides better results than KCCA, KCFA and SKPCCA because it can estimate a fullcovariance matrix representing all information about the residual dimensions. Evidently, KPCCA provides similar results to SKPCCA in highdimensions.

5.2 Results on emotion data

Emotion databases (eINTERFACE and RML) consist of six basic human prototypical emotions (anger, disgust, fear, happy, sad and surprise) [43, 44]. The eINTERFACE database contains 44 subjects who showing 5 believable reactions to each emotion with the acoustic sampling rate of 48 KHz and visual frame rate of 25. Ryerson database contains 8 subjects, speaking 6 languages that generate 3 believable reactions to all the situations with acoustic sampling rate of 22050 Hz, and visual frame rate of 30. In eINTERFACE and RML datasets, each sample is truncated to 2 and 1.2-second-long, respectively, and divided into

10 segments. The dimensionality of the audio and visual features is empirically set to 200. To overcome the high complexity and memory requirements of kernel methods, 10 subjects are randomly selected for each experiment. These subjects are divided in two sets: 70 % for training 30 % for testing. This process is repeated 10 times, and the average results are presented in Figs. 10 and 11.

The emotion recognition accuracy of the conventional methods together with this proposed method for eINTERFACE and Ryerson (RML) databases are presented in Figs. 10 and 11. In SKPCCA the different dimension sizes are considered at 0, 25, 50 and 100 for the ν space and the results indicate that for $\nu = 50$, the best result is achieved.

Although KPCCA provides proper results when handling low dimensional data, it cannot perform well when encountering high dimensional inputs. This deficiency arises from the estimation of great number of parameters (fullcovariance Ω) with small number of train samples. It is observed that SKPCCA outperforms KCCA, KCFA and KPCCA

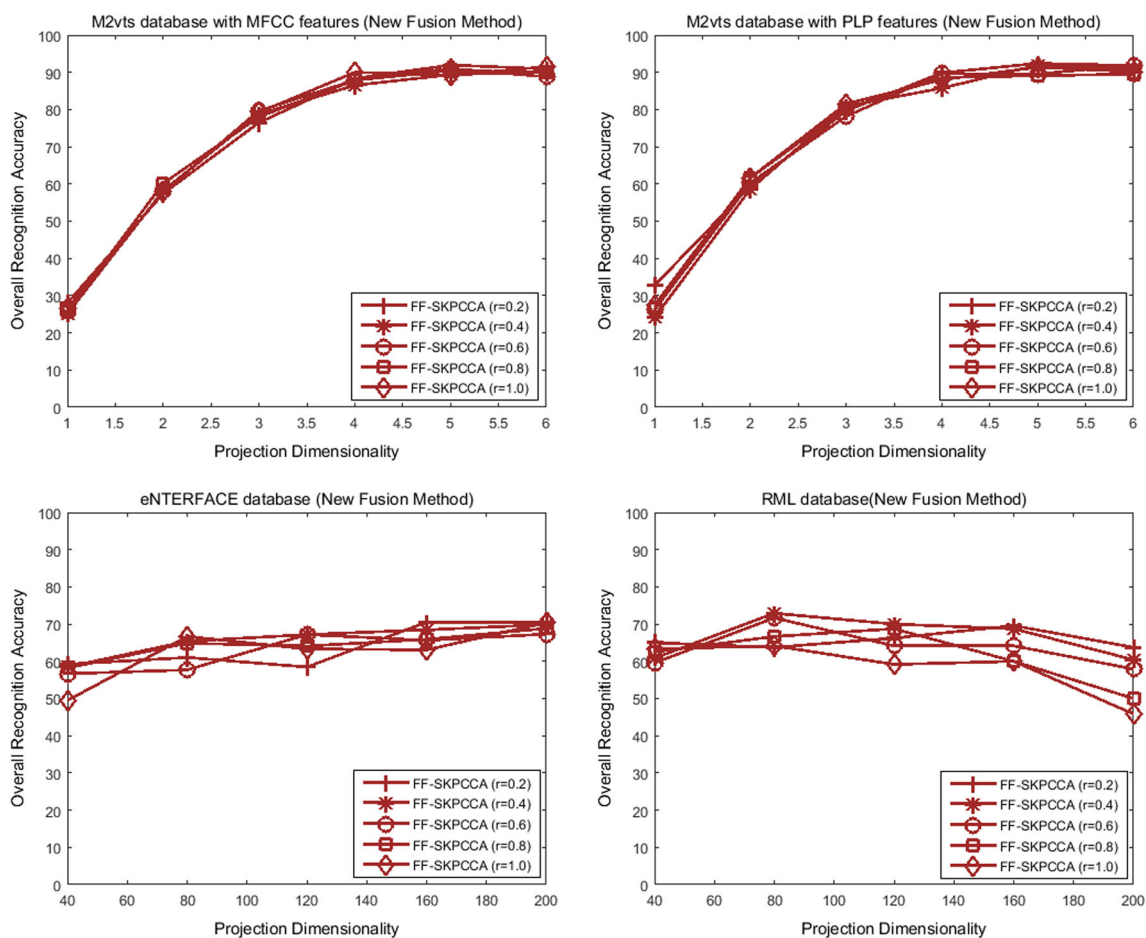


Fig. 12 Experimental results of New Fusion method. Top Left: M2vts database with MFCC features. Top Right: M2vts database with PLP features. Bottom Left: eINTERFACE database. Bottom Right: RML database

due to estimating a low rank (sparse) covariance matrix in highdimensional space with adequate number of samples.

5.3 The results of the proposed FF-SKPCCA method

As explained before, the conventional feature fusion methods synchronously concatenate latent variables of audio and visual modalities (Fig. 3a) which might lead to the curse of dimensionality problem. For example, in RML database which contains low reactions of each situation, the number of samples is not enough for the training phase; therefore, a decrease is expected in the results when conventional feature fusion methods are applied (Fig. 10).

To solve the above mentioned problems, the proposed feature fusion for KPCCA is applied to RML and other datasets. According to Figs. 12 and 13, it is observed that this proposed method overcomes both the problems of redundant features and curse of dimensionality, in comparison with the conventional feature fusion methods (Figs. 8 and 10). This proposed FF-SKPCCA reveals that fusion of latent variables into a unified set, instead of concatenating

them, can generate informative and lowdimensional feature vectors for describing the subjects' state.

The recognition accuracy of all conventional and proposed methods for emotion databases (eINTERFACE and RML) and speech database (M2vts with MFCC and PLP acoustic features) are demonstrated in Fig. 13. To make a fair comparison, maximum accuracy in each dimension for regularization parameters is shown. It is observed that at lowdimensions, FF-SKPCCA method does not provide proper results, in contrast when input dimension increases, this algorithm improves the performance considerably. In fact fusing different feature vectors is totally different from concatenating them; because by increasing both the number of modalities and the number of features, the recognition accuracy declines. Thus, FF-SKPCCA is designed to conceptually fuse the latent variables from different modalities instead of concatenating them, which leads to a reduction in both the redundancy and curse of dimensionality.

To have a better comparison among these methods, their results for each emotion class (eINTERFACE and RML contain 6 classes) and speech class (M2vts contain 10 classes)

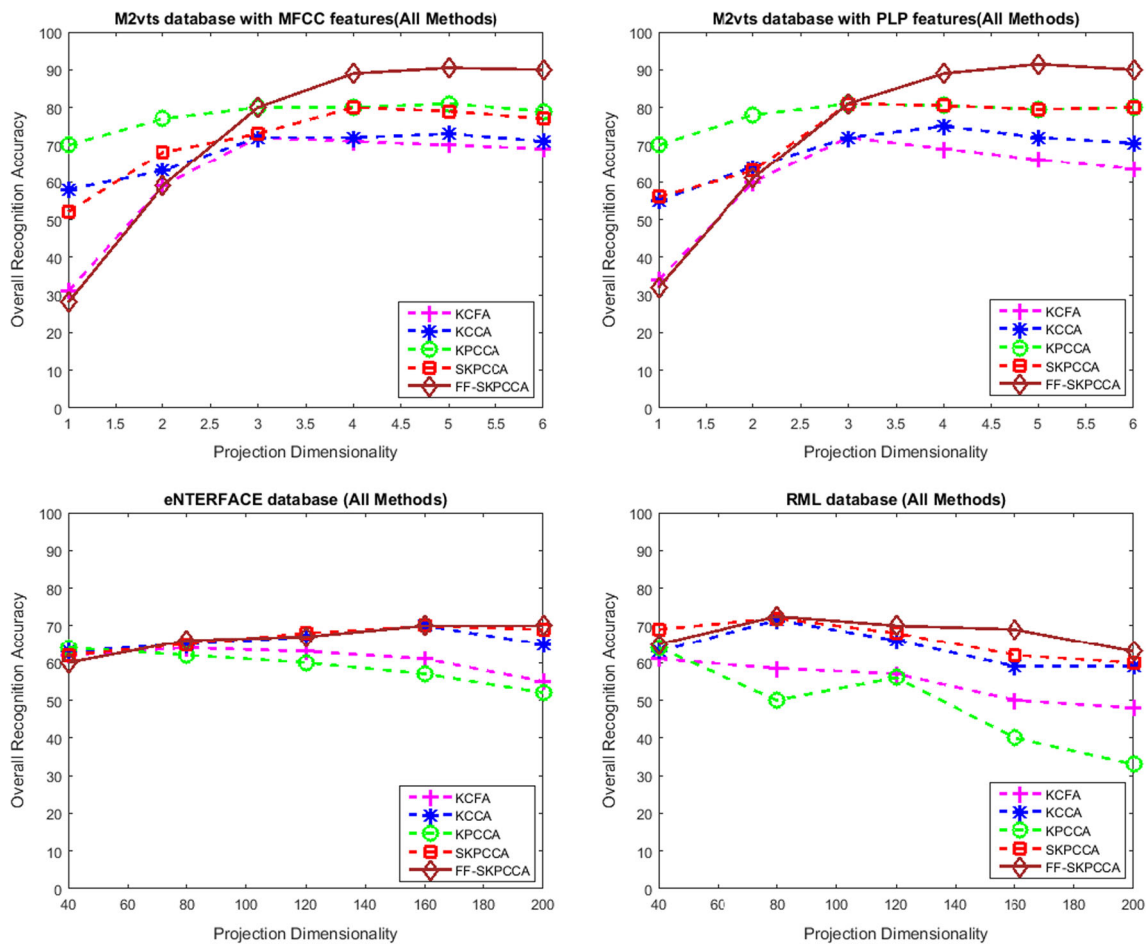


Fig. 13 Comparison overall recognition accuracy between KCCA, KCFA, KPCCA, SKPCCA and FF-SKPCCA. Top Left: M2vts database with MFCC features. Top Right: M2vts database with PLP features Bottom Left: eINTERFACE database. Bottom Right: RML database

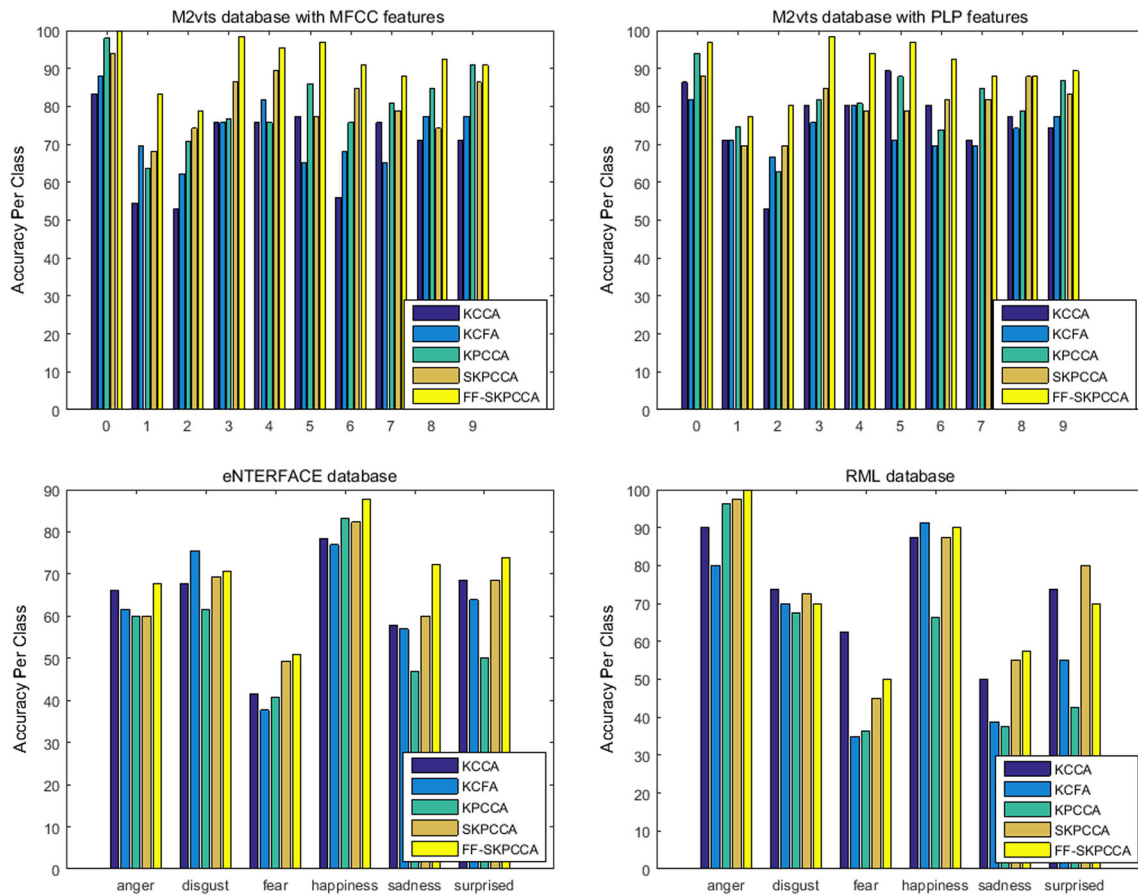


Fig. 14 Comparison between KCCA, KCFA, KPCCA, SKPCCA and FF-SKPCCA. Top Left: M2vts database with MFCC features. Top Right: M2vts database with PLP features. Bottom Left: eNTERFACE database. Bottom Right: RML database

are illustrated in bar-charts in Fig. 14. Note that, for both the emotion datasets, the FF-SKPCCA in average provides better performance than the conventional methods. However, regarding M2vts database, the FF-SKPCCA outperforms the compared methods for all cases significantly by improving the speech recognition accuracy in a significant manner.

6 Conclusion

Information fusion systems are in their infancy with respect to artificial intelligence. Naturally, they encounter different problems like coping with input variability, lack of detecting nonlinear dependencies and sensitivity to high dimensional inputs. To overcome these drawbacks, an almost coverall fusion algorithm for bimodal emotion and speech recognition systems is developed mathematically. This newly introduced KPCCA method provides good results on low-dimensional inputs but its performance is not sufficient for high dimensional ones. SKPCCA is capable of better handling high-dimensional inputs when the covariance matrix is sparse and its results on the emotion datasets confirm this claim. FF-SKPCCA, by fusing the latent variables

of two modalities and generate a low-dimensional set of latent variables solves the problems of redundancy and curse of dimensionality. Experimental results on several datasets demonstrated that FF-SKPCCA outperforms its counterparts in most cases.

Acknowledgments The authors of this paper acknowledge Dr. Homayounpour, professor of AmirKabir University, to let us using their M2VTS dataset in order to develop the experimental results.

Appendix A

A-1. CCA Method

A proposed statistical method named Canonical Correlation Analysis (CCA) is proposed by [14], in order to find a shared structure between two sources of data. CCA is closely related to the mutual information method [45] but it has some differences in terms of objective function. A pair of feature vectors with zero is considered in the method as follows:

$$(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (15)$$

where x_i and y_i are the observation data (original features) of the two modalities with dimensions of p and q , respectively. CCA seeks to develop two transformation matrices of \mathbf{W}_x and \mathbf{W}_y with dimensions of $p \times d$ and $q \times d$ respectively, where $d \leq \min(pq)$. The original features of these modalities are projected to the correlation subspace by \mathbf{W}_x and \mathbf{W}_y in a manner that the correlation between $\hat{x} = \mathbf{x}\mathbf{W}_x$ and $\hat{y} = \mathbf{y}\mathbf{W}_y$ is maximized. Maximizing the correlation between the projected feature vectors of \hat{x} and \hat{y} is the same as maximizing ρ (correlation coefficient) between them as follows:

$$\begin{aligned} \rho &= \max_{\mathbf{W}_x, \mathbf{W}_y} \frac{E[\hat{x}^T \hat{y}]}{\sqrt{E[\hat{x}^2] E[\hat{y}^2]}} \\ &= \max_{\mathbf{W}_x, \mathbf{W}_y} \frac{E[\mathbf{W}_x^T \mathbf{x}^T \mathbf{y} \mathbf{W}_y]}{\sqrt{E[\mathbf{W}_x^T \mathbf{x}^T \mathbf{x} \mathbf{W}_x] E[\mathbf{W}_y^T \mathbf{y}^T \mathbf{y} \mathbf{W}_y]}} \\ &= \max_{\mathbf{W}_x, \mathbf{W}_y} \frac{\mathbf{W}_x^T C_{xy} \mathbf{W}_y}{\sqrt{\mathbf{W}_x^T C_{xx} \mathbf{W}_x \mathbf{W}_y^T C_{yy} \mathbf{W}_y}} \end{aligned} \tag{16}$$

where C_{xy} is the cross-covariance matrix of (\mathbf{x}, \mathbf{y}) and C_{xx}, C_{yy} are the covariance matrices of \mathbf{x} and \mathbf{y} respectively.

The above equation can be solved as an Eigen-value problem like:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} \mathbf{W}_x = \rho^2 \mathbf{W}_x \tag{17}$$

$$C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} \mathbf{W}_y = \rho^2 \mathbf{W}_y \tag{18}$$

A-2. CFA Method

The cross-modal factor analysis (CFA) method is proposed by [15], where, the features from different modalities are treated as two subsets where the same patterns between these two subsets are discovered. In this method, it is assumed that a pair of normalized feature vectors \mathbf{x} and \mathbf{y} with zero means are linearly projected into a joint space applying \mathbf{W}_x and \mathbf{W}_y transforms, in a manner that the following criterion can be minimizing:

$$\min_{\mathbf{W}_x, \mathbf{W}_y} \|\mathbf{x}\mathbf{W}_x - \mathbf{y}\mathbf{W}_y\|_F^2 \tag{19}$$

where, $\mathbf{W}_x^T \mathbf{W}_x$ and $\mathbf{W}_y^T \mathbf{W}_y$ are unit matrices and F is the Frobenius norm and is calculated by $\|\mathbf{W}\|_F = \sqrt{\sum_{ij} w_{ij}^2}$.

By solving the above equation for optimal transformation matrices \mathbf{W}_x and \mathbf{W}_y and decomposing cross-covariance matrix C_{xy} through Singular Value Decomposition (SVD) method, the following equation is obtained:

$$C_{xy} = S_{xy} \Lambda_{xy} D_{xy} \tag{20}$$

Consequently,

$$\mathbf{W}_x = S_{xy} \& \mathbf{W}_y = D_{xy} \tag{21}$$

A-3. Probabilistic CCA

To deal with the uncertainty problem in the CCA performance, the probabilistic CCA (PCCA) is introduced by [24] through the projected latent variables provide maximum variance in the joint correlation space. To do this, they defined a Gaussian model for every single source of data as follow:

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \quad 1 \leq d \leq \min(p, q) \\ \mathbf{x}|z &= \mathcal{N}(z\mathbf{W}_x^T + \mu_x, \varphi_x) \\ \mathbf{y}|z &= \mathcal{N}(z\mathbf{W}_y^T + \mu_y, \varphi_y) \end{aligned} \tag{22}$$

where, z is the latent variable, shared between the two modalities \mathbf{x} and \mathbf{y} and μ and φ are the mean and covariance of each data, respectively. Here, by maximizing the probability functions, the φ_x and φ_y should be minimized. By considering $\mathbf{O} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$, $\mathbf{W} = [\mathbf{W}_x \mathbf{W}_y]$,

$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and $\varphi = \begin{bmatrix} \varphi_x & \mathbf{0} \\ \mathbf{0} & \varphi_y \end{bmatrix}$ both probabilistic functions are merged into the following joint probabilistic function as:

$$\mathbf{p}(\mathbf{O}, z) = \mathbf{p}(\mathbf{O}|z)\mathbf{p}(z) = N(z\mathbf{W}^T + \mu, \varphi) N(0, I_d) \tag{23}$$

They indicate that the posterior expectation of z given \mathbf{x} and \mathbf{y} are:

$$\begin{aligned} \varphi_x &= C_{xx} - \mathbf{W}_x \mathbf{W}_x^T \\ \varphi_y &= C_{yy} - \mathbf{W}_y \mathbf{W}_y^T \\ E(z|\mathbf{x}) &= \mathbf{x}\mathbf{W}_x M_x^{-1}, M_x = I + \mathbf{W}_x^T \varphi_x^{-1} \mathbf{W}_x \\ E(z|\mathbf{y}) &= \mathbf{y}\mathbf{W}_y M_y^{-1}, M_y = I + \mathbf{W}_y^T \varphi_y^{-1} \mathbf{W}_y \end{aligned} \tag{24}$$

where, \mathbf{W}_x and \mathbf{W}_y are the first d canonical directions of \mathbf{x} and \mathbf{y} . The parameters C_{xx} and C_{yy} are the covariances of \mathbf{x} and \mathbf{y} , respectively.

However, this new method named the unified latent variable can identify a latent variable, given \mathbf{x} and \mathbf{y} as :

$$\begin{aligned} E(z|\mathbf{x}, \mathbf{y}) &= [\mathbf{x}\mathbf{W}_x \quad \mathbf{y}\mathbf{W}_y] \begin{bmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} M_x^{-1} \\ M_y^{-1} \end{bmatrix} \end{aligned} \tag{25}$$

where, $P_d = M_x^{-1} * (M_y^{-1})^T$.

Another solution for (22) is based on the expectation maximization (EM) algorithm. Similarly, the Probabilistic Principle Component Analysis (PPCA) is proposed by [46] which iterates through the following expectation maximization (EM).

- **Expectation-step:** finds the sufficient statistics of the latent variables given the current estimated parameter:

$$\begin{aligned} M_t &= I + \mathbf{W}_t^T \varphi_t^{-1} \mathbf{W}_t \\ E(z_t) &= M_t^{-1} \mathbf{W}_t \varphi_t^{-1} \mathbf{O} \\ E(z_t z_t^T) &= M_t^{-1} + E(z_t) E(z_t)^T \end{aligned} \tag{26}$$

where subscript t indicate the iteration number.

- **Maximization-step:** updates the estimated parameter to maximize the likelihood function:

$$\begin{aligned} \mathbf{W}_{t+1} &= \left[\mathbf{O} E(z_t)^T \right] \left[E(z_t z_t^T) \right]^{-1} \\ \varphi_{t+1} &= \mathbf{O} \mathbf{O}^T - 2 \mathbf{O} E(z_t)^T \mathbf{W}_{t+1}^T \\ &\quad + \text{trace}(E(z_t z_t^T) \mathbf{W}_{t+1}^T \mathbf{W}_{t+1}) \end{aligned} \tag{27}$$

By inserting (26) into (27), this method provides a general solution for PCCA scheme which yields the following updated equation:

$$\begin{aligned} \mathbf{W}_{t+1} &= C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \left(M_t^{-1} + M_t^{-1} \mathbf{W}_t^T \varphi_t^{-1} C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \right)^{-1} \\ \varphi_{t+1} &= \begin{pmatrix} (C - C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \mathbf{W}_{t+1}^T)_{11} & 0 \\ 0 & (C - C \varphi_t^{-1} \mathbf{W}_t M_t^{-1} \mathbf{W}_{t+1}^T)_{22} \end{pmatrix} \end{aligned} \tag{28}$$

where $M_t = I + \mathbf{W}_t^T \varphi_t^{-1} \mathbf{W}_t$.

A-4. KCCA Method

Kernel Canonical Correlation Analysis (KCCA) [21] is the kernelized version of CCA method that projects data into higher dimensional feature spaces and applies CCA to the data in the kernel space in order to find a nonlinear correlation between the two modalities. Let us consider ϕ and ψ as two mapping functions that map the input data into a space of higher dimension:

$$\begin{aligned} (\phi(\mathbf{x}), \psi(\mathbf{y})) &= \{(\phi(x_1), \psi(y_1)), (\phi(x_2), \psi(y_2)), \\ &\quad \dots, (\phi(x_n), \psi(y_n))\} \end{aligned} \tag{29}$$

The KCCA seeks to develop the two matrices α and β that are applied in the following equations:

$$\mathbf{W}_x = \phi(\mathbf{x})^T \alpha \tag{30}$$

$$\mathbf{W}_y = \psi(\mathbf{y})^T \beta \tag{31}$$

This means that \mathbf{W}_x and \mathbf{W}_y are the projections of $\phi(\mathbf{x})$ and $\psi(\mathbf{y})$ onto α and β , respectively. By inserting ϕ and ψ , into (16), the correlation function is applied as:

$$\begin{aligned} \rho &= \max_{\alpha, \beta} \frac{E[\alpha^T \phi(\mathbf{x}) \cdot \phi(\mathbf{x})^T \psi(\mathbf{y}) \cdot \psi(\mathbf{y})^T \beta]}{\sqrt{E[\alpha^T \phi(\mathbf{x}) \cdot \phi(\mathbf{x})^T \phi(\mathbf{x}) \cdot \phi(\mathbf{x})^T \alpha] E[\beta^T \psi(\mathbf{y}) \cdot \psi(\mathbf{y})^T \psi(\mathbf{y}) \cdot \psi(\mathbf{y})^T \beta]}} \\ &= \max_{\alpha, \beta} \frac{\alpha^T \mathbf{K}_x \mathbf{K}_y \beta}{\sqrt{[\alpha^T \mathbf{K}_x \mathbf{K}_x \alpha] [\beta^T \mathbf{K}_y \mathbf{K}_y \beta]}} \end{aligned} \tag{32}$$

where, $\mathbf{K}_x = E[\phi(\mathbf{x}) \cdot \phi(\mathbf{x})^T]$ and $\mathbf{K}_y = E[\psi(\mathbf{y}) \cdot \psi(\mathbf{y})^T]$.

This optimization problem can be solved through the generalized Eigen-value decomposition method. When kernel functions are non-invertible, conventional regularization technique can be applied, therefore, the following Equation is yield [30]:

$$\max_{\alpha, \beta} \frac{\alpha^T \mathbf{K}_x \mathbf{K}_y \beta}{\sqrt{[\alpha^T (\mathbf{K}_x^2 + \tau \mathbf{K}_x) \alpha] [\beta^T (\mathbf{K}_y^2 + \tau \mathbf{K}_y) \beta]}} \tag{33}$$

where, $0 \leq \tau \leq 1$.

A-5. KCFA Method

Kernel CFA [12] approach can provide correct information association provided that the two modalities are not linearly related. To illustrate this fact $X = (\phi(x_1), \phi(x_2), \dots, \phi(x_n))^T$, and $Y = (\psi(y_1), \psi(y_2), \dots, \psi(y_n))^T$ represent the two matrices with each row representing a sample in the nonlinearly mapped feature space; next, the $X^T Y = S_{xy} \Lambda_{xy} D_{xy}$ should be solved through kernel method. The kernel matrices of the two subsets of features can be computed as $K_x = X X^T$ and $K_y = Y Y^T$. By performing eigenvalue decomposition on the product of the kernel matrices $K_x K_y$, it becomes obvious that

$$\begin{aligned} (K_x K_y) \beta_i &= \lambda_i \beta_i \\ (X X^T Y Y^T) \beta_i &= \lambda_i \beta_i \\ (Y^T X X^T Y) Y^T \beta_i &= \lambda_i Y^T \beta_i \end{aligned} \tag{34}$$

Since the right singular vectors of the SVD of $X^T Y$, D_{xy} correspond with the eigenvectors of $Y^T X X^T Y = (X^T Y)^T (X^T Y) Y^T \beta_i$ corresponds to the columns of D_{xy} , which can be further normalized into a unit norm as:

$$v_i = \frac{Y^T \beta_i}{\|Y^T \beta_i\|} = \frac{Y^T \beta_i}{\sqrt{\beta_i^T Y Y^T \beta_i}} = \frac{Y^T \beta_i}{\sqrt{\beta_i^T K_y \beta_i}} \tag{35}$$

For a feature vector y' with nonlinear mapping $\psi(y')$, the projection can be computed as

$$\begin{aligned} v_i^T \psi(y') &= \left(\frac{Y^T \beta_i}{\sqrt{\beta_i^T K_y \beta_i}} \right)^T \psi(y') \\ &= \frac{1}{\sqrt{\beta_i^T K_y \beta_i}} \beta_i^T \begin{bmatrix} K(y', y_1) \\ K(y', y_2) \\ \dots \\ K(y', y_n) \end{bmatrix} \end{aligned} \tag{36}$$

Similarly, it can be illustrated that

$$(K_y K_x) \alpha_j = (X^T Y Y^T X) X^T \alpha_j = \lambda_j X^T \alpha_j \quad (37)$$

The left singular vectors S_{xy} are the eigenvectors of $X^T Y Y^T X = (X^T Y)(X^T Y)^T$, hence $X^T \alpha_j$ corresponds to the S_{xy} columns, which can be normalized into a unit norm as:

$$\mu_j = \frac{X^T \alpha_j}{\|X^T \alpha_j\|} = \frac{X^T \alpha_j}{\sqrt{\alpha_j^T X X^T \alpha_j}} = \frac{X^T \alpha_j}{\sqrt{\alpha_j^T K_x \alpha_j}} \quad (38)$$

By allowing x' to be a feature vector in the original domain where the nonlinear mapping is $\phi(x')$, the feature vector in the cross-modal associated domain can be computed as:

$$\begin{aligned} \mu_j^T \phi(x') &= \left(\frac{X^T \alpha_j}{\sqrt{\alpha_j^T K_x \alpha_j}} \right)^T \phi(x') \\ &= \frac{1}{\sqrt{\alpha_j^T K_x \alpha_j}} \alpha_j^T \begin{bmatrix} K(x', x_1) \\ K(x', x_2) \\ \dots \\ K(x', x_n) \end{bmatrix} \end{aligned} \quad (39)$$

References

- Shivappa S, Trivedi M, Rao B (2010) Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. *Proc IEEE* 98(10):1692–1715
- Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans Pattern Anal Mach Intell*:39–58
- Ayadi ME, Kamel M, Karray F (2011) Survey on speech emotion recognition: features, classification schemes and databases. *Pattern Recogn* 44(3):572–587
- Pradeep KA, M.Anwar H, Abdulmotaleb ES, Mohan SK (2010) Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*:345–379
- Galatas G, Potamianos G, Makedon F (2012) Audio-visual speech recognition incorporating facial depth information captured by the Kinect. In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp 2714–2717
- Gupta R, Malandrakis N, Xiao B, Guha T, Van Segbroeck M, Black M, Potamianos A, Narayanan S (2014) Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 33–40. Orlando Florida, USA: ACM
- Taouche C, Batouche MC, Berkane M, Taleb-Ahmed A (2014) Multimodal biometric systems. In: *International Conference on Multimedia Computing and Systems (ICMCS)*, pp 301–308
- Xu C, Hero AO (2012) Savarese, s multimodal video indexing and retrieval using directed information. *IEEE Trans Multimedia*:3–16
- Ercan AO, Gamal AE, Guibas LJ (2013) Object tracking in the presence of occlusions using multiple cameras: a sensor network approach. *ACM trans Sen Netw*:16:1–16:36
- Wagner J, Andre E, Lingenfelter F, Jonghwa K (2011) Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans Affect Comput*:206–218
- Wang Y, Guan Y (2008) Recognizing human emotional state from audiovisual signals. *IEEE Trans Multimedia* 10(5):936–946
- Wang Y, Guan Y, Venetsanopoulos AN (2012) Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Trans Multimedia*:597–607
- Li B, Qi L, Gao L (2014) Multimodal emotion recognition based on kernel canonical correlation analysis
- Hotelling H (1936) Relations between two sets of variates. *Biometrika*:321–377
- Li D, Dimitrova N, Li N, Sethi IK (2003) Multimedia content processing through cross-modal association. In: *Proceedings ACM International Conference*, pp 604–611
- Bredin H, Chollet G (2007) Audio-visual speech synchrony measure for talking-face identity verification. In: *Acoustics, Speech and Signal Processing, ICASSP 2007*, pp II–233
- Abo-Zahhad M, Ahmed SM, Abbas SN (2014) PCG biometric identification system based on feature level fusion using canonical correlation analysis. In: *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp 1–6
- Metallinou A, Lee S, Narayanan S (2010) Decision level combination of multiple modalities for recognition and analysis of emotional expression. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp 2462–2465
- Li D, Taskiran C, Dimitrova N, Wang W, Li M, Sethi IK (2005) Cross-modal analysis of audio-visual programs for speaker detection. In: *Proceedings IEEE Workshop Multimedia Signal Process.*, Shanghai, China, pp 1–4
- Kumar K, Potamianos G, Navratil J, Marcheret E, Libal V (2011) Audio-visual speech synchrony detection by a family of bimodal linear prediction models. *Multibiometrics for Human Identification*:31–50
- Lai PL, Fyfe C (2000) Kernel and nonlinear canonical correlation analysis. *Int J Neural Syst*:365–377
- Shi Y, Ji H (2014) Kernel canonical correlation analysis for specific radar emitter identification. *Electron Lett*:1318–1320
- Chetty G, Göcke R, Wagner M (2009) Audio-Visual mutual dependency models for biometric liveness checks. *AVSP 2009*, Norwich, pp. 32–37
- Bach F, Jordan MI (2005) A probabilistic interpretation of canonical correlation analysis. Technical Report 688 Department of Statistics, University of California, Berkeley
- Archambeau C, Bach FR (2009) Sparse probabilistic projections. *Adv Neural Inf Proces Syst* 21:73–80
- Klami A, Virtanen S, Kaski S (2010) Bayesian exponential family projections for coupled data sources. In: *26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp 286–293
- Koskinen M, Viinikanoja J, Kurimo M, Klami A, Kaski S, Hari R (2013) Identifying Fragments of natural speech from the listener's MEG signals. *Hum Brain Mapp* 34(6):1477–1489
- Rudovic O, Petridis S, Pantic M (2013) Bimodal log-linear regression for fusion of audio and visual features. *21st ACM Int Conf Multimedia*:789–792
- Wu CH, Lin JC, Wei WL (2014) Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, e12
- Hardoon D, Szedmak S, Shawe-taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. *Neural Comput*:2639–2664
- Blaschko M, Lampert CH (2008) Correlational spectral clustering. *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*:1–8
- Golub GH, Hansen PC, O'Leary DP (1999) Tikhonov Regularization and total least squares. *SIAM J Matrix Anal Appl* 21(1):185–194

33. Rohani R, Sobhanmanesh F, Alizadeh S, Boostani R (2011) Lip processing and modeling based on spatial fuzzy clustering in color images. *Int J Fuzzy Syst* 13(2):65–73
34. Hermansky H, Hanson BA, Wakita H (1985) Perceptually-based linear predictive analysis of speech. In: *Proceedings IEEE ICASSP*, vol 2, pp 509–512
35. Bartlett A, Evans V, Frenkel I, Hobson C, Sumera E (2004) *Digital Hearing Aids* [Online]. Available www.clear.rice.edu/elec301/Projects01/dig_hear_aid
36. Wu CH, Lin JC, Wei WL (2013) Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. *IEEE Trans Multimedia*:1880–1895
37. Jiang D, Cui Y, Zhang X, Fan P, Gonzalez I, Sahli H (2011) Audiovisual emotion recognition based on triple-stream dynamic Bayesian network models. *Affective Computing and Intelligent Interaction*:609–618
38. Sing V, Shokeen V, Singh B (2013) Face detection by haar cascade classifier with simple and complex backgrounds images using opencv implementation. *International Journal of Advanced Technology in Engineering and Science*:33–38
39. Lyons MJ, Budynek J, Plante A, Akamatsu S (2000) Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis. *4th Int Conf Automatic Face and Gesture Recognition*:202–207
40. Manjunath BS, Ma WY (1996) Texture features for browsing and Texture features for browsing and *IEEE Trans. Pattern Anal Machine Intell*, pp 837–842
41. Pigeon S, Vandendorpe L (1997) The M2VTS multimodal face database (release 1.00). In *Audio-and Video-Based Biometric Person Authentication*. Springer, Berlin Heidelberg, pp 403–409
42. Martin O, Kotsia I (2006) Macq, B Pitas, I The eNTERFACE05 audiovisual emotion database. In: *Proc. ICDEW*, p 8
43. Ekman P, Friesen WV, Press CP (1975) *Pictures of facial affect*. consulting psychologists press
44. Ekman P (1993) Facial expression and emotion. *Am Psychol*:384
45. Sun QS, Zeng SG, Liu Y, Heng PA, Xia DS (2005) A new method of feature fusion and its application in image recognition. *Pattern Recogn*:2437–2448
46. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society B* 61(3):611–622
47. Li YO, Eichele T, Calhoun VD, Adali T (2012) Group study of simulated driving fMRI data by multiset canonical correlation analysis. *Journal of signal processing systems*:31–48
48. Lin JC, Wu CH, Wei WL (2012) Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Trans Multimedia*:142–156
49. Mello S, Kory J (2012) Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp 31–38
50. Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun*:98–112
51. Muramatsu D, Iwama H, Makihara Y, Yagi Y (2013) Multi-view multi-modal person authentication from a single walking image sequence. *2013 International Conference on Biometrics (ICB)*:1–8

Reza Rohani Sarvestani He received a B.Sc. in computer science from Bahonar University, Iran, and a M.Sc. in artificial intelligence from Shiraz University, Iran, in 2006 and 2009, respectively. In 2009 he started Ph.D. degree in artificial intelligence from Shiraz University, Iran. His research interests include speech processing, data fusion and biomedical signal processing. Now he is a faculty of member of Shahrekord University at Computer Department, Engineering Faculty of Shahrekord University, Iran.

Reza Boostani He is born in Shiraz in 1973. He got his B.Sc. in Electronic Engineering at Shiraz University in 1996. He then continued his M.Sc. and Ph.D. in Biomedical Engineering at AmirKabir University of Technology (Tehran Polytechnic) in 1999 and 2005, respectively. From 2005 till present, he is a faculty of member of Shiraz University at CSE & IT Department, Electrical and Computer Engineering Faculty of Shiraz University. At the moment, he is an Associate Professor and is the head of Biomedical Engineering Group. His interested topics is Biomedical Signal Processing, Statistical Pattern Recognition and Bioinstrument.