Incomplete Cholesky Decomposition based Kernel Cross Modal Factor Analysis for Audiovisual Continuous Dimensional Emotion Recognition

Xia Li[1,2], Guanming Lu[1*], Jingjie Yan[1], Haibo Li[1], Zhengyan Zhang[1,3], Ning Sun[4], Shipeng Xie[1]

[1]College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[e-mail: xialiahut@163.com, lugm@njupt.edu.cn, yanjingjie1212@163.com, lihb@njupt.edu.cn, zhangzhengyan@just.edu.cn, xie@njupt.edu.cn]
[2]School of Mathematics and Physics, Anhui University of Technology, Maanshan 243000, China
[3]School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang 212003 China,
[4]Engineering Research Center of Wideband Wireless Communication Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[e-mail: sunning@njupt.edu.cn]
*Corresponding author: Guanming Lu

Received June 7, 2017; revised June 5, 2018; accepted September 4, 2018; published February 28, 2019

Abstract

Recently, continuous dimensional emotion recognition from audiovisual clues has attracted increasing attention in both theory and in practice. The large amount of data involved in the recognition processing decreases the efficiency of most bimodal information fusion algorithms. A novel algorithm, namely the incomplete Cholesky decomposition based kernel cross factor analysis (ICDKCFA), is presented and employed for continuous dimensional audiovisual emotion recognition, in this paper. After the ICDKCFA feature transformation, two basic fusion strategies, namely feature-level fusion and decision-level fusion, are explored to combine the transformed visual and audio features for emotion recognition. Finally, extensive experiments are conducted to evaluate the ICDKCFA approach on the AVEC 2016 Multimodal Affect Recognition Sub-Challenge dataset. The experimental results show that the ICDKCFA method has a higher speed than the original kernel cross factor analysis with the comparable performance. Moreover, the ICDKCFA method achieves a better performance than other common information fusion methods, such as the Canonical correlation analysis, kernel canonical correlation analysis and cross-modal factor analysis based fusion methods.

Keywords: continuous dimensional emotion recognition, incomplete Cholesky decomposition, kernel cross-modal factor analysis, multimodal information fusion

http://doi.org/10.3837/tiis.2019.02.018 **ISSN : 1976-7277**

1. Introduction

The human emotion can be expressed and inferred through various modalities such as voice, facial expression, body gesture, and various physiological indices [1]-[3]. It is known that using multiple modalities can improve emotion recognition performance [2]-[6]. However, how to make full use of the multimodal information to implement emotion recognition is still a challenging problem [2]-[6].

In the past few decades, there have been a lot of studies on multimodal emotion recognition [3][15]. However, most of them have been focused on multimodal discrete emotion recognition [9][10]. However, the discrete categorical representation of emotion could neither distinguish the subtle difference of the emotion nor describe the evolvement of the emotion [1]. In order to overcome these shortcomings, more attention has recently been focused on continuous dimensional emotion recognition . Especially, since 2012, multimodal continuous dimensional emotion recognition has always been a major task in the Audio-Visual Emotion recognition Challenge (AVEC) [11]-[15].This has greatly boosted the development of multimodal continuous dimensional recognition in various aspect such as feature extraction, feature fusion and recognition method. What was of interest to us was the feature fusion. In general, the fusion methods used in multimodal continuous dimensional emotion recognition can be divided into feature level, decision level, model level fusion, and mixed approaches [1][7].

For feature level fusion, the information from multiple modalities is combined to generate the recognition feature [1][7]. The simplest method is to construct a joint feature as the input of a regression model by concatenating the features from all modalities [1][7][11][16]-[19]. Additionally, many other feature-level fusion strategies have been proposed. Eyben et al. [16] proposed a string-based audiovisual fusion method based on the simple feature level fusion idea to fuse audiovisual behavior events such as head gestures, facial action unit, laughter and sighs. Soladié et al. [17] proposed a radial basis function (RBF) system, in which the fusion of the input relevant features was implemented via the k-mean clustering algorithm to generate a set of representative samples.

For decision level fusion, the predictions from various models based on each single modality are combined by various strategies to obtain the final prediction [1][7]. By far, there have been many strategies to combine the predictions. The traditional methods include weighted summing[20][21], averaging[13][22] and median calculating[22]. Linear regression is also a common method that has been used in [15][16][22][23]. These methods can also be improved, e.g. Nicolle et al. [24] expanded the linear regression to the local linear regressions.

For model level fusion, a designed model itself could not only combine the multimodal information and the other information of the emotion, but also obtain the emotion recognition results [1][7]. Soladié et al. [25] introduced a fuzzy inference system to fuse multimodal features and recognize dimensional emotion. Metallinou et al. [26] proposed a Gaussian Mixture Model (GMM) to fuse the visual and audio features as well as track the emotion.

Except for the three types of fusion above-mentioned, some hybrid approaches have also been proposed. Sayedelahl et al. [20] presented a combined bi-modal feature-decision fusion approach to improve the performance of emotion recognition. Tian et al. [27] proposed a hierarchical fusion strategy to combine features from different modalities at different layers of a hierarchical structure.

Feature-level and decision-level fusion are the most popular fusion strategies due to their simplicity in theory and flexibility in application. Since the emotion information is imbedded in the audio and visual modal with a complementary redundant manner, identifying the relationship between the two modalities before implementing feature-level or decision-level fusion is expected to effectively improve emotion recognition performance [1][4][5]. Canonical correlation analysis (CCA) [28] is a popular method to analyze the relationship between two modalities. The CCA, especially its kernel version, i.e. kernel canonical correlation analysis (KCCA) [28], has often been used in multimodal recognition. For example, Song et al. [29] used KCCA to compute the visual and audio transformed features before implementing feature-level fusion. However, the implementation of CCA involves the calculation of the inverse of the covariance matrices from two modalities. When at least one of two matrices is non-invertible or even just close to non-invertible, the method will result in large errors [5][6]. In order to break through the restriction, Li et al. [6] proposed cross-modal factor analysis (CFA), that could effectively analyze the linear relationship between two modalities and avoid the calculation of the inverse matrix. To analyze the non-linear relationship between two modalities, Wang et al. [4][5] expanded the CFA to the kernel cross-modal factor analysis (KCFA) using the kernel trick.

By KCFA transforming, we could obtain transformed visual and audio features. Thus, it could be followed by various feature-level and decision-level fusion strategies. It has been successfully used in discrete emotion recognition [4][5]. However, the effectiveness of the KCFA used for continuous dimensional emotion recognition has yet to be investigated, which is what we wanted to do. The implementation of the KCFA involves eigenvalue decomposition of two square matrices, where their dimensions are the number of samples n. The computational complexity of each of the implementation is $O(n^3)$. If n is large, which is the status encountered in the continuous dimensional emotion recognition, then the computation speed will be very slow.

Motivated by the fact that given a precision η, the kernel matrix can be approximated by a low-rank matrix with rank $M \ll n$, we used the incomplete Cholesky decomposition to acquire the low-rank approximation matrices of the kernel matrices involved in the KCFA, and further reduced the computational complexity of the KCFA. In this paper, we provided a theorem and its corollary to describe the properties of the solutions of KCFA, which makes it possible to solve the KCFA problem with incomplete Cholesky decomposition. Based on these discoveries, we presented a novel algorithm named the incomplete Cholesky decomposition based kernel cross factor analysis (ICDKCFA). The overall computational complexity of ICDKCFA is $O(M^2 n)$, which is far lower than that of the original KCFA. Finally, the ICDKCFA was employed for continuous dimensional audiovisual emotion recognition. After the ICDKCFA feature transformation, two basic fusion strategies, i.e., feature-level fusion and decision-level fusion, were explored to combine the transformed visual and audio features for continuous dimensional emotion recognition. Extensive experiments were conducted to evaluate the ICDKCFA approach on the AVEC 2016 Multimodal Affect Recognition Sub-Challenge dataset. The experimental results confirmed that compared to KCFA, our method significantly reduced the computational complexity while maintaining a comparable performance. Moreover, the ICDKCFA method achieved a better performance than other common information fusion methods such as CCA, KCCA，and CFA based fusion methods. Particularly, in the feature level fusion, the ICDKCFA method had a significant advantage in the recognition performance.

2. Kernel Cross-Model Factor Analysis

The kernel cross modal factor analysis (KCFA) was proposed by Wang et al. in [4][5]. It is the kernel version of the linear cross modal factor analysis (CFA), which was proposed by Li et al.[6] to overcome the shortcomings of the canonical correlation analysis (CCA). Therefore, the KCFA can be seen as the development of the KCCA. Moreover, the four transformations are similar in both definition and computation.

2.1 Notations

For the convenience of discussion, we give the meanings of the symbols used in this paper.

Suppose $\{x_i\}_{i=1}^n \in R^p$ and $\{y_i\}_{i=1}^n \in R^q$ are zero mean sample vectors from two modalities respectively, $x' \in R^p, y' \in R^q$ are two arbitrary sample vectors from the two modalities. Let $X=(x_1,x_2,\cdots,x_n)^T$, $Y=(y_1,y_2,\cdots,y_n)^T$, and $C_{xx}=X^TX, C_{yy}=Y^TY$ are the within-sets covariance matrices, $C_{xy}=X^TY$ is the between-sets covariance matrices. Let ψ and φ be two nonlinear functions that map the vectors in R^p and R^q to a higher dimensional space, respectively. Applying ψ, φ to $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, respectively, we can get $\{\psi(x_i)\}_{i=1}^n$ and $\{\varphi(y_i)\}_{i=1}^n$, and let $\Psi=(\psi(x_1),\cdots,\psi(x_n))^T$, $\Phi=(\varphi(x_1),\cdots,\varphi(x_n))^T$. Let $K_x=\Psi\Psi^T=(k_x(x_i,x_j))_{n\times n}$, $K_y=\Phi\Phi^T=(k_y(y_i,y_j))_{n\times n}$ be the kernel matrices of the two set of samples, respectively, where $k_x(\cdot,\cdot)$, $k_y(\cdot,\cdot)$ are the kernel functions corresponding to the maps ψ and φ.

Suppose $\{u_i\}_{i=1}^d \in R^p$, $\{v_i\}_{i=1}^d \in R^q$, $(d \le \min(p,q))$ are two set of vectors, and let $U=(u_1,u_2,\cdots,u_d)$, $V=(v_1,v_2,\cdots,v_d)$. Similarly, suppose $\{a_i\}_{i=1}^d$, $\{b_i\}_{i=1}^d$ are two set of vectors in the image space of ψ and φ, respectively, and $A=(a_1,a_2,\cdots,a_d)^T, B=(b_1,b_2,\cdots,b_d)^T$.

2.2 Kernel Canonical Correlation Analysis

With the notations given above, the objective function and constraint condition of CCA and KCCA are shown in **Table 1**.

Table 1. The objective function and constraint condition of CCA and KCCA.

	CCA	KCCA	
		Original form	Regularization form
Objective function	$\max\limits_{\substack{(u_1,\cdots,u_d),\\(v_1,\cdots,v_d)}} \sum\limits_{i=1}^{d}(u_i)^T C_{xy} v_i$ or $\min\limits_{U,V} \lVert XU-YV \rVert_F$	$\max\limits_{\substack{(u_1,\cdots,u_d),\\(v_1,\cdots,v_d)}} \sum\limits_{i=1}^{d} u_i^T K_x K_y v_i$	$\max\limits_{\substack{(u_1,\cdots,u_d),\\(v_1,\cdots,v_d)}} \sum\limits_{i=1}^{d} u_i^T K_x K_y v_i$
Constraint condition	$U^T C_{xx} U = I,$ $V^T C_{yy} V = I,$ $u_i^T C_{xy} v_j = 0,$ $i,j=1,\cdots,d, i \ne j.$	$U^T K_x^2 U = I,$ $V^T K_y^2 V = I,$ $u_i^T K_x K_y v_j = 0,$ $i,j=1,\cdots,d, i \ne j.$	$U^T((1-\tau)K_x^2+\tau K_x)U = I,$ $V^T((1-\tau)K_y^2+\tau K_y)V = I,$ $u_i^T K_x K_y v_j = 0,$ $i,j=1,\cdots,d, i \ne j,$ $0 \le \tau \le 1$ is the regularization parameter

where $\|M\|_F$ denotes the Frobenius norm of the matrix M and can be expressed as $\|M\|_F = \sqrt{\sum_i \sum_j |m_{ij}|^2}$. When C_{xx} and C_{yy} are invertible, the optimization problem of CCA can be solved as a series of eigenvalue problems:

$$(C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-1} C_{yx} C_{xx}^{-\frac{1}{2}}) \cdot C_{xx}^{\frac{1}{2}} u_i = \lambda_i (C_{xx}^{\frac{1}{2}} u_i), \tag{1}$$

$$v_i = C_{yy}^{-1} C_{yx} u_i \Big/ \lambda_i, \; i = 1, \cdots, d,$$

where λ_i is the i the largest nonzero eigenvalue of $C_{xx}^{-\frac{1}{2}} C_{xy} C_{yy}^{-1} C_{yx} C_{xx}^{-\frac{1}{2}}$ [28].

The optimization problem of KCCA could be solved similarly. In order to overcome the impracticality of the original form of KCCA, a regularization parameter τ was used to construct the regularization form and the incomplete Cholesky decomposition with a precision parameter η could be used to solve the computational issues [28][30].

2.3 Kernel Cross Modal Factor Analysis

The objective function and constraint condition of CFA and KCFA are shown in **Table 2**.

Table 2. The objective function and constraint condition of CFA and KCFA.

	CFA	**KCFA**
Objective function	$\min_{U,V} \|XU - YV\|_F^2$	$\min_{A,B} \|\Psi A - \Phi B\|_F^2$
Constraint condition	$U^T U = I,$ $V^T V = I.$	$A^T A = I,$ $B^T B = I.$

Comparing the objective function and constraint condition of CCA and CFA, we found that the two transformations had the same objective function, and different constraint condition. The constraint condition of CCA $U^T C_{xx} U = I$, $V^T C_{yy} V = I$, $u_i^T C_{xy} v_j = 0$, $i, j = 1, \cdots, d, i \neq j$ is changed as $U^T U = I, V^T V = I$. It is this difference that results in the solution of CFA free from the restriction of the invertibility of C_{xx} and C_{yy}. The optimization problem of CFA is equivalent to

$$\begin{cases} X^T Y = U \Lambda V^T \\ U^T U = I \\ V^T V = I \end{cases} \tag{2}$$

Then, U, V can be obtained by implementing singular value decomposition (SVD) on $X^T Y$. Let the SVD of $X^T Y$ be

$$X^T Y = S_{xy} \cdot \Lambda_{xy} \cdot D_{xy}^T. \tag{3}$$

Then, $U = S_{xy}, V = D_{xy}$, consequently, the representation of X, Y in the transformed domain is [4][5][6]:

$$\hat{X} = XU, \hat{Y} = YV. \tag{4}$$

Similarly, for KCFA, suppose the singular value decomposition of $\Psi^T \Phi$ is [4][5]

$$\Psi^T \Phi = A \Lambda B^T, \tag{5}$$

Then, the representations of Ψ, Φ in the transformed domains are [4][5]

$$\hat{\Psi} = \Psi A, \hat{\Phi} = \Phi B. \quad (6)$$

Furthermore, the representations of $\psi(x')$ and $\varphi(y')$ (the nonlinear function of x' and y') in the transformed domains are [4][5]

$$\hat{\psi}(x')^T = \psi(x')^T A, \hat{\varphi}(y')^T = \varphi(y')^T B. \quad (7)$$

During the process of solving the KCFA, the key problem is to acquire $\hat{\psi}(x'), \hat{\varphi}(y')$ without knowing the explicit expressions of ψ and φ. The problem was solved by the kernel trick in [4][5]. Suppose β is an eigenvector of $K_x K_y$, then $b = \Phi^T \beta / \left\| \Phi^T \beta \right\| = \Phi^T \beta / \sqrt{\beta^T K_y \beta}$ is a column vector of B. Furthermore,

$\varphi(y')^T \cdot \left[(\Phi^T \beta) / \left\| \Phi^T \beta \right\| \right] = \beta / \sqrt{\beta^T K_y \beta} \cdot \left(k_y(y', y_1), \cdots, k_y(y', y_1) \right)^T$ is a component of $\hat{\varphi}(y')$.

Similarly, suppose α is an eigenvector of $K_y K_x$, then $a = \Psi^T \alpha / \sqrt{\alpha^T K_x \alpha}$ is a column vector of A, and, $\alpha / \sqrt{\alpha^T K_x \alpha} \cdot \left(k_x(x', x_1), \cdots, k_x(x', x_1) \right)^T$ is a component of $\hat{\psi}(x')$.

3. Incomplete Cholesky Decomposition Based Kernel Cross Modal Factor Analysis

The algorithm used in [4][5] needs implementation of eigenvalue decomposition on $K_x K_y$ and $K_y K_x$, whose dimensions are the number of the samples n. The complexity of each implementation is $O(n^3)$, which is a serious burden for a large data set. Batch et al. [30] pointed out that given a precision η, the kernel matrix can be approximated by a low-rank matrix with rank $M = h(n/\eta)$. The function $h(t)$ is determined by the kernel function and the decay of the distribution of the data. For the case of Gaussian kernels, when the decay is exponential, $h(t) = O(\log t)$, and when the decay is polynomial (e.g., x^{-d}), $h(t) = O(t^{1/d+\varepsilon})$, where ε is an arbitrary positive real number. These mean that, in general, the low-rank approximation matrix has a rank $M \ll n$. Based on this discovery, we employed the incomplete Cholesky decomposition to acquire the low-rank approximation of the kernel matrices, and present an efficient algorithm for KCFA, which is named as the incomplete Cholesky decomposition based KCFA (ICDKCFA) algorithm in this paper.

3.1 Incomplete Cholesky Decomposition

Incomplete Cholesky decomposition is often used to acquire the low-rank approximation of a kernel matrix. It is the incomplete form of the Cholesky decomposition. The Cholesky decomposition on a kernel matrix (take K_x as an example) can be seen as the dual of Gram-Schmidt orthonormalization on the vectors $\{\psi(x_i)\}_{i=1}^n$. If the rank of K_x is n, using Gram-Schmidt orthonormalization, we can get the standard orthogonal basis $\{q_1, q_2, \cdots, q_n\}$ of the space spanned by $\{\psi(x_i)\}_{i=1}^n$. We denoted $Q = \left[q_1, q_2, \cdots, q_n \right]$, then, $\Psi^T = QG^T$, where G^T is an upper triangular matrix with its ith column is the representation coeffcient of $\psi(x_i)$ by $\{q_1, q_2, \cdots, q_i\}$. Consequently, $K_x = \Psi\Psi^T = GG^T$, this is the Cholesky decomposition of K_x.

The ith element of the diagonal of G^T, G_{ii} is the residual norm of the representation of $\psi(x_i)$ by $\{q_1, q_2, \cdots, q_{i-1}\}$. So, G_{ii} demonstrates how independent $\psi(x_i)$ is from $\{\psi(x_1), \cdots, \psi(x_{i-1})\}$ [31]. In order to acquire the low-rank approximation matrix of K_x and at the same time maintain the most important dimensions of $\{\psi(x_i)\}_{i=1}^n$, we could vary the order of $\{\psi(x_i)\}_{i=1}^n$ processed in Gram-Schmidt orthonormalization, such that the residual norm is always the largest [31]. Furthermore, if the residual norm is below a certain threshold, it will be ignored. Then, we obtained a new upper triangular matrix G_x^T, with its ith column is the representation coefficient of $\psi(x_i)$ by the standard orthogonal vectors acquired before, and perhaps the residual norm is ignored. Then, we got an approximation matrix $G_x G_x^T$ of K_x, with its rank $m_x \ll n$ and the norm of $K_x - G_x G_x^T$ was less than a given value η. The decomposition is referred to as the incomplete Cholesky decomposition, and the corresponding processing on $\{\psi(x_i)\}_{i=1}^n$ is called partial Gram-Schmidt orthonormalization [31].

Table 3 gives the pseudocode of incomplete Cholesky decomposition or dual partial Gram-Schmidt orthogonalization from [28] and [31] (with slightly changed).

Table 3. Incomplete Cholesky decomposition algorithm

Algorithm 1: Pseudocode for Incomplete Cholesky Decomposition /dual partial Gram-Schmidt orthongonalisation
Input $n \times n$ kernel matrix K and a precision parameter η
Initialization: j=0; G is a zeros matrix with the size $n \times n$; d is a vector formed by the diagonal elements of K; a is the maximum element of d, and $I(j+1)$ is the index of a in d. while $\sum_{i=1}^{n} d(i) > \eta$ j=j+1; nu(j)= $\sqrt{a}$; G(:,j)=(K(:,I(j))-G(I(j),:)*G^T)/nu(j); d=d-G(:,j).^2; I(j+1)=argmax(d); a=d(I(j+1)); end while M=j; G=G(:,1:M);
Output: an $n \times M$ lower triangular matrix G with $\left\|K - G \cdot G^T\right\| \le \eta$.

3.2 Incomplete Cholesky Decomposition Based Kernel Cross Modal Factor Analysis

As mentioned in Sub-Section 2.2, during the process of solving the KCFA, the key problem is to acquire $\hat{\psi}(x'), \hat{\varphi}(y')$ without knowing the explicit expressions of ψ and φ. To accomplish this task, the key problem is to solve the singular value decomposition problem of $\Psi^T\Phi$. In this sub-section, we present a theorem and its corollary to describe the property of the right and left singular vector of $\Psi^T\Phi$ corresponding to a nonzero singular value. Based on these discoveries, we could use the incomplete Cholesky decomposition to solve the KCFA, and present the ICDKCFA algorithm. (we still use the notations given in the previous section, but the meanings of U, V, A, B and their column vectors are redefined).

Theorem Let Ψ,Φ be two matrices with sized of $n\times p, n\times q$, respectively, and $K_x=\Psi\Psi^T$, $K_y=\Phi\Phi^T$. Suppose, σ is a nonzero singular value of $\Psi^T\Phi$. Then, u,v are the left and right singular vectors of $\Psi^T\Phi$ corresponding to the singular value σ if and only if there exists an eigenvector α of K_yK_x corresponding to nonzero eigenvalue σ^2 such that $u=\Psi^T\alpha,\ v=(1/\sigma)\cdot\Phi^TK_x\alpha$; Or there exists an eigenvector β of K_xK_y corresponding to nonzero eigenvalue σ^2 such that $v=\Psi^T\beta,\ u=(1/\sigma)\cdot\Phi^TK_y\beta$.

Proof:

Suppose, u,v are the left and right singular vectors of $\Psi^T\Phi$ corresponding to singular value σ, then,

$$\begin{cases} u^T\Psi^T\Phi=\sigma v^T \\ \Psi^T\Phi v=\sigma u \end{cases}. \tag{8}$$

Taking transpose on the first equation of Equation (8), and left multiplying both side by $\Psi^T\Phi$, we have

$$\Psi^T\Phi\cdot\Phi^T\Psi u=\sigma\Psi^T\Phi v=\sigma^2u \tag{9}$$

The second equality is by the second equation of Equation (8). Let $a=\Psi u/\sigma^2$ for the reason of $\sigma\neq 0$. So, the left side of Equation (9) can be rewritten as

$$\Psi^T\Phi\Phi^T\Psi u=\sigma^2\Psi^T\Phi\Phi^Ta=\sigma^2\Psi^TK_ya. \tag{10}$$

So, $\sigma^2u=\sigma^2\Psi^TK_ya$, consequently, $u=\Psi^TK_ya$.

Let $\alpha=K_ya$, then

$$u=\Psi^T\alpha. \tag{11}$$

Left multiplying Equation (9) by Ψ, we have $\Psi\Psi^T\Phi\cdot\Phi^T\Psi u=\sigma^2\Psi u$, hence,

$$K_xK_ya=\sigma^2a. \tag{12}$$

Combining the definition of α with Equation (12) we have

$$K_yK_x\alpha=K_yK_xK_ya=K_y\sigma^2a=\sigma^2\alpha. \tag{13}$$

This shows that, α is an eigenvector of K_yK_x corresponding to the eigenvalue σ^2. From the first equation of Equation (8) and $\sigma\neq 0$, we have

$$v=(1/\sigma)\cdot\Phi^T\Psi u=(1/\sigma)\cdot\Phi^T\Psi\Psi^T\alpha=(1/\sigma)\cdot\Phi^TK_x\alpha. \tag{14}$$

Conversely, suppose α is an eigenvector of K_yK_x corresponding to eigenvalue σ^2, and $u=\Psi^T\alpha$, $v=(1/\sigma)\cdot\Phi^TK_x\alpha$, then,

$$u^T\Psi^T\Phi=\alpha^T\Psi\Psi^T\Phi=(\Phi^TK_x\alpha)^T=\sigma v^T, \tag{15}$$

$$\Psi^T\Phi v=\Psi^T\Phi\cdot(1/\sigma)\cdot\Phi^TK_x\alpha=(1/\sigma)\cdot\Psi^TK_yK_x\alpha=(1/\sigma)\cdot\Psi^T\sigma^2\alpha=\sigma\Psi^T\alpha=\sigma u. \tag{16}$$

The two equations illustrate that u,v are the left and right singular vectors of $\Psi^T\Phi$ corresponding to the singular value σ, respectively.

The proof of the second equivalent conditions is similar to the above.

From the theorem, the algorithm introduced in [4][5] can be perfected by doing eigenvalue decomposition once, and is not further mentioned. We introduce another algorithm here.

From the theorem, we can have the following corollary.

Corollary Suppose u, v are the left and right singular vectors of $\Psi^T\Phi$ corresponding to the nonzero singular value σ, respectively. Then, there exist vectors α, β satisfies, $\begin{cases} u = \Psi^T\alpha \\ v = \Phi^T\beta \end{cases}$.

Suppose,

$$\begin{cases} \Psi^T\Phi = U\Lambda U^T \\ U^TU = I \\ V^TV = I \end{cases} \quad (17)$$

is the SVD of $\Psi^T\Phi$, where $U = [u_1, \cdots, u_T], V = [v_1, \cdots, v_T], \Lambda = diag(\sigma_1, \cdots, \sigma_T)$, and T is the number of the nonzero singular value (multiple singular value calculated on multiplicity), σ_i is the nonzero singular values, and u_i, v_i are the corresponding left and right singular vectors. From the corollary, there exist $A = [\alpha_1, \cdots, \alpha_T], B = [\beta_1, \cdots, \beta_T]$ such that,

$$\begin{cases} U = \Psi^T A \\ V = \Phi^T B \end{cases}. \quad (18)$$

The only work we need to do is to try to seek A and B. Substituting Equation (18) to Equation (17), we have

$$\begin{cases} A^T K_x K_y B = \Lambda \\ A^T K_x A = I \\ B^T K_y B = I \end{cases}. \quad (19)$$

Approximating the kernel matrices K_x and K_y via the incomplete Cholesky decomposition gives

$$\begin{cases} K_x \approx G_x G_x^T \\ K_y \approx G_y G_y^T \end{cases}, \quad (20)$$

where G_x, G_y are lower triangular matrices with size $n \times m_x, n \times m_y$ ($m_x << n$, $m_y << n$). Substituting Equation (20) to Equation (19), we have:

$$\begin{cases} A^T G_x G_x^T G_y G_y^T B = \Lambda \\ A^T G_x G_x^T A = I \\ B^T G_y G_y^T B = I \end{cases}. \quad (21)$$

Computing the SVD of G_x, G_y, we have $G_x = U_x \Sigma_x V_x^T, G_y = U_y \Sigma_y V_y^T$, where Σ_x, Σ_y are diagonal matrixes whose elements of the main diagonal are the nonzero singular values of G_x, G_y. Then

$$\begin{cases} A^T U_x \Sigma_x^2 U_x^T U_y \Sigma_y^2 U_y^T B = \Lambda \\ A^T U_x \Sigma_x^2 U_x^T A = I \\ B^T U_y \Sigma_y^2 U_y^T B = I \end{cases} \quad (22)$$

Let

$$\begin{cases} U_1 = \Sigma_x U_x^T A \\ U_2 = \Sigma_y U_y^T B \end{cases}, \tag{23}$$

Then,

$$\begin{cases} U_1^T \Sigma_x U_x^T U_y \Sigma_y U_2 = \Lambda \\ U_1{}^T U_1 = I \\ U_2{}^T U_2 = I \end{cases}. \tag{24}$$

Let $R = \Sigma_x U_x^T U_y \Sigma_y$, then U_1, U_2 can be obtained by computing the SVD of R, and A, B can be solved from Equation (23)

$$\begin{cases} A = U_x \Sigma_x^{-1} U_1 \\ B = U_y \Sigma_y^{-1} U_2 \end{cases}. \tag{25}$$

Then, the representations of Ψ, Φ in the transformed domain, i.e., Equation (6) can be written as

$$\begin{cases} \hat{\Psi} = \Psi U = \Psi \Psi^T A = K_x A \\ \hat{\Phi} = \Phi V = \Phi \Phi^T B = K_y B \end{cases}. \tag{26}$$

Furthermore, the representations of $\hat{\psi}(x'), \hat{\varphi}(y')$ in the transformed domain, i.e., Equation (7) can be written as

$$\begin{cases} \hat{\psi}(x') = U^T \psi(x') = A^T \Psi \psi(x') = A^T \cdot \left(k_x(x_1, x'), \cdots, k_x(x_n, x')\right)^T \\ \hat{\varphi}(y') = V^T \varphi(y') = B^T \Phi \varphi(y') = B^T \cdot \left(k_y(y_1, y'), \cdots, k_y(y_n, y')\right)^T \end{cases}. \tag{27}$$

Table 4. ICDKCFA algorithm

Algorithm 2: ICDKCFA Algorithm
Input: the matrixes X and Y with their rows are the samples of the visual and audio features respectively, an arbitrary visual feature x' and audio feature y', the parameters of the kernels, the precision parameter of the incomplete Cholesky decomposition η; 1) Using the input kernel parameters, compute the kernel matrixes K_x, K_y of the visual and audio features respectively, and $K_{xx'} = (k_x(x_1, x'), \cdots, k_x(x_n, x'))$, $K_{yy'} = (k_y(y_1, y'), \cdots, k_y(y_n, y'))$; 2) Given the precision parameter η, compute the incomplete Cholesky decomposition of K_x, K_y $K_x \approx G_x G_x^T, K_y \approx G_y G_y^T$; 3) Implement SVD on G_x, G_y respectively, and obtain $G_x = U_x \Sigma_x V_x{}^T, G_y = U_y \Sigma_y V_y{}^T$; 4) Implement SVD on $R = \Sigma_x U_x^T U_y \Sigma_y$, and obtain $R = U_1 \Sigma V_1^T$; 5) Compute the representation coefficients: $A = U_x \Sigma_x^{-1} U_1, B = U_y \Sigma_y^{-1} U_2$; 6) Acquire the representations of the nonlinear function of x', y' in the transformed domain: $x' \to A^T K_{xx'}^T, y' \to B^T K_{yy'}^T$;
Output: the representations of the nonlinear function of x', y' in the transformed domain.

3.3 Performance Analysis Compared to the Original KCFA

As discussed in Sub-Section 2.2, the solving of KCFA is to decompose $\Psi^T\Phi$ as $\Psi^T\Phi = A\Lambda B^T$. So, if a and b are the same columns of A and B, respectively, then

$$a^T\Psi^T\Phi b = \sigma \tag{28}$$

Furthermore, if a and b are the different columns of A and B, then

$$a^T\Psi^T\Phi b = 0 \tag{29}$$

where σ is the corresponding singular value of $\Psi^T\Phi$.

In the implementation of the original KCFA, the eigenvalue decomposition on K_xK_y and K_yK_x was done separately to acquire the eigenvector β of K_xK_y and the eigenvector α of K_yK_x. Based on this, the column of B could be constructed as $b = \Phi^T\beta \Big/ \sqrt{\beta^T K_y \beta}$, and the column of A could be constructed as $a = \Phi^T\alpha \Big/ \sqrt{\alpha^T K_y \alpha}$. This algorithm could not guarantee the establishment of Equations (28) and (29). Even if β and α are selected as the eigenvectors of K_xK_y and K_yK_x corresponding to the same eigenvalue, respectively, when the dimension of the eigenvector space is greater than 1, the establishment of the Equations (28) and (29) are still not guaranteed. This will hurt the performance of the original KCFA. Additionally, in the implementation of the ICDKCFA, Equations (28) and (29) are guaranteed automatically.

The above analysis seems to indicate that the ICDKCFA outperforms the original KCFA in performance. However, in practice, this is not the case. In practice, due to the complexity and noise of the data, the situation that the dimension of the eigenvector space of K_xK_y corresponding to a nonzero eigenvalue is greater than 1 rarely occurs. In addition, in the implementation of ICDKCFA, the low-rank approximation of the kernels will also hurt the performance of the ICDKCFA. Therefore, in practice, the performance of the original KCFA and the ICDKCFA are comparable.

4. Experiments

In order to investigate the effectiveness of the ICDKCFA used for continuous dimensional emotion recognition, we conducted experiments on the datasets used in the AVEC 2016 Multimodal Affect Recognition Sub-Challenge [15], which is a subset of the RECOLA database [32]. The dataset is referred to as the AVEC 2016 dataset. By performing ICDKCFA on the visual features $\{x_i\}_{i=1}^n \in R^p$ and audio features $\{y_i\}_{i=1}^n \in R^q$, we could obtain the transformed features for arbitrary visual or audio feature by Equation (27). Based on these transformed features, we used feature-level and decision-level fusion strategies, respectively, to evaluate the effectiveness of the ICDKCFA for continuous dimensional emotion recognition. Furthermore, we compared the performance of the ICDKCFA based fusion method with other common information fusion methods such as the CCA, KCCA, and CFA based fusion methods. In order to make our results comparable, except for the different feature transformation methods, all of the methods or rules were the same throughout our experiment.

4.1 Dataset and Features

The AVEC 2016 dataset contains multimodal recordings (including audio, video, electro-cardiogram, elector-dermal activity, etc.) from 27 subjects with five minutes for each. All of these signals were synchronously recorded. In this paper, we only used the audio and video data. The videos were recorded at 25 FPS, i.e., the interval between successive frames was 40ms. The ground truth labels were the time-continuous values (i.e., frame by frame , every 40ms) of the arousal and valence dimensions. The recordings in the datasets were evenly divided into three subsets to train, develop, and test the system, respectively. As we had no test labels, our systems were trained on the training set and evaluated on the development set.

The features we used in this paper were the baseline features given in the AVEC 2016 [15]. For the video features, two types of features were given in [15]: appearance and geometric features. The appearance features were derived from LGBP_TOP (Local Gabor Binary Patterns from Three Orthogonal Planes) features. The geometric features originated from 49 facial landmarks. The audio features were extracted by the OPENSMILE toolkit and based on the EGEMAPS (extended Geneva Minimalistic Acoustic Parameter Set) file. All of the features were provided separately for arousal and valence dimensions, and all of the features were given frame by frame (i.e., every 40 ms). As a result, the appearance, geometric and audio feature included 168 features × 7501 frames, 632 features × 7501 frames and 88 features × 7501 frames per file, respectively [15].

In our experiment, we combined the appearance and geometric features with the audio features using various fusion strategies, respectively.

4.2 Performance Metric

Following AVEC 2016 [15], we used the Concordance Correlation Coefficient (CCC) as the performance metric of the continuous dimensional emotion recognition. Suppose, x, y are two series, σ_x^2 , σ_y^2 are the corresponding variance, μ_x , μ_y are their mean value respectively, and ρ is the Pearson correlation coefficient between them. Then, the definition of the CCC of the two series is [15]

$$\rho_c = (2\rho\sigma_x\sigma_y) \Big/ (\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2) . \tag{30}$$

4.3 Data Processing

Before the features were input into a regression model, we had implemented a series of processing including (1) normalizing the data; (2) transforming to the transformation space; (3) delaying in the time; and (4) the dimension selection in the transformation space.

Two normalizations were done on all of the three sets (training, development, testing) in our systems. One is that the features were normalized with a z-score before they were transformed; the other is that they were normalized to $[-1,1]$ before they were input into a regression model.

When transforming the features to the transformation space with various methods, if the kernel method was used, then the RBF kernel $k(x_i, x_j) = \exp(-\|x_i - x_j\| / 2\sigma^2)$ was used as the kernel function, and the kernel width σ was selected from 2^k $(k = 3,5,7,9)$ based on the final recognition performance. Furthermore, if the incomplete Cholesky decomposition was used, the precision was set as $\eta = 0.1$. Additionally, when performing KCCA, the regularization parameter τ was selected from 0.05, 0.2, 0.6, and 1 based on the recognition performance.

Considering the time delay of the annotations to compensate the time reaction of the raters could improve the recognition performance was proved in [33]. The time delay could be done in various ways. One method was to directly drop the first N (N is the delay time measured by frame) labels and the last N feature frames before regression training [33]. Another method was used in [15], where the first N labels were dropped and the last label was duplicated, and the features remained unchanged. The predictions produced by the regressors trained by the above two kinds of training data had N frames ahead of the ground truth. In order to align with the ground truth, the predictions should be processed. There is no need to make the approximation of the training label in the first method, which is necessary for the second method, thus, it is expected to obtain better performance in the first approach. However, the operation process is too complex as it involves the processing of both the labels and features. To obtain better performance meanwhile simplifying the operation process, we implemented time delay by dropping the last N feature frames (the first feature frame was duplicated). Our method avoided the processing of both the training and predicted labels, so it is more efficient. Experiments with the audio feature and the arousal dimension label showed that using the same parameters, the CCC obtained on the development set were 0.787, 0.607, and 0.787 for the first, second, and our method, respectively. This showed that our method was an effective time delay method.

With regard to the visual and audio mono-modal regressions, the best time delays (denoted as t_1, t_2, respectively) were selected from 1.2 s to 4 s by a step 0.4 s based on the final recognition performance [15]. For the bimodal regression, since the fusion strategy may change the mono-modal delay nature slightly, the best pair of delay times was selected from the various pairwise combination of $\{t_1 - 0.4, t_1, t_1 + 0.4\}$ and $\{t_2 - 0.4, t_2, t_2 + 0.4\}$.

In order to obtain better recognition performance, the representation dimension of the features in the transformation space should be optimized. In this paper, the coarse optimal dimension (denoted as m_0) was obtained by increasing the dimension from 10 by a step 20 and there was no improvement over the best performance after two iterations. Furthermore, we selected the best dimension from $\{m_0 - 10, m_0, m_0 + 10\}$ based on the recognition performance. The exception was for the KCCA based fusion method, by experience, a satisfactory performance was achieved when the representation dimension was high. So, the coarse optimal dimension was selected by decreasing the dimension from 300 by a step 20 and there was no improvement over the best performance after two iterations.

4.4 Audiovisual information fusion

After obtaining the transformed features which could be seen as the recognition features, both feature level and decision level fusion could be used to fuse the audiovisual information. For feature level fusion, the recognition features of the two modalities are concatenated as the input of a regression model [4][5]. For the decision level fusion, the predictions of the two single modalities were combined by linear regression [15]. Additionally, the predictions from the audio features were used as the final predictions when the visual features were missing.

4.5 Regression Model

The prediction of the continuous dimensional emotion is a regression problem [15][29]. The Support Vector Regression (SVR) with the RBF kernel was used as our regression model. LibSVM for Matlab Toolbox [34] was used to train the SVR model. During the parameter selection of SVR, for convenience, we set the cost term as $C = 1$. For the RBF kernel width G,

we first selected the best one (denoted as 2^{k_0}) from $2^k, k \in \{-12,-10,\cdots,-2\}$, and then selected the best one from $\{2^{k_0-1}, 2^{k_0}, 2^{k_0+1}\}$ based on the recognition performance.

4.6 Post Processing

In order to improve the prediction performance, a series of post processing were used in our system including median filtering, centering, and scaling [35].

For the median filter, the filter width was optimized by increasing the width from 10 (frames) by a step 10 and there was no improvement over the best performance after two iterations. The centering was realized by computing the bias between the predicted and the ground truth labels, and then subtracting the bias from the prediction. The scaling was realized by computing the ratio of standard deviation of the ground truth and the predicted labels, and then multiplying the prediction by the ratio [35].

4.7 Experimental Setup

In our work, all the recognition systems were trained on the training set, and the parameters were optimized on the development set. When training the regression models, in order to reduce the memory requirement and the computation time, we concatenated the frames of all nine recordings (7501 * 9 frames) in the training set, then extracted one frame out of every 20 frames. The number of frames actually used in the training set was 3375. The frames of all nine recordings (7501 * 9) in the development set were concatenated to evaluate the performance.

When training the ICDKCFA, in order to reduce the memory requirement and the computation time while maintaining the reliability of the results, we omitted the frames where the visual data were missing, then extracted one frame out of every 15 frames.

4.8 Experimental Results

To show the effectiveness of the ICDKCFA, we fused the visual appearance features and visual geometric features with the audio features, respectively and the ICDKCFA were followed by both feature level and decision level fusion. Other common fusion strategies (including direct fusion method, i.e., the original features were used as the recognition features; original KCFA based fusion method; CFA based fusion method; CCA based fusion method; and KCCA based fusion method), and the mono-modal recognition were also used as the contrast methods..

From the results in **Tables 5-7**, it can be seen that a good fusion method is crucial for multimodal emotion recognition. An inappropriate fusion method cannot improve the recognition performance, on the contrary, it might even degrade the performance. Using CCA based fusion as an example, except for the valence dimension recognition results based on the decision level fusion, all the others were inferior to the corresponding mono modal recognition results. This indicates that CCA based fusion was not suited to continuous dimensional emotion recognition. The reason for these results is that the computation of CCA involves the calculation of the inverse of C_{xx} and C_{yy}, when at least one of the two matrices is non-invertible or close to non-invertible, a large error will be produced, and the recognition performance is poor.

Table 5. The recognition results on the development set based on feature-level fusion combined with various transformation methods. The results are measured by the CCC between the predicted and ground truth labels. (a) and (b) are the Visual-appearance Audio Fusion results and the Visual-geometric Audio Fusion results, respectively.

Method	Visual-appearance Audio Fusion		
	Arousal	Valence	Mean
ICDKCFA	0.825	**0.587**	0.706
KCFA	**0.826**	**0.587**	**0.707**
Direct Fusion	0.805	0.552	0.679
CFA	0.810	0.578	0.694
CCA	0.433	0.387	0.410
KCCA	0.811	0.585	0.698

(a)

Method	Visual-geometric Audio Fusion		
	Arousal	Valence	Mean
ICDKCFA	**0.838**	**0.719**	**0.779**
KCFA	**0.838**	0.718	0.778
Direct Fusion	0.804	0.657	0.731
CFA	0.833	0.690	0.762
CCA	0.432	0.387	0.4095
KCCA	0.808	0.675	0.742

（b）

Table 6. The recognition results on the development set based on decision-level fusion combined with various transformation methods. The results are measured by the CCC between the predicted and ground truth labels. (a) and (b) are the Visual-appearance Audio Fusion results and the Visual-geometric Audio Fusion results, respectively.

Method	Visual-appearance Audio Fusion		
	Arousal	Valence	Mean
ICDKCFA	**0.805**	**0.611**	**0.708**
KCFA	**0.805**	**0.611**	**0.708**
Direct Fusion	0.801	0.551	0.676
CFA	0.798	0.576	0.687
CCA	0.772	0.516	0.644
KCCA	0.807	0.598	0.703

（a）

Method	Visual-geometric Audio Fusion		
	Arousal	Valence	Mean
ICDKCFA	0.816	**0.700**	**0.758**
KCFA	**0.817**	0.698	**0.758**
Direct Fusion	0.803	0.679	0.741
CFA	0.817	0.692	0.755
CCA	0.780	0.618	0.699
KCCA	0.810	0.671	0.741

（b）

Table 7. Mono-modal recognition results on the development set. The results are measured by the CCC between the predicted and ground truth labels.

Features	Arousal	Valence	mean
Visual appearance	0.542	0.485	0.514
Visual geometric	0.483	0.579	0.531
Audio	0.787	0.465	0.626

Comparing the recognition results based on direct fusion and the other fusion strategies, it can be seen that with the exception of the CCA based recognition results the latter were all not inferior to the former on average (average over the arousal and valence dimension). This indicates that identifying the relationship between the audio and visual modalities properly could improve the recognition performance in continuous dimensional emotion recognition.

Comparing the recognition results based on the KCCA and ICDKCFA with the CCA and CFA based fusion strategies, respectively, it can be seen that the former were all superior to the latter on average. That is to say, the kernel versions of CCA and CFA were more effective in identifying the correlated information contained in the audio and visual modalities than the corresponding linear versions in the continuous dimension emotion recognition context.

Comparing the recognition results based on the ICDKCFA and CFA with the KCCA and CCA based fusion strategies, respectively, it can be seen that the former had a significant advantage over the latter on average. This indicates that in the continuous dimensional emotion recognition context, the ICDKCFA, CFA had a significant advantage in extracting

the correlated information from the audio and visual modalities over the corresponding KCCA, CCA. Furthermore, the recognition performance of the CFA was comparable or even better than the KCCA. The reason for these phenomena is that the CFA and ICDKCFA involve no computation of inverse matrices, but CCA and KCCA do. When the matrices needed to inverse are non-invertible, a large error will be produced. Meanwhile, the regularization parameter in the KCCA will also produce errors. Therefore, the recognition performance improved by the advantage of analyzing the non-linear relationship of the KCCA is offset by the errors resulting from the regularization parameter and the computation of inverse matrices. Thereby, the recognition performance of the KCCA is only comparable or even worse than that of the CFA.

In conclusion, in the continuous dimensional emotion recognition context, compared with other common fusion strategies, the ICDKCFA based fusion has an obvious advantage.

The dimension at which the optimal performance was obtained for ICDKCFA based fusion method is shown in **Table 8**. In general, the feature level fusion would cause a large dimension, thereby increasing the computational cost and also degrading the recognition performance [1]. So, decision level fusion has been used more than feature level fusion in the literature, although the assumption of independence within different modalities is improper and would cause a loss of the relevant information within different modalities [1]. However, with the ICDKCFA transformation, the dimensions of the recognition features are small. The advantage of the feature level fusion is shown vividly. Except for the valence dimension recognition result for the Visual-appearance Audio fusion, feature level fusion had the better recognition result. For the Visual-appearance Audio fusion on valence dimension, at the dimension 100, the decision level fusion achieved its best performance of 0.611, which is higher than the best performance of the feature level fusion. This again verifies that when the representation dimension is high, the feature level fusion has few advantages.

Table 8. The dimension at which the optimal performance was obtained for ICDKCFA based fusion. (a) and (b) are for the Visual-appearance Audio Fusion and the Visual-geometric Audio Fusion, respectively.

Fusion level	Visual-appearance Audio Fusion	
	Arousal	Valence
Feature level	30	40
Decision level	40	100

（a）

Fusion level	Visual-geometric Audio Fusion	
	Arousal	Valence
Feature level	20	70
Decision level	30	50

（b）

Comparing the recognition performance of the ICDKCFA with the original KCFA, we can see that the two algorithms had a comparable recognition performance. Sometimes, the ICDKCFA performed some less well e.g., the arousal prediction from the Visual-appearance Audio feature level fusion framework. Sometimes, the ICDKCFA performed better than the original KCFA, e.g., the performance of the valence prediction from the Visual-geometric Audio in both the feature level fusion and decision level fusion framework. These phenomena confirmed the analysis in Sub-Section 3.4.

Comparing the recognition performance of the ICDKCFA based fusion approaches with the post processing mentioned in Sub-Section 4.6 with that and without it, we can see that the post processing could effectively improve the performance. The effect of the post processing on the prediction is shown in **Fig. 1**. From **Fig. 1**, it can be seen that, with the median filtering,

the prediction was smoothed, then the random noise was reduced. Combined with the centering and scaling, the predicted and ground truth label obtained a better match.

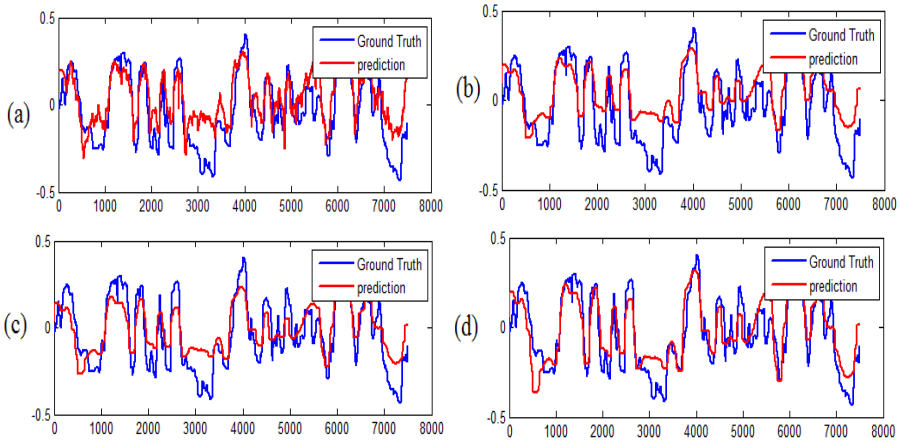

Fig. 1. Effect of post processing on the predicted arousal label of the first recording in development set obtained from the Visual-geometric Audio feature fusion. (a) Prediction without post processing; (b) Prediction with the median filtering; (c) Prediction with the median filtering and centering; and (d) Prediction with median filtering, centering, and scaling.

Table 9. The baseline results of the AVEC 2016 on the development set

Features	Arousal	Valence
Visual appearance	0.483	0.474
Visual geometric	0.379	0.612
Audio	0.796	0.455
Multimodal Fusion	0.820	0.702

Comparing our results with the AVEC 2016 baseline results, which are shown in **Table 9**, it can be seen that although the mono-modal recognition performance of ours was not better than that of the baseline, the ICDKCFA based feature level Visual-geometric Audio fusion results were better than the baseline multimodal fusion results. The unsatisfactory performance of the mono modal recognition showed that the regression model had room for improvement. If a better regression model is carefully selected, the results could be significantly better. Furthermore, only two features were used in our multimodal recognition, however, eight features were used in the AVEC 2016 baseline, if the other modalities could also be used, the recognition results could be better. However, our ICDKCFA model could not implement multimodal analysis of more than two modalities, which will be an interesting work to do in the future.

4.9 Running Time and Computational Complexity Analysis

The above experiments showed that the ICDKCFA had a comparable performance with the original KCFA in the final recognition results, however, it had a faster running speed. In order to show this, we removed the frames with missing data from the visual appearance features and the corresponding frames from the audio features of the arousal dimension, and extracted the first 5000 frames for experiment. The experiment setup and the running time are shown in **Table 10**.

Table 10. The experiment setup and running time for training the representation coefficients

Method	Computer Configuration	Software	RBF kernel width	Incomplete Cholesky Decomposition Precision Parameter	Running Time
ICDKCFA	Intel(R) Core(TM) i7-4790 CPU @ 3.60 GHz	Matlab R2014a	2^7	0.1	15s (appr.)
KCFA				-	350s (appr.)

The results in **Table 10** indicate that ICDKCFA runs much faster than the original KCFA. In fact, the results can also be inferred from the computational complexity analysis. Let $M = \max(m_x, m_y)$, where m_x, m_y are the number of columns of G_x, G_y or the rank of the low-rank approximation matrix of K_x, K_y. Then, $M \ll n$, where n is the number of samples.

In practice, using the incomplete Cholesky decomposition to approximate the kernel matrix, the full calculation of the matrix is actually avoided [30]. To acquire the low-rank approximation of K_x and K_y from the sample matrices X and Y, the overall complexity was $2 \times O(M^2 n) = O(M^2 n)$. The sizes of $G_x (G_y)$ did not exceed $n \times M$, then the implementations of SVD on G_x and G_y have the cost of $O(M^2 n)$. The size of $R = \Sigma_x U_x^T U_y \Sigma_y$ is not exceed $M \times M$, then the complexity of performing SVD on R is $O(M^3)$. The size of $U_x (U_y)$ does not exceed $n \times M$, $\Sigma_x (\Sigma_y)$ is a diagonal matrix with size not exceeding $M \times M$, and the size of $U_1 (U_2)$ does not exceed $M \times M$, consequently, the acquisitions of A and B have the cost of $O(M^2 n)$. Finally, the acquisitions of the representation of x', y' have the cost of $O(Mn)$. In conclusion, the overall complexity of the ICDKCFA is $O(M^2 n)$, which is far lower than $O(n^3)$, which is the complexity of the original KCFA.

5. Conclusions

In this paper, a novel ICDKCFA was presented to allow the KCFA to handle the large amount of data encountered in continuous dimensional emotion recognition. Based on the presented ICDKCFA, the visual and audio features were fused to recognize the continuous dimensional emotion. As shown in the experiment and analysis, the presented ICDKCFA performed faster than the original KCFA while maintaining a comparable performance. Compared with other common fusion methods, the ICDKCFA performed the best. The performance result of the ICDKCFA based audiovisual recognition was even superior to the multimodal (four modalities, eight features) recognition. The SVR model was used as the regression model in this paper. If the regression model is selected carefully, the performance would be better.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61501249, No. 61071167 and No. 61471206, the Key Research and Development

Program of Jiangsu Province under Grant No. BE2016775, the Natural Science Foundation of Jiangsu Province under Grant No. BK20150855 and No. BK20141428, the Natural Science Foundation for Jiangsu Higher Education Institutions under Grant No. 15KJB510022, the Project funded by China Postdoctoral Science Foundation under Grant 2018M63234 and the Postgraduate Innovation Project of Jiangsu Province under Grant No. KYLX15_0827 and No. KYLX16_0660.

References

[1] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," *IEEE Transactions on Pattern Analysis amd Machine Inteligencel*, Vol. 31, no. 1, pp. 39-58, Jan. 2009. Article (CrossRef Link)

[2] J. Yan, W. Zheng, M. Xin, and J. Yan, "Integrating facial expression and body gesture in videos for emotion recognition," *IEICE Transactions on Information and Systerms*, vol. E97.D, no. 3, pp. 610-613, Mar. 2014. Article (CrossRef Link)

[3] J. Yan, W. Zheng, Q. X, G. L, H. Li, and B. W, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319-1329, Jul. 2016. Article (CrossRef Link)

[4] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Audiovisual emotion recognition via cross-modal association in kernel space," in *Proc. of IEEE International Conference on Multimedia & Expo*, pp. 1-6, Jul. 2011. Article (CrossRef Link)

[5] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transations on Multimedia*, vol. 14, no. 3, pp. 597-607, Jun. 2012. Article (CrossRef Link)

[6] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. of 11th ACM International Conference on Multimedia*, pp. 604-611, Nov. 2003. Article (CrossRef Link)

[7] C. H. Wu, J. C. Lin, and W.L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *Apsipa Transactions on Signal & Information Processing*, vol. 3, pp. 1-18, 2014. Article (CrossRef Link)

[8] C. Vinola, and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *Electronic Letters on Computer Vision & Image Analysis*, vol. 14, no. 2, pp. 24-44, 2015. Article (CrossRef Link)

[9] L. Pang, S. Zhu, and C. W. Ngo, "Deep Multimodal Learning for Affective Analysis and Retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008-2020, Nov. 2015. Article (CrossRef Link)

[10] C. H. Wu, J. C. Lin, and W. L. Wei, "Two-level hierarchical alignment of semi-coupled HMM-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880-1895, Dec. 2013. Article (CrossRef Link)

[11] B.Schuller, M. Valstar, F. Eyben, R. Cwie, and M. Pantic, "AVEC 2012 - the continuous audio/visual emotion challenge," in *Proc. of 14th ACM International Conference on Multimodal Interaction*, pp. 449-456, Oct. 2012. Article (CrossRef Link)

[12] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge," in *Proc. of the 3rd ACM International international worshop on Audio/visusl emotion challenge*, pp. 3-10, Oct. 2013. Article (CrossRef Link)

[13] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014-3D dimensional affect and depression recognition challenge," in *Proc. of 4th International Workshop on Audio/Visual Emotiona Challenge*, pp. 3-10, Nov. 2014. Article (CrossRef Link)

[14] F. Ringeval, B. Schuller, M. Valster, S. Jaiswal, E. Marchi, D. Lalanne R. Cowie, and M. Pantic, "AV+EC 2015-the first affect recognition challenge bridging across audio, video, and

physiological data," in *proc. of 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 3-8, Oct. 2015. Article (CrossRef Link)

[15] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 – depression, mood, and emotion recognition workshop and challenge," in *Proc. of 6th International Workshaop on Audio/Visual Challenge*, pp. 3-10, Oct. 2016. Article (CrossRef Link)

[16] F. Eyben, M. Wöllmer, M.F. Valstar, H.Gunes, B.Schuller, and M.Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *Proc. of IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 322-329, Mar. 2011. Article (CrossRef Link)

[17] C. Soladié, H. Salam, N. Stoiber, and R. Seguier, "Continuous facial expression representation for multimodal emotion detection," *International Journal of Advanced Computer Science*, vol. 3, no. 5, pp. 202-216, May. 2013. Article (CrossRef Link)

[18] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. of 5th International Workshop on Audo/Visual Emotion Challenge*, pp. 65-72, Oct. 2015. Article (CrossRef Link)

[19] S. Chen, and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. of 5th International Workshop on Audo/Visual Emotion Challenge*, pp. 49-56, Oct. 2015. Article (CrossRef Link)

[20] A. Sayedelahl, R. Araujo, and M. S. Kamel, "Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversation," in *Proc. of IEEE International Conference on Multimedia and Expo Workshops*, pp. 1-6, Oct. 2013. Article (CrossRef Link)

[21] Y. Falinie, A. Gaus, H. Meng, A. Jan, F. Zhang, and S. Turabzadeh, "Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and PLS regression," in *Proc. of IEEE International Conference and Workshop on Automatic Face and Gesture Recognition*, pp. 1-6, Oct. 2015. Article (CrossRef Link)

[22] M. Kächele, M. Schels, P. Thiam, and F. Schwenker, "Fusion mappings for multimodal affect recognition," in *Proc. of IEEE Symposium Series on Computional Intelligence*, pp. 307-313, Jan. 2015. Article(CrossRef Link)

[23] P. Cardinal, M. Dehak, A. Lameiras, J. Alam, and P. Boucher, "ETS system for AV+EC 2015 challenge," in *Proc. of 5th International Workshop on Audo/Visual Emotion Challenge*, pp. 17-23, Oct. 2015. Article (CrossRef Link)

[24] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proc. of 14th ACM International Conference on Multimodal Interaction*, pp. 501-508, Oct. 2012. Article (CrossRef Link)

[25] C. Soladié. H. Salam, C. Pelachaud, N. Stoiber, and R. Séguier, "A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection," in *Proc. of 14th ACM International Conference on Multimodal Interaction*, pp. 493-500, Oct. 2012. Article(CrossRef Link)

[26] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion state using body language and prosodic cues," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2288-2291, Jul. 2011. Article (CrossRef Link)

[27] L. Tian, J. D. Moore, and C. Lai, "Recognizing emotions in dialogues with acoustic and lexical features," in *Proc. of IEEE International Conference on Affective Computing and Intelligent Interaction*, pp. 737-742, Dec. 2015. Article (CrossRef Link)

[28] D. R.Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning methods," *Neural Computation*, vol. 16, no. 12, pp.2639-2664, Dec. 2004. Article (CrossRef Link)

[29] Y. Song, L. P. Morency, and R. Davis, "Learning a sparse codebook of facial and body microexpressions for emotion recognition," in *Proc. of 15th ACM on International Conference on Multimodal Interaction*, pp. 237-244, Dec. 2013. Article (CrossRef Link)

[30] F. R. Bach, and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1-48, Jul. 2002. Article(CrossRef Link)

[31] J. Shawe-Taylor and N. Cristianini, *Kernel Method for Pattern Analysis*, Cambridge, New York, 2004. Article (CrossRef Link)

[32] F. Ringeval, A. Sondergger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. of IEEE International Conference and Workshop on Automatic Face and Gesture Recognition*, pp. 1-8, Jul. 2013. Article (CrossRef Link)

[33] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. of 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 41-48, Oct. 2015. Article (CrossRef Link)

[34] C. C. Chang, and C. J. Lin, "LibSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Apr. 2011 Article (CrossRef Link)

[35] G. Tigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," In *Proc.of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5200-5203, Mar. 2016. Article(CrossRef Link)

Xia Li received the B.S. degree in Mathematics and Applied Mathematics from Qufu Normal University and the M.S. degree in Applied Mathematics from Nanjing University, in 2002 and 2005, respectively. She is currently pursuing the Ph.D. degree at the College of Telecommunication and Information Engineering in Nanjing University of Posts and Telecommunications. Her current research interests include pattern recognition, affective computing, machine learning and computer vision.

Guanming Lu received the B.E. degree in radio engineering and the M.S. degree in communication and electronic systems from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 1985 and 1988, respectively, and the Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is currently a Professor with the College of Communication and Information Engineering, NUPT. His current research interests include image processing, affective computing, and machine learning.

Jingjie Yan received the B.E. degree in electronic science and technology and the M.S. degree in signal and information processing from the China University of Mining and Technology, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in signal and information processing from Southeast University, Nanjing, China, in 2014. Since January 2015, he has been with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, as a Lecturer. His current research interests include pattern recognition, affective computing, computer vision, and machine learning.

Haibo Li received the B.E. degree in wireless engineering and the M.S. degree in communication and electronic systems from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 1985 and 1988, respectively, and the Ph.D. degree in information theory in 1993 from Linöping University, Linöping, Sweden. He is a Professor of Innovative Media Technology with the KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include mainly media signal processing, including facial and hand gesture recognition and invisible interaction technology

Zhengyan Zhang received the B.S. degree in electronic information engineering and the M.S. degree in signal and information processing from Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, Chian, in 2004 and 2007, respectively. He is currently pursuing the Ph.D. degree at the College of Telecommunications an Information Engineering in Nanjing University of Posts and Telecommunications. His current reearch interests include pattern recogition, machine learning and computer vision.

Ning Sun received the B.S., M.S. and Ph.D. degrees from Guilin University of Electronic Technology, Nanjing Institute of Electronic Technology and Southeast University, in 2000, 2004 and 2007, respectively. Since 2012, he has been with Nanjing University of Posts and Telecommunications, Nanjing, China, where he is currently an Associate Professor in the Engineering Research Center of Wide Band Wireless Communication Technology, Ministry of Education. His current research interests include deep learning, pattern recognition and embedded platform based video analysis.

Shipeng Xie received the B.S. degree in Mathematical Sciences in June 2003 from Anhui University, and the Ph.D. degree in Computer Science and Engineering in July 2012, from Southeast University, Nanjing, China. From June 2006 to February 2013, he served as an officer in Anhui University. He is currently an Associate Professor in Nanjing University of Posts and Telecommunications, working in the fields of computational imaging and computer vision harnessing both variational and learning-based methods.