IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 29, 2020, accepted March 16, 2020, date of publication March 18, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981760

Outlier Processing in Multimodal Emotion Recognition

GE ZHANG[1], TIANXIANG LUO[2], WITOLD PEDRYCZ[3], (Fellow, IEEE), MOHAMMED A. EL-MELIGY[4], MOHAMED ABDEL FATTAH SHARAF[5], AND ZHIWU LI[1,6], (Fellow, IEEE)

[1]Institute of Systems Engineering, Macau University of Science and Technology, Macau, China
[2]Department of Electronic Engineering, Xidian University, Xi'an 710071, China
[3]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T5J 4P6, Canada
[4]Advanced Manufacturing Institute, King Saud University, Riyadh 11421, Saudi Arabia
[5]Industrial Engineering Department, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia
[6]Department of Mechanical and Electrical Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Zhiwu Li (zhwli@xidian.edu.cn)

This work was supported by the Macau University of Science and Technology under the Faculty Research Grant FRG-19-005-MISE.

ABSTRACT Automatic emotion recognition plays a key role in human-computer interactions. Multimodal emotion recognition has attracted much attention in recent years. When multimodalities are used, different modalities interact with each other and the obtained results tend to be accurate in general. However, there are also cases of unimodal anomalies. Most of the existing studies do not take into account the existence of outliers in the multimodality, which leads to low accuracy of the prediction results. This paper proposes fuzzy weighted support vector machine for regression (FWSVR) to deal with outliers and prediction errors. We design an automatic affective recognition model structure to analyze continuous dimension emotions based on multimodality (audio and visual). The LIRIS-ACCEDE database is used in this work. Experimental results indicate that the concordance correlation coefficient (CCC) is 0.9456 for arousal and 0.9183 for valence on the test set. The fusion result obtained when using fuzzy weighting is much better than the direct fusion one.

INDEX TERMS Multimodal emotion recognition, fuzzy weighted support vector machine for regression (FWSVR), continuous dimension emotional space, outlier processing.

I. INTRODUCTION

Artificial intelligence will greatly change our life in the coming years. People are beginning to realize the importance of human-computer interactions. However, computers are limited in the cognitive and response levels of human emotions, leading to computers unable to make the right instructions based on human emotions. Once machines understand human emotions, artificial intelligence will rise to a new level. In recent years, researchers continuously carry out scientific research and obtain rich theoretical results. Tran and Cambria [2] perform real-time multimodal sentiment analysis, collect and process the massive information of multimedia websites, such as, Facebook and YouTube. In addition, emotion recognition has some applications in many fields, such as education, medicine, industry, computerized psychological counseling, therapy [1] and stock market [3].

According to different theories of psychology, emotions are divided into discrete categorical emotions and continuous dimension emotions. American psychologist Ekman [4] classifies basic emotions into six categories, including happiness, sadness, fear, surprise, anger and disgust. This emotion classification can roughly cover human emotional changes. Each of emotions is independent of the others. The continuous dimension emotion analysis [5] indicates that in fact emotions are interrelated rather than being independent of each other. Many researchers have divided continuous dimension space, and established multiple dimension representation methods.

Russell proposes a two-dimensional bipolar space (i.e. arousal and valence) to represent the cognitive of emotions [5]. The horizontal dimension is the valence dimension that represents pleasure-displeasure. The vertical dimension is arousal dimension that stands for arousal-sleep.

The associate editor coordinating the review of this manuscript and approving it for publication was Le Hoang Son.

Russell organizes emotions in a circular arrangement, in which all emotions are represented at different quadrants. The division of circular region is based on the ambiguity of emotions [11]. In Russell's work, it is shown that emotions are not independent, but highly correlated with each other [6]. In this paper, we recognize video emotions in a continuous dimension (valence-arousal) space.

In recent years, researchers often use one or multimodal features to analyze emotions in the field of human emotion recognition (e.g., audio, visual, facial expression, body gesture and physiological signals) [24]. Many people believe that the information in audio can express the emotion of video content. Cambria *et al.* [8] propose sentic blending which enables the continuous interpretation of semantics and sentics. In [7] the authors show that the audio feature has good recognisability in arousal dimension. The visual feature is superior in recognizing in valence emotion space. The advantage of emotional analysis by means of audio and visual fusion lies in the different expressive abilities of audio and visual content in the continuous dimension of emotional space. The disadvantage is that there exist outliers in multimodality emotion recognition. Since visual and audio are divided into frames, the features of a single frame are extracted. A unimodal emotion analysis may produce misleading information. For instance, the characters may try to conceal their emotions, or display abnormal facial expressions and body movements due to extreme sadness. In this case, the unimodal cannot correctly recognize an emotion [10]. The fusion of multimodalities may reduce the recognition accuracy due to outliers [9]. In this paper, the unimodal is used to predict emotion value, and then we can obtain the discrepancies between audio and visual features. However, due to the existence of outliers in the unimodal recognition process, this paper proposes a FWSVR algorithm to blur the outliers in fusion layer. It will reduce the influence of the overall system of outliers.

The major contributions made in this study are presented as follows: (1) In addition to focusing on the complementary relationship of multimodality, this paper also pays attention to the negative effects between visual and audio signals, and provides an adaptive solution. The outliers of a certain modality will affect the accuracy of the system identification, which has a great negative impact on the emotion recognition. (2) FWSVR is proposed to alleviate the above-mentioned problem of modal relationship to solve the interference of outliers and noise on the system during the regression process. (3) We establish a multilevel regression model for emotion prediction in multimodal continuous dimensions. Based on the unimodal sentiment prediction of audio or visual signal, a multimodal emotion recognition method is proposed, and a multi-layer emotion recognition model structure is established. This method of establishing a degree of membership not only considers the similarity between audio and visual prediction results and actual values, but also blurs the part that deviates from the actual value, which improves the recognition accuracy.

The rest of the paper is organized as follows. Section 2 reviews related works, including the method of continuous multimodal recognition and related technical innovation. The FWSVR and a multimodal regression structure are proposed in Section 3. In Section 4 the proposed model is tested on a database. Section 5 concludes this paper and discusses future work.

II. RELATED WORK

Multimodal features have been proposed to recognize human emotions. Commonly used modalities are visual [16], audio [18], text [14], facial action [15], posed versus spontaneous expression and multiple physiological parameters [13]. In the multimodal emotion recognition, audio-visual contents are most studied [12], [19]. Zhang *et al.* [20] propose to bridge the emotional gap based on a multimodal deep convolution neural network, which fuses the audio and visual cues in a deep model. A strength model is designed for real-world automatic continuous affect recognition from audio-visual signals by Han *et al.* [21], where support vector machine for regression and long short-term memory recurrent neural networks are used to build the strength modelling. Ringeval *et al.* [22] add physiological parameters (EGG, EDA), audio and visual signals, and use deep learning algorithms to predict emotions. Ranganathan *et al.* [23] use the emoFBVP database of multimodal (face, body gesture, voice and physiological signals) and propose convolutional deep belief network models that learn salient multimodal features of expressions of emotions.

In the real world, individuals tend to have partial or mixed sentiments about an opinion target [27]. Due to the complexity and uncertainty of human emotions and the fact that the emotional features extracted in audio and visual have certain nonlinear mapping relationship with the fuzzy space of emotion, fuzzy logic is applied to continuous dimension emotion recognition. Russell proposes a continuous dimension emotion space, where the valence-arousal space is organized in a circular order and fuzzy sets are employed to blur their boundary. Chaturvedi *et al.* [27] propose the combined model of deep convolutional neural networks and fuzzy logic to predict the degree of a particular emotion. Sun *et al.* [28] propose fuzzy C-means to divide continuous emotion intervals into seven parts and then the membership degree of each interval is determined, in order to conduct emotion analysis of video contents. Ioannou *et al.* [29] present a fuzzy neural network system based on particular rules to predict continuous emotion space by facial expression. Rani *et al.* [30] propose a real-time anxiety detection method based on fuzzy logic and regression tree.

Multimodal fusion is important because of the complex interrelationships between multimodal information. In a fusion strategy, the fusion method of feature-level and decision-level is frequently used [22]. Feature-level fusion is completed at the feature data level. Multimodal features are fused according to some rules of addition or some weighted factors. The fusion feature is used as input to the recognition

algorithm [24]. Feature-level fusion allows a classifier to take advantage of the complementarity among signals. Some experiments indicate that the result of feature-level fusion is better than that of unimodal recognition [25]. Decision-level fusion takes the features extracted from different modalities as the input of the emotion recognition algorithm and then fuses the output of the algorithm according to the strategies of principal component analysis or linear discriminant analysis.

Ringevala *et al.* find that the decision-level fusion is superior to the feature-level fusion [22]. The advantage of using the decision-level fusion is that the amount of data for fusion is reduced to a large extent. It can effectively solve the problem of abnormal situation in a particular modality. In [26], the authors achieve great recognition effects using the decision-level fusion. In our paper, we find that the results predicted by different modalities will affect the fusion result. Hence a novel fusion method, namely FWSVR, in the decision-level fusion is proposed.

However, before proceeding decision-level fusion, each modal feature passes through the classifier alone. After predicting the emotional features of each modal, a decision-level fusion strategy is used to fuse the prediction results. In this process, the interrelationship between modalities becomes especially important. If two modalities accurately express the same emotion and the classifier prediction is accurate, the fused predicted value will be more accurate. If a modality does not express the emotion correctly at that moment or intentionally misleads the emotional information at that moment, or has a problem in unimodal prediction, the result of the fusion will reduce the accuracy.

The above problems arising at the decision-level fusion can be attributed to the existence of processing outliers. Fuzzy logic can effectively deal with uncertainty, complexity and ill-posed problems. It has been proposed in the literature that a support vector machine (SVM) endowed with a suitable membership function can reduce the influence of outliers on prediction results and solve the above problems [32]. The SVM is first proposed by Vapnik [40]. It is based on the idea of structural risk minimization and used for classification [34]. A support vector machine for regression (SVR) is a derivative of SVM, which is used for regression analysis in the continuous space. SVM and SVR are widely used because of their power in predictions. In conventional SVR, the training process is very sensitive to noise or outliers in the training samples. However, the limitations can be overcome by introducing fuzzy logic [33].

Aydilek and Arslan [35] dwell upon a hybrid method, using fuzzy C-means with SVR and a genetic algorithm to reduce the impact of missing values on prediction. Jiang *et al.* [32] report a new method for calculating fuzzy membership degree. A two-layer fuzzy multiple random forest is proposed for speech emotion recognition in [50]. Differences between different categories of people were taken into account. Juang and Hsieh [36] develop the Takagi–Sugeno fuzzy system-SVR, which improves generalization capability. Furthermore an interval type-2 fuzzy-neural network with SVR is proposed to solve noise and outlier problems in [37]. Chen *et al.* [38] propose a three-layer weighted fuzzy SVR model that includes adjusted weighted kernel fuzzy C-means for emotion data clusting to determine the customer's preference for alcohol. Liu *et al.* [17] touch upon a local least squares SVR, where the Gustafson-Kessel clusting algorithm is used to reduce the size of subsets.

Fuzzy SVR has made great progress and solved the problem of outliers in many fields. However, in the multimodal sentiment analysis, the problem of outliers between modalities is not well noticed. To this end, we establish a regression network structure and propose a decision-level fusion structure between modalities (as shown in Fig. 1). A new membership function is formulated, which is determined based on the residual of the unimodal predicted value and the actual value.

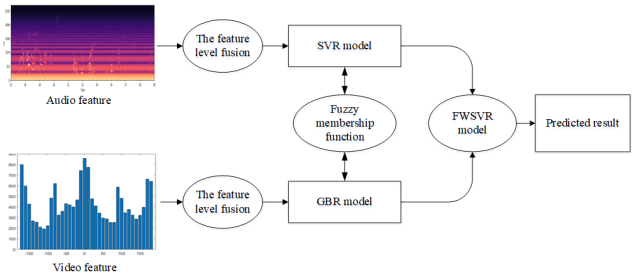


FIGURE 1. Regression network model structure with membership function.

The proposed regression network consists of three layers. The first layer is feature processing of audio and visual signals. It is mainly responsible for audio and visual feature extraction, preprocessing, dimensionality reduction and feature-level fusion. After that the second layer is a single modal regression layer that is selected according to the fitness of algorithm of different models. The second layer consists of SVR and GBR to predict continuous dimension emotions by visual and audio modality. Finally, the third layer is fusion layer. We propose a new fusion method, namely FWSVR, to process the outlier problem. The fuzzy membership function is selected according to the residual error of modal prediction such that the outliers have a small weight, which reduces the influence of outliers on emotion prediction.

In this fusion layer, the fuzzy membership function comes from the prediction result of the SVR and the GBR. In other words, we employ the prediction results of the SVR, the GBR and the actual label value to obtain the residual error, which is used as the standard for establishing the fuzzy membership function. The structure of the regression model is considered and improved, which includes decision level fusion and feature level fusion. It is concatenated with the original feature space utilized as the basis for regression analysis in a regression network.

III. THE FRAMEWORK STRUCTURE OF THE PREDICTION MODEL

A. MULTIMODALITY REGRESSION MODEL

1) SVR MODEL

SVR is a regression method based on SVM. It tries to find the optimal regression hyperplane such that most of the training samples lie within a margin ε around this hyperplane [40]. Applying SVR for a regression task, the target is to optimise the generalisation bounds for regression in the high-dimension feature space by using an ε-insensitive loss function which measures the cost of the predicted value. Suppose that we have a database, including the feature data x_i $(i = 1, 2, \ldots, n)$, $x_i \in R$, and the target data y_i $(i = 1, 2, \ldots, n)$, $y_i \in R$. In the ε-insensitive model, the aim is to find a function $f(x)$ that has at most a deviation ε from the actually obtained targets y_i for all the training data, and at the same time it is as flat as possible [40], [41]. In other words, when $f(x) = y_i$, the deviation is zero, otherwise, we want the deviation to be less than ε, i.e., $|f(x) - y_i| < \varepsilon$. We begin by describing the case of linear functions $f(x)$

$$f(x) = \langle w, x \rangle + b, w \in X, b \in R \tag{1}$$

where w and b are determined parameters, which are weights and bias, respectively. $\langle w, x \rangle$ is the inter product of w and x.

This problem can be formulated as a convex optimization task:

$$\begin{aligned} \min \ & \frac{1}{2}||w||^2 + C\sum_{i=1}^{\iota}(\xi_i + \xi_i^*) \\ \text{s.t.} \ & y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \\ & \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{2}$$

where the parameter ε specifies the ε-tube within which no penalty is associated in the training loss function with points predicted within a distance ε from the actual value. The parameter C describes penalty. Through the above convex optimization problems, we can reasonably predict the value of y and fix it in the range of ε. The ε-insensitive loss function $|\xi|_\varepsilon$ is described by

$$|\xi|_\varepsilon : \begin{cases} 0, & |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & \text{otherwise} \end{cases} \tag{3}$$

2) GRADIENT BOOSTED REGRESSION (GBR) MODEL

GBR is an integrated model that combines multiple weak classifiers. Its basic model is a binary tree structure that is a weak classifier. The accuracy is low. Through the integration of multiple weak classifiers, the gradient descent of residual error is carried out to obtain the optimal fitting state [31].

Suppose that the train data set is $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \subseteq R$, $y_i \subseteq R$, and x_i is the input data and y_i is the true data. We use the Huber function as the loss function. This loss function can be expressed as

$$L_\delta(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2, & |y_i - \hat{y}_i| \le \delta \\ \delta|y_i - \hat{y}_i| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{y}_i$ is the predicted value from visual and audio prediction, and δ is shown in (5).

In the Huber loss function, the mean square error is used to calculate δ by

$$\delta = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

The Huber loss function is used for robust regression. It is less sensitive to outliers than the squared error loss.

The GBR algorithm is described as a sequence of steps:

step 1: Initialization.

$$c_{mj} = \arg\min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

step 2: The model uses the value of the negative gradient of the loss function as an estimate of the residual, i.e.,

$$r = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

step 3: The input space is divided into J_m disjoint areas, which can be regarded as $R_{m1}, R_{m2}, \ldots, R_{mj}, j = 1, 2, \ldots, J$.

step 4: For $j = 1, 2, \ldots, J$,

$$c_{mj} = \arg\min \sum_{x \in R_{mj}} L(y_i, f_m - 1(x_i) + c)$$

step 5: Update

$$f(x) = \sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I, (x \in R_{mj})$$

step 6: Assume that the constant value of the output is c_j within each region. The gradient boosting regression function $f(x)$ can be shown as

$$f(x) = \sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj} I, (x \in R_{mj})$$

B. FUSION MODEL – FWSVR

For regression prediction using the SVR model, three problems have been considered. First, noise can lead to false regression. Second, the closer the sample is to the predicted point, the greater its impact on the regression. Third, the sample point farther away from the regression line, the greater the error of regression prediction [39]. Motivated by those problems, this paper proposes the FWSVR model that adds the fuzzy membership function. This is equivalent to providing adaptive weights for sample points. When the residual is large, the value of membership is small. When the residual is small, the membership is large. Thus, it reduces the negative impact of outliers on prediction. The fuzzy weighting process

of the SVR model can reduce the impact of noise and perform fuzzy weighting on the sample according to the distance, thereby improving the prediction accuracy. Next, we make a detailed description about the idea and formulation of an FWSVR model.

The FWSVR model is modeled by minimizing the following constrained cost function

$$\begin{aligned} \min\ & \frac{1}{2}||w||^2 + C\sum_{i=1}^{\iota}(\xi_i + \xi_i^*) * \mu_i \\ \text{s.t.}\ & y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ & \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \tag{6}$$

where the penalty parameter C is a regularization constant, controlling a compromise between maximizing the margin and minimizing the number of training set errors. The variable y_i is the actual affective value. Through the above convex optimization problem, we can reasonably predict the value of y_i and fix it in the range of ε. The variable μ_i is the fuzzy membership function as (17). The variable x_i is input data, serving the prediction result of GBR and SVR models through audio and visual modalities. The parameters ξ_i and ξ_i^* represent upper and lower constrains on the outputs of model, respectively. The ε-insensitive loss function $|\xi|_\varepsilon$ is described by (3).

We use the Lagrange multiplier method to solve the constraint optimization problem in (7), while the parameter α_i, α_i^*, γ_i and γ_i^* are multipliers.

$$\begin{aligned} L = & \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)\mu_i - \sum_{i=1}^{N}\gamma_i\xi_i + \gamma_i^*\xi_i^* \\ & - \sum_{i=1}^{N}\alpha_i(\varepsilon_i + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{N}\alpha_i^*(\varepsilon_i + \xi_i^* + y_i - \langle w, x_i\rangle - b) \end{aligned} \tag{7}$$

Next, we optimize (7) that must meet the conditions as stated in (8), (9), (10) and (11). To formulate the corresponding dual problem of (7), we use the substitute conditions of (8), (9), (10), (11) in (7).

$$\frac{\partial L}{\partial w} = w_i - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)x_i = 0 \tag{8}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\alpha_i^* = 0 \tag{9}$$

$$\frac{\partial L}{\partial \xi} = C\mu_i - \alpha_i - \gamma_i = 0 \tag{10}$$

$$\frac{\partial L}{\partial \xi^*} = C\mu_i - \alpha_i^* - \gamma_i^* = 0 \tag{11}$$

We can find the new optimization equations in (12) through these constraints.

$$\begin{aligned} \max\ & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i \\ & + \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\varepsilon_i \\ \text{s.t.}\ & \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ & 0 < \alpha_i < \mu_i C \\ & 0 < \alpha_i^* < \mu_i C \end{aligned} \tag{12}$$

Then, we use the Karush-Kuhn-Tucker (KKT) condition to solve the quadratic programming problem as shown in (13). We know that $\alpha_i\alpha_i^* = 0$, and the parameters α_i and α_i^* cannot be non-zero at the same time. Only when $\alpha_i = \mu_i C$ and $\alpha_i^* = \mu_i C$, the error between the estimated value $f(x)$ and the actual value y_i is more than ε. Thus, we can obtain the bias as (14).

$$\begin{aligned} & \alpha_i(\varepsilon_i + \xi_i - y_i + wx_i + b) = 0 \\ & \alpha_i^*(\varepsilon_i^* + \xi_i^* - wx_i - b + y_i) = 0 \\ & \gamma_i\xi_i = 0 \\ & \gamma_i^*\xi_i = 0 \end{aligned} \tag{13}$$

$$b_i = \begin{cases} y_i - wx_i - \varepsilon_i, & 0 < \alpha_i < \mu_i C \\ y_i - wx_i + \varepsilon_i, & 0 < \alpha_i^* < \mu_i C \end{cases} \tag{14}$$

Then, α_i and α_i^* are constrained by the KKT condition again, and the fitting interval is continuously decided by the sample points. Finally, the function $f(x)$ can be shown as

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)x_i + b \tag{15}$$

Considering the interaction between the two modalities, fuzzy weighted adjustment is made for the SVR results of the decision-level fusion. Based on the relationship between the predicted result and the actual value, the membership function of gaussian distribution is established. Suppose that the outputs after SVR and GBR are $\hat{y_a}$ and $\hat{y_v}$. The variable $\hat{y_a}$ is the predicted value of audio feature, and the variable $\hat{y_v}$ is the predicted value of visual. Then, we use $\hat{y_i}$ to stand for $\hat{y_a}$ and $\hat{y_v}$. The variable y_i is the actual label value, $\hat{y}$ is prediction value, and the variable e_i is error or residual, which can be described by

$$e_i = y_i - \hat{y_i} \tag{16}$$

We establish the relation function between residual error and fuzzy membership, i.e.,

$$\mu_{e_i} = \begin{cases} e^{-\frac{\theta e_i^2}{2\sigma^2}}, & -\sigma < e_i < \sigma \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where the parameter θ is a constant that can control the extent to which outliers affect the results. The parameter σ is the root

mean square error, which can be described as

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i)^2} \tag{18}$$

where the constant m is the number of samples. We use this fuzzy membership to adjust the predicted result of audio and visual.

Some reasons for establishing fuzzy membership degree are as follows:

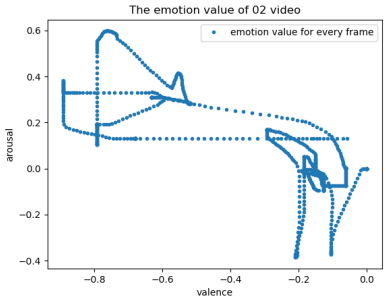
(a) The residual needs to meet three conditions: first, when the residuals are small, a large weight is selected. When the residuals are large, a small weight is selected. Second, the sum of the weights is one. Third, the residual error approximates normal distribution.

(b) Since the predicted value is distributed on both sides of the true value, it is approximately symmetric.

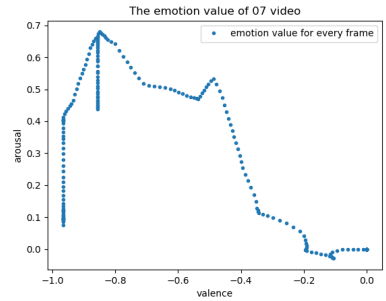
(c) The parameter σ represents the deviation between the predicted value and the true value, which can limit the range of outlier points well.

IV. EXPERIMENT AND ANALYSIS

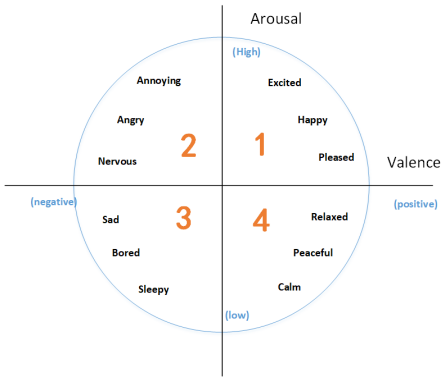
This paper uses the LIRIS-ACCEDE database [44], provided by the French Central Institute of Technology [42]. We select the official data set, which is provided in 2018. Thirteen videos are selected as the experimental subjects. Those videos vary in length and emotion tone as shown in Table 1. They contain audio messages that can be used to determine the overall emotional tone of a character's voice as shown in Figs. 2(a) and 2(b). Here several videos are listed for emotional tone analysis. The right of Fig. 2(c) is the emotional space description in [5], which is convenient for comparing the corresponding video emotional tone. For example, the most points of Fig. 2(a) lie in the negative axis, thus the emotion tone is unhappy. There are several movies selected based on the audio and visual differences. The unimodal of audio and visual signals are predicted by SVR and GBR regression network, respectively.



(a) The emotion values of video 02.



(b) The emotion values of video 07.



(c) The emotion space of arousal-valence.

FIGURE 2. The tone of 13 videos.

TABLE 1. The length of 13 videos.

video	01	02	03	04	05	06	
length (/second)	3516	844	4693	763	5593	5000	
video	07	08	09	10	11	12	13
length (/second)	209	210	210	210	210	210	210

Ren *et al.* use long short-term memory recurrent neural networks to extract audio features [49]. However, the audio features extracted by this method cannot be determined. The data dimension is large. Thus, audio features are extracted by the openSMILE toolbox in our paper. There are 1,582 dimensionalities. These segments are extracted every second. A total of 1,582 features are extracted from the audio signal, including a base of 34 low-level descriptors (LLD), 34 corresponding delta coefficients appended, as well as 21 functions applied to each of LLD contour. In addition, 19 functions are applied to the 4 pitch-based LLD and their four delta coefficient contours (152 features). Finally, the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features). We perform the feature-level fusion for features.

Various types of visual features are extracted. The video frame rate is 29.97 frame/s in those videos. Therefore, an emotion is easy to change in a few seconds. For each movie, 1 frame/s is extracted using the video feature extraction software – Ffmpeg.

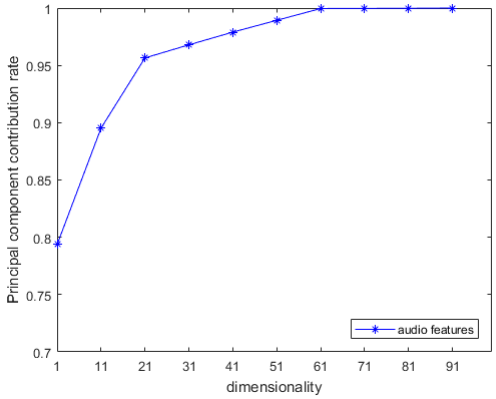
For each of these images, several general-purpose visual features are provided. Visual features include color features, texture features, color texture joint features, and other features. Color features include auto color correlogram and color layout and scalable color. Texture features include edge histogram and local binary patterns. Color texture joint features include color and edge directivity descriptor (CEDD), fuzzy

color and texture histogram (FCTH), fuzzy color and texture histogram and joint descriptor joining CEDD and FCTH in one histogram. Other features include gabor and tamura. In addition, we learn and extract visual features through convolutional neural networks (CNN) as high-level features [43]. They have been extracted using LIRE library [43], except CNN features that have been extracted using Matlab neural networks toolbox [45].

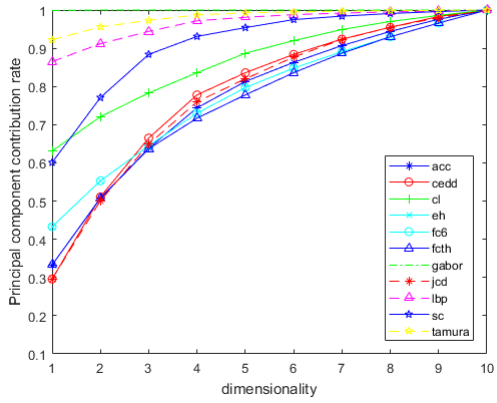
The principal component analysis (PCA) is a linear transformation-independent transformation of raw data into a set of dimensions that can be used to extract the main feature components of the data. We use PCA to reduce the dimensionality of audio and visual features. Since there are many visual features, we use the method of dimensionality reduction for each type of feature first before the feature-level fusion. Finally, the dimensionality of visual feature is 110 as depicted in Fig. 3(b). For audio features, all features are fused, and then PCA is used to reduce dimensionality. The audio features include prosody features, sound quality features, and spectral features. The 1582 dimensional features are fusion of all the features of the audio signal, and the fusion model is the splicing of multiple features.

It is proved by experiments that the contribution rate of principal component has reached 99% when features are reduced to 60 dimensions as portrayed in Fig. 3(a). However, due to the complexity of the video, features can be fully characterized when they reach 110 dimensions. Therefore, audio features are reduced to 110. This situation is also convenient for comparing model regression accuracy in audio and visual signals.

This paper selects the best segmentation method between the test set and the training set through experiments in Fig. 4. We compare the method of 10-fold cross-validation and the simple method of dividing the training set and the test set. Finally, we find that the 10-fold cross-validation result is not obvious in the case of large data sets, but the effect is significantly improved in the case of expanding the test set. Thus, on the regression network, 50% of the data are used for training and 50% for testing.



(a) The contribution rate of principal component in audio features.



(b) The contribution rate of principal component in visual features.

FIGURE 3. The principal component contribution rate of PCA.

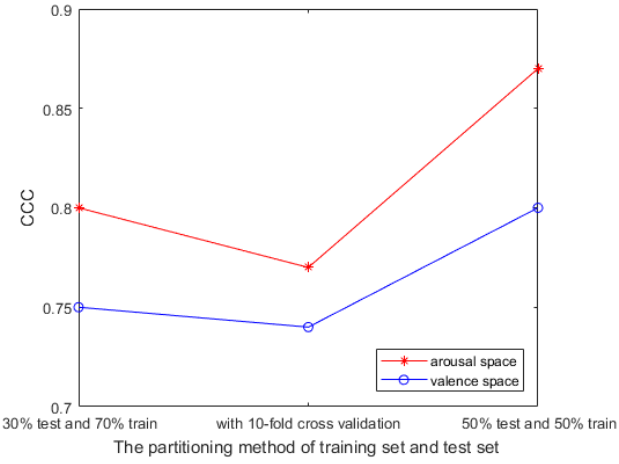


FIGURE 4. The partition method of training set and test set.

A. OUTLIERS CASES ANALYSIS

In this paper, visual and audio signals are used to predict emotions. Most general fusion methods are used in many works, such as linear fusion. However, we take into consideration the errors of different modalities in the recognition process and the possible outliers of a particular modality. Due to the emotional concealment in the audio (for example, people usually deliberately cover up their emotions or laughing wildly due to extreme sadness), or the fact that the emotional tone of the scene in video does not match the reality, outliers exist.

In Fig. 5, we find that single frame audio and visual can affect people's decision of overall video emotion. Therefore, we need to choose the appropriate emotional features, and have an appropriate treatment for abnormal situations in order to predict emotions more accurately. As seen from Fig. 5, the left and right figures come from same video, and the emotions are the same. However, a certain frame of the video will lead us to misunderstand.

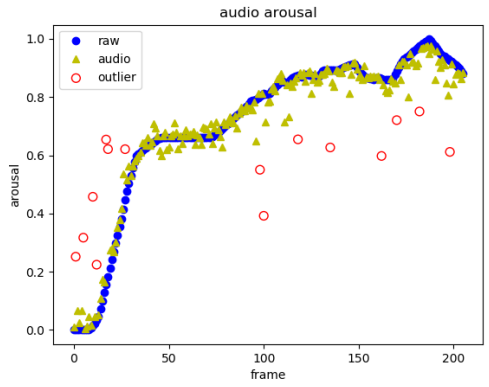
In Fig. 6, *raw* represents the original emotional data, the label of the experiment, *audio* represents the emotional result predicted using audio information, *video* represents the emotional result predicted using video information, and

(a) A mixture of love and hate.

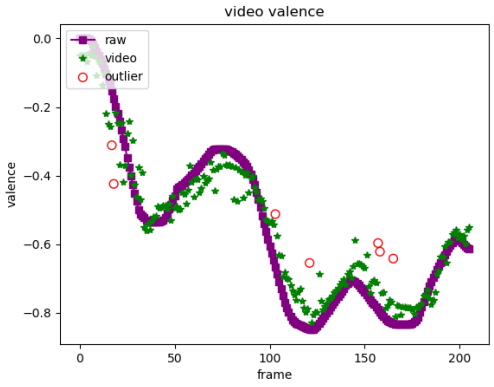
(b) The miscalculation of truth and falsehood.

(c) Grief beyond control.

FIGURE 5. **Different movies clips.**



(a) Outlier of audio prediction result in arousal space.



(b) Outlier of visual prediction result in valence space.

FIGURE 6. **Marking outliers.**

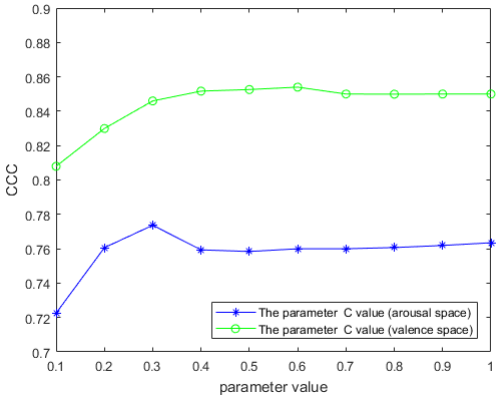


FIGURE 7. **The value of parameter *C*.**

outlier represents the prediction result that deviates from the original sentiment value. Red circles are marked as the position of the outliers. Here we use the threshold method to detect the outliers. When the residual of the predicted value and the actual sentiment value is greater than 0.2, it is thought of as an outlier [9]. When the distance between the predicted value and the actual value is greater than this threshold, the predicted value is marked as an outlier.

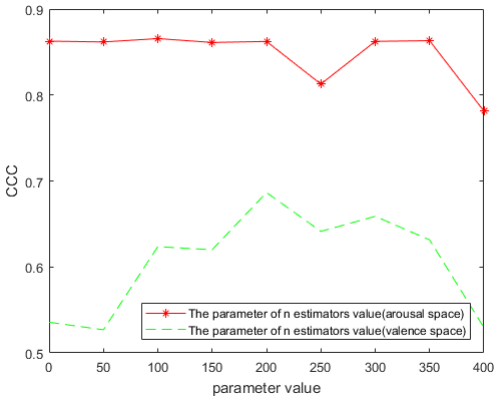
According to Fig. 6, it can be clearly seen that the number of outliers in the prediction result of visual is less than the prediction result of audio, showing that in the process of emotional change, the change of phonetic and intonation will have a great impact on the emotional prediction and the image is not misleading to the emotional recognition. Based on the above situation, we propose a fusion strategy that is an FWSVR model.

B. COMPARISON OF UNIMODAL AND MULTIMODAL EMOTION RECOGNITION

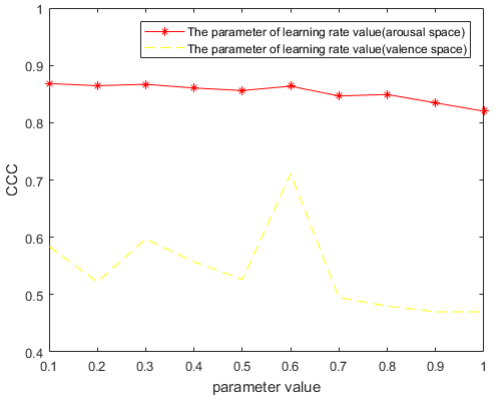
Although there is a mutual misleading possibility between the two modalities, the complementary relationship between them is more important. By regression prediction analysis on audio and visual separately and comparing the results of the two modal fusions, we find that the fusion results are better than the unimodal prediction results.

SVR is used to process audio features. In the SVR model, we use a kernel function to map data to a high dimensional space, to make better regression prediction for nonlinear problems. In this model, we choose the radial basis function (RBF) as the kernel function. Through a lot of experiments, we can obtain that the parameter C is 0.3 as shown in Fig. 7.

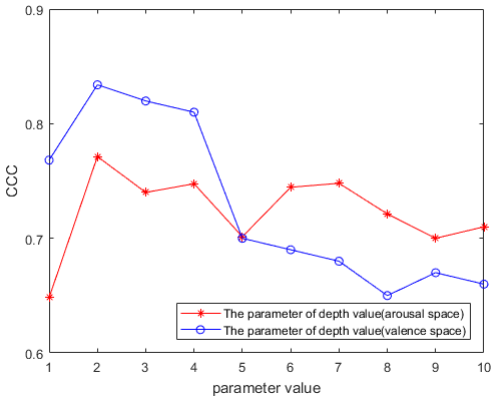
We use the GBR model to process visual features. In the parameter selection, we consider the prediction results of the arousal space and the valence space. The experimental results are shown in Fig. 8. In this model, the learning rate α is set to be 0.6 as in Fig. 8(b). The parameter $n-estimator$ is 200 as in Fig. 8(a), the *max depth* is 2 as in Fig. 8(c) and the loss function is the Huber one.



(a) The parameter of $n-estimator$ in emotion space.



(b) The parameter of $learning\ rate$ in emotion space.



(c) The parameter of $max\ depth$ in emotion space.

FIGURE 8. Selection of several parameters.

The result is shown in Table 2. The evaluation indexes include explained variance (EV), mean absolute error (MAE), mean squared error (MSE), R^2 and concordance correlation coefficient (CCC). In the arousal space, the result of fusion is 31.28% higher than the prediction of audio and 16.83% higher than the visual result in EV. The fusion results in other evaluation indicators are better than the unimodal prediction results. In the valence space, the result of fusion is 9.18% higher than the audio result and 13.15% higher than the visual result. Thus, the interrelationship between multimodalities is important and cannot be ignored. As long as the modal relationship is handled reasonably, emotion recognition will lead to the better results.

TABLE 2. Results for the different modalities.

ES	modality	EV	MAE	MSE	R2	CCC
AR	audio	0.581	0.087	0.013	0.580	0.764
	visual	0.729	0.050	0.009	0.705	0.858
	a+v	**0.894**	**0.03**	**0.004**	**0.888**	**0.946**
VA	audio	0.736	0.097	0.015	0.736	0.860
	visual	0.697	0.076	0.018	0.680	0.841
	a+v	**0.828**	**0.049**	**0.009**	**0.823**	**0.918**

[1] ES – emotion space.
[2] AR – arousal space, VA – valence space.

TABLE 3. Comparison of three fusion methods.

ES	Fusion	EV	MAE	MSE	R2	CCC
Arousal	LR	0.733	0.049	0.008	0.714	0.860
	LWLR	0.849	0.031	0.005	0.843	0.922
	FWSVR	**0.894**	**0.03**	**0.004**	**0.888**	**0.946**
Valence	LR	0.694	0.076	0.018	0.676	0.839
	LWLR	0.806	**0.047**	0.011	0.800	0.900
	FWSVR	**0.828**	0.049	**0.009**	**0.823**	**0.918**

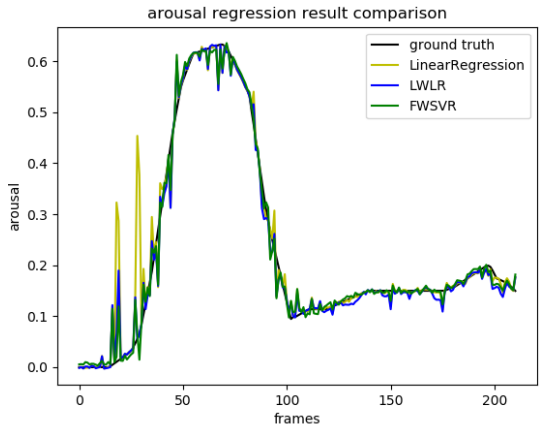
C. COMPARISON OF MULTIMODAL FUSION METHODS

In this paper, we try to use different fusion methods to predict the final result in emotion space, including the linear regression model (LR), local weight linear regression model (LWLR) and FWSVR model. Table 3 is the result of three fusion methods and we can find that the weighted fusion method is superior to linear regression. Comparing the two weighted fusion methods, LWLR considers the correlation between adjacent points of the test data and the FWSVR takes into account the differences between predicted values and actual value. The result indicates that the FWSVR fusion method is better than LWLR fusion one. Fig. 9 shows the fitting effect of the three fusion methods.

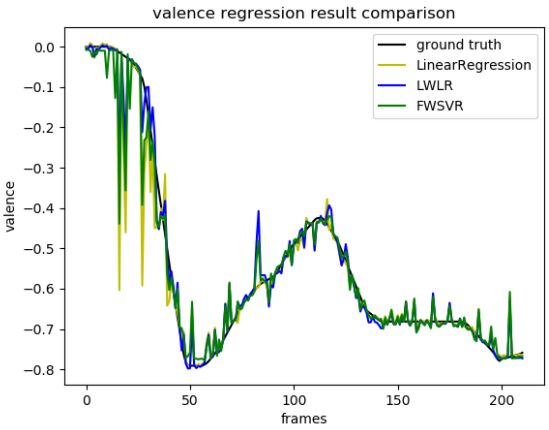
In addition, we select two films for experiment and compare them with LIRIS-ACCEDE dataset. The video 1 is *You again*, and the video 2 is *Dimensional meltdown*. FWSVR is performed on those videos and LIRIS-ACCEDE dataset, and the experimental results can be presented in Table 4, indicating that FWSVR has similar prediction results for different videos.

D. COMPARISON OF SVR, FUZZY WEIGHT LINEAR REGRESSION (FWLR) AND FWSVR

Considering the impact of outliers on prediction results, we propose a way to alleviate it. FWSVR is a fuzzy membership weighted regression process by regression of the audio and visual features obtained in the second layer of the regression network. The experimental results of SVR and FWSVR for the converged network are shown in Tables 5 and 6.



(a) Comparison of three fusion methods in arousal space.



(b) Comparison of three fusion methods in valence space.

FIGURE 9. Results obtained for fusion methods.

TABLE 4. Comparison of emotional prediction results in multiple videos.

Video	ES	MAE	CCC
Video 1	Arousal	0.047	0.923
	Valence	0.060	0.912
Video 2	Arousal	0.224	0.964
	Valence	0.082	0.880
Dataset	Arousal	0.030	0.946
	Valence	0.049	0.918

The prediction result of FWSVR is better than the results produced for SVR.

Many advantages exist in the FWSVR model. First, it can solve outliers and noise interference problems in the prediction results, and overcome the limitations of SVR. Second, considering the relationship between audio and visual contents, the fuzzy weighted processing of different modalities is carried out to predict the emotion trend more accurately. Third, the model effectively improves the evaluation indexes, EV, MSE, MAE, etc., which reflects the accuracy at global prediction and prevents the occurrence of overfitting. At the same time, the model also has some limitations. Due to the fuzzy processing of outliers and noise, the model is blurred at

TABLE 5. The comparison of SVR, FWLR and FWSVR (test set).

ES	model	EV	MAE	MSE	R2	CCC
Arousal	SVR	0.817	0.065	0.007	0.792	0.940
	FWLR	0.573	0.089	0.031	0.572	0.760
	FWSVR	**0.894**	**0.03**	**0.004**	**0.888**	**0.946**
Valence	SVR	0.807	0.078	0.011	0.799	**0.934**
	FWLR	0.724	0.099	0.016	0.723	0.854
	FWSVR	**0.828**	**0.049**	**0.009**	**0.823**	0.918

TABLE 6. The comparison of SVR, FWLR and FWSVR (training set).

ES	model	EV	MAE	MSE	R2	CCC
Arousal	SVR	0.917	0.055	0.004	0.906	0.997
	FWLR	0.810	0.074	0.007	0.810	0.918
	FWSVR	**0.999**	**0.005**	**0.00003**	**0.999**	**0.999**
Valence	SVR	0.954	0.055	0.004	0.941	0.995
	FWLR	0.871	0.077	0.008	0.866	0.946
	FWSVR	**0.993**	**0.012**	**0.0004**	**0.993**	**0.997**

the independent point of emotional mutation, which leads to the decrease of fitting accuracy of emotional mutation point. For the non-emotional saltus, FWSVR has better fitting effect than SVR.

Through the comparison of SVR and FWSVR, it is found that the fuzzy weighting has a certain improvement on the prediction effect. In addition, we also verify the effect of other fuzzy algorithms through experiments. This paper uses the same membership function, and the same data set for comparison experiments in FWLR and FWSVR. Tables 5 and 6 indicate that the FWSVR prediction results are better than those due to the FWLR, regardless of the test set or the training set.

E. OUTLIER REMOVAL ANALYSIS

In this experiment, we compare the recognition rates before and after the removal of the outliers. First, we calculate the regression recognition rate of audio and visual predicted results, which are the case where the outliers are not removed. Then we apply the proposed algorithm and the existing fuzzy theory algorithm to remove the outliers, and represent the outliers removal effect by the regression recognition rate. The final results indicate that the FWSVR regression prediction results are significantly better than the unimodal (audio or visual) regression prediction.

The regression recognition rate, denoted by *RRR*, is used to compare the removal of outliers.

$$RRR = \widetilde{n}/n \tag{19}$$

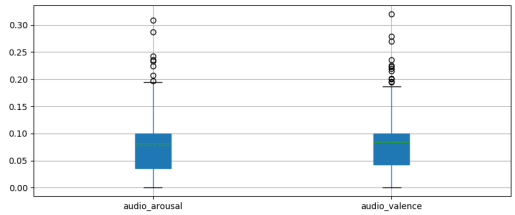
where $\widetilde{n}$ is the number of outliers, and n is total number of video frames.

In Table 7, the prediction result of FWSVR is better than that of audio modal by 25.118%, and the prediction result of FWSVR is better than that of visual modal by 9.04% in the arousal space. In the valence space, the prediction result

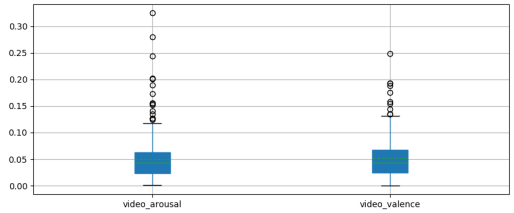
TABLE 7. Outlier removal rate.

ES	SVR	GBR	FWLR	FWSVR
Arousal	72.986%	89.10%	72.986%	**98.104%**
Valence	64.929%	93.365%	94.313%	**95.261%**

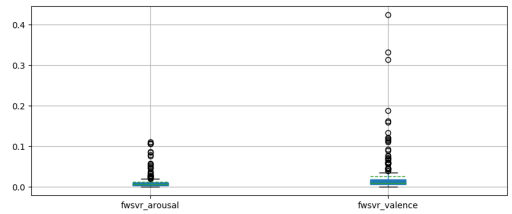
of FWSVR is better than that of audio modal by 30.332%, and the prediction result of FWSVR is better than that of visual modal by 1.896%. Although the FWLR model has a certain degree of improvement in the removal of outliers, the overall effect is unstable, especially in the arousal space. The prediction effect has a large gap between arousal and valence space.



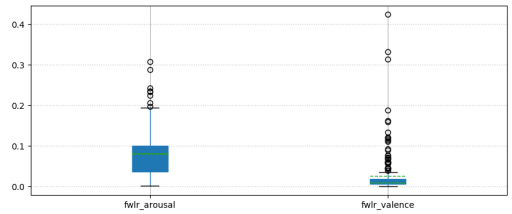
(a) The box plot of the audio prediction results.



(b) The box plot of the visual prediction results.



(c) The box plot of the fusion prediction results by FWSVR.



(d) The box plot of the fusion prediction results by FWLR.

FIGURE 10. Visualize predicted values and represent outliers using box plots.

As shown in Fig. 10, the box plot is used to explain the results of FWSVR, which is a statistics plot that shows the dispersion of a set of data. The box plot includes the lower edge (Q_1), the lower quartile (Q_2), the median (Q_3), the upper quartile (Q_4), and the upper edge (Q_5). The interquartile range (IQR) can be calculated by [48]

$$IQR = Q_3 - Q_1 \tag{20}$$

The box plot has three functions [47]: identifying data outliers, deciding data skewness and tail weight, and comparing data shapes.

The box plot provides a special standard for identifying outliers. The outlier is defined as a value less than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$. The box plot relies on real data without presupposing a specific distribution of data, just a real and intuitive display of the data itself. In addition, the box plot decides the outliers based on the quartile and the interquartile range. It has the advantage that the quartile has certain resistance. As up to 25% of the data can be arbitrarily far away without disturbing the quartile to a large extent, the outliers will not have a significant impact on the quartiles. There is a distinguisged advantage that the box plot is used for outlier recognition. In addition, since the evaluation index is derived from the residual, that is, the predicted value minus the actual sentiment value, the maximum and minimum values in the data and the statistics of the quartile are favorable for deciding the distribution of data. The outliers are concentrated on the side of the smaller value, and the distribution is left skewed. The outliers are concentrated on the side of the larger value, and the distribution exhibits a right skewed [46].

In this experiment, we use the error as the standard shown in (16). There are four box plots, including the unimodal regression result of audio and video by SVR and GBR models. The results of multimodal fusion by FWSVR and FWLR models are shown in Fig. 10. In Fig. 10(a), we can find that the error is in [0, 0.35], the outlier is above 0.18, and the standard value is under 0.18. In Fig. 10(b), we can find that the error is in [0, 0.35], and the outlier is above 0.12. Comparing Figs. 10(a) and (b), we can clearly find that the prediction result of the visual is better than that of the audio, since the error value of the outliers is reduced and the prediction accuracy is improved. In Fig. 10(c), we can see that the error is in [0, 0.12], which is lower than those shown in Figs. 10(a) and (b). In addition, most prediction results of FWSVR are concentrated within 0.04. When we set the tolerance to 0.2, there are only a few points in the FWSVR prediction that deviate from this threshold. From those figures, the proposed fusion algorithm minimizes errors and removes most of the outliers.

V. CONCLUSION

In this paper, the LIRIS-ACCEDE dataset is used for continuous dimensional emotion recognition. We pay attention to the problem that in the multimodality emotion recognition, the prediction result is inaccurate due to the existence of abnormal frames in the decision-level fusion. FWSVR algorithm is proposed to improve the fusion precision for multimodality emotion recognition.

The contributions of this research are summarized as follows.

(a) Mutual relationships between modalities are presented and verified. The relationships between modalities are complementary and restrictive at the same time. In experiment B,

modality complementarity is proved. The experiments verify that multimodal emotion recognition is superior to unimodal. In addition, it is found through experiments that outliers and noise after unimode prediction greatly restrict the decision-level fusion. Therefore, solving outliers and noise plays a key role in improving the recognition accuracy.

(b) We conduct multiple decision-level fusion experiments. The final results show that the method of adding weights to test data can effectively reduce the impact of errors and outliers on the model. LWLR is better than LR in predicting emotion. The prediction effect of FWSVR is better than LWLR, since the recognition errors of different modalities are considered. For the case where the recognition result deviates from the true value, by adding weights to the membership, the impact of outliers is reduced and the accuracy of prediction is improved.

(c) FWSVR and SVR models are compared, and we find that FWSVR has advantage and disadvantage. The advantage is that FWSVR has a good degree of blurring for outliers, thus the data well fit under normal conditions and there is no overfitting. The shortcoming stems mainly from the fact that the emotion of mutation is fuzzified, and then the recognition of saltus emotion needs to be improved.

There is a vast research space in emotion analysis which is worthy of our future study. In this work, we just investigate two modalities. In the future, one could add physiological signals in emotion recognition. In addition, the regression model will be updated, and the contextual semantics will be considered. We will deal with outliers and solve the impact of emotional saltus emotion on predictions. It is interesting to apply the outlier processing technique in this paper to automated manufacturing systems [51].

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group number RG-1440-048.

REFERENCES

[1] A. Chakraborty, A. Konar, U. K. Chakraborty, and A. Chatterjee, "Emotion recognition from facial expressions and its control using fuzzy logic," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 39, no. 4, pp. 726–743, Jul. 2009.

[2] H.-N. Tran and E. Cambria, "Ensemble application of ELM and GPU for real-time multimodal sentiment analysis," *Memetic Comput.*, vol. 10, no. 1, pp. 3–13, Mar. 2018.

[3] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Syst. Appl.*, vol. 135, pp. 60–70, Nov. 2019.

[4] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symp. Motivat.*, vol. 19, 1972, pp. 207–282.

[5] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[6] S. Anders, M. Lotze, M. Erb, W. Grodd, and N. Birbaumer, "Brain activity underlying emotional valence and arousal: A response-related fMRI study," *Hum. Brain Mapping*, vol. 23, no. 4, pp. 200–209, Dec. 2004.

[7] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, Jan. 2019.

[8] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Hum.-Like Intell. (CIHLI)*, Singapore, Apr. 2013, pp. 108–117.

[9] C. Wang, P. Lopes, T. Pun, and G. Chanel, "Towards a better gold standard: Denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation," in *Proc. Audio/Vis. Emotion Challenge Workshop*, New York, NY, USA, 2018, pp. 73–81.

[10] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human-robot interaction," *Expert Syst. Appl.*, vol. 66, pp. 42–61, Dec. 2016.

[11] H. M. Hersh and A. Caramazza, "A fuzzy set approach to modifiers and vagueness in natural language," *J. Exp. Psychol., Gen.*, vol. 105, no. 3, pp. 254–276, 1976.

[12] Y. Tang, Q. Mao, H. Jia, H. Song, and Y. Zhan, "An emotion-embedded visual attention model for dimensional emotion context learning," *IEEE Access*, vol. 7, pp. 72457–72468, 2019.

[13] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[14] R. Satapathy, Y. Li, S. Cavallari, and E. Cambria, "Seq2Seq deep learning models for microtext normalization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1–8.

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2019.

[16] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu, "EmoCo: Visual analysis of emotion coherence in presentation videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 927–937, Jan. 2020.

[17] X. Liu, A. Ouyang, and Z. Yun, "Fuzzy weighted least squares support vector regression with data reduction for nonlinear system modeling," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Dec. 2018.

[18] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019.

[19] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3695–3699.

[20] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal deep convolutional neural network for audio-visual emotion recognition," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2016, pp. 281–284.

[21] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image Vis. Comput.*, vol. 65, pp. 76–86, Sep. 2017.

[22] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, Nov. 2015.

[23] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–9.

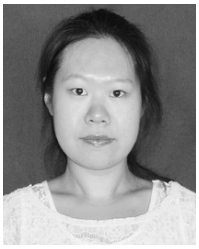
[24] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Brisbane, QLD, Australia, 2015, pp. 65–72.

[25] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, New York, NY, USA, 2015, pp. 41–48.

[26] A. Sayedelahl, R. Araujo, and M. S. Kamel, "Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, San Jose, CA, USA, Jul. 2013, pp. 1–6.

[27] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, Jul. 2019.

[28] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, New York, NY, USA, Jun. 2009, pp. 566–569.
[29] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Neural Netw.*, vol. 18, no. 4, pp. 423–435, May 2005.
[30] P. Rani, N. Sarkar, and J. Adams, "Anxiety-based affective communication for implicit human–machine interaction," *Adv. Eng. Informat.*, vol. 21, no. 3, pp. 323–334, Jul. 2007.
[31] R. E. Schapire, "The boosting approach to machine learning: An overview," *Nonlinear Estimation and Classification*. New York, NY, USA: Springer, 2003, pp. 149–171.
[32] X. Jiang, Z. Yi, and J. C. Lv, "Fuzzy SVM with a new fuzzy membership function," *Neural Comput. Appl.*, vol. 15, nos. 3–4, pp. 268–276, Jun. 2006.
[33] J. E. R. Dhas and S. Kumanan, "Evolutionary fuzzy SVR modeling of weld residual stress," *Appl. Soft Comput.*, vol. 42, pp. 423–430, May 2016.
[34] Z. Sun and Y. Sun, "Fuzzy support vector machine for regression estimation," in *Proc. IEEE Int. Conf. Theme-Syst. Secur. Assurance, Man Cybern.*, Washington, DC, USA, vol. 4, Oct. 2003, pp. 3336–3341.
[35] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy C-means with support vector regression and a genetic algorithm," *Inf. Sci.*, vol. 233, pp. 25–35, Jun. 2013.
[36] C.-F. Juang and C.-D. Hsieh, "TS-fuzzy system-based support vector regression," *Fuzzy Sets Syst.*, vol. 160, no. 17, pp. 2486–2504, Sep. 2009.
[37] C.-F. Juang, R.-B. Huang, and W.-Y. Cheng, "An interval type-2 fuzzy-neural network with support-vector regression for noisy regression problems," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 4, pp. 686–699, Aug. 2010.
[38] L. Chen, M. Zhou, M. Wu, J. She, Z. Liu, F. Dong, and K. Hirota, "Three-layer weighted fuzzy support vector regression for emotional intention understanding in human–robot interaction," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2524–2538, Oct. 2018.
[39] Y. Fan, H. Yang, Z. Li, and S. Liu, "Predicting image emotion distribution by emotional region," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Beijing, China, Oct. 2018, pp. 1–9.
[40] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[41] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[42] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Deep learning vs. Kernel methods: Performance for emotion prediction in videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Xi'an, China, Sep. 2015, pp. 77–83.
[43] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 43–55, Jan. 2015.
[44] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challengev (AVEC)*, New York, NY, USA, 2016, pp. 75–82.
[45] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Feature selection in multimodal continuous emotion prediction," in *Proc. IEEE 7th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, San Antonio, TX, USA, Jul. 2017, pp. 30–37.
[46] M. D. A. Pranatha, N. Pramaita, M. Sudarma, and I. M. O. Widyantara, "Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting," in *Proc. 4th Int. Conf. Wireless Telematics (ICWT)*, Jul. 2018, pp. 1–4.
[47] N. C. Schwertman, M. A. Owens, and R. Adnan, "A simple more general boxplot method for identifying outliers," *Comput. Statist. Data Anal.*, vol. 47, no. 1, pp. 165–174, Aug. 2004.
[48] W. Xie, O. Chkrebtii, and S. Kurtek, "Visualization and outlier detection for multivariate elastic curve data," *IEEE Trans. Vis. Comput. Graphics*, to be published.
[49] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal correlated network for emotion recognition in speech," *Vis. Informat.*, vol. 3, no. 3, pp. 150–155, Sep. 2019.
[50] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Inf. Sci.*, vol. 509, pp. 150–163, Jan. 2020.
[51] X. Zan, Z. P. Wu, C. Guo, and Z. H. Yu, "A Pareto-based genetic algorithm for multi-objective scheduling of automated manufacturing systems," *Adv. Mech. Eng.*, vol. 12, no. 1, pp. 1–15, 2020, doi: 10.1177/1687814019885294.



GE ZHANG was born in Daqing, Heilongjiang, China, in 1993. She received the bachelor's degree from the School of Electrical Information Engineering, Northeast Petroleum University, in 2016, and the master's degree from the School of Mechanical and Electrical Engineering, Xidian University, in 2019. She is currently pursuing the Ph.D. degree with the Institute of Systems Engineering, Macau University of Science and Technology. Her research interests include sentiment analysis, granular computing, and intelligent computing in human–computer interaction.

TIANXIANG LUO was born in Ankang, Shanxi, China, in 1993. He received the bachelor's degree from the School of Electrical and Information Engineering, Shaanxi University of Science and Technology, in 2015. He is currently pursuing the master's degree with the School of Electronic Engineering. His research interests include video emotion analysis and multimodal emotion recognition.

WITOLD PEDRYCZ (Fellow, IEEE) is currently a Professor and the Canada Research Chair (CRC) of computational intelligence with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. He is also with the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. In 2009, he was elected as a Foreign Member of the Polish Academy of Sciences. He is also the author of 15 research monographs covering various aspects of computational intelligence, data mining, and software engineering. His main research directions involve computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data mining, fuzzy control, pattern recognition, knowledge-based neural networks, relational computing, and software engineering. He has published numerous articles in this area.

Dr. Pedrycz was elected as a Fellow of the Royal Society of Canada, in 2012. He has been a member of numerous program committees of the IEEE conferences in the area of fuzzy sets and neurocomputing. In 2007, he received the prestigious Norbert Wiener Award from the IEEE Systems, Man, and Cybernetics Society. He was a recipient of the IEEE Canada Computer Engineering Medal, the Cajastur Prize for Soft Computing from the European Centre for Soft Computing, the Killam Prize, and the Fuzzy Pioneer Award from the IEEE Computational Intelligence Society. He is also intensively involved in editorial activities. He is also a member of a number of editorial boards of other international journals. He is also the Editor-in-Chief of *Information Sciences*, *WIREs Data Mining and Knowledge Discovery* (Wiley), and the *International Journal of Granular Computing* (Springer). He currently serves on the Advisory Board for the IEEE TRANSACTIONS ON FUZZY SYSTEMS.

MOHAMMED A. EL-MELIGY received the B.Sc. degree in information technology from the Menoufia University of Egypt in 2005. He has been a Software Engineer with King Saud University, Riyadh, Saudi Arabia, since 2009. His research interests include Petri nets, supervisory control of discrete event systems, database software, and network administration.

MOHAMED ABDEL FATTAH SHARAF received the degree in industrial engineering from Chiba University, Japan. He is the Head of the Development and Quality Unit, College of Engineering, King Saud University. He has Published more than 30 papers in the areas of spare parts control, quality management, maintenance, six sigma methodology, and academic accreditation.

ZHIWU LI (Fellow, IEEE) received the B.S. degree in mechanical engineering, the M.S. degree in automatic control, and the Ph.D. degree in manufacturing engineering from Xidian University, Xi'an, China, in 1989, 1992, and 1995, respectively.

He joined Xidian University in 1992. He is also currently with the Institute of System Engineering, Macau University of Science and Technology, Macau, China. His research interests include discrete event systems, data mining, and Petri nets. He was a recipient of an Alexander von Humboldt Research Grant, Alexander Von Humboldt Foundation, Germany. He is listed in Marquis Who's Who in the World, 27th Edition, 2010. He serves as a reviewer for 90+ international journals. He is the Founding Chair of Xi'an Chapter of IEEE Systems, Man and Cybernetics Society. He currently chairs Discrete-Event Systems Technical Committee of the IEEE Systems, Man and Cybernetics Society.