IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 24, 2019, accepted November 15, 2019, date of publication November 25, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955637

A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition

SHAHLA NEMATI[1], (Member, IEEE), REZA ROHANI[1], MOHAMMAD EHSAN BASIRI[1], (Member, IEEE), MOLOUD ABDAR[2], NEIL Y. YEN[3], (Member, IEEE), AND VLADIMIR MAKARENKOV[2]

[1]Department of Computer Engineering, Shahrekord University, Shahrekord 8818634141, Iran
[2]Department of Computer Science, Université du Québec à Montréal, Montréal, QC H2X 3Y7, Canada
[3]School of Computer Science and Engineering, University of Aizu, Aizu 965-8580, Japan

Corresponding author: Shahla Nemati (s.nemati@sku.ac.ir)

This work was supported by the research deputy of Shahrekord University under Grant 97GRN1M44264.

ABSTRACT Multimodal emotion recognition is an emerging interdisciplinary field of research in the area of affective computing and sentiment analysis. It aims at exploiting the information carried by signals of different nature to make emotion recognition systems more accurate. This is achieved by employing a powerful multimodal fusion method. In this study, a hybrid multimodal data fusion method is proposed in which the audio and visual modalities are fused using a latent space linear map and then, their projected features into the cross-modal space are fused with the textual modality using a Dempster-Shafer (DS) theory-based evidential fusion method. The evaluation of the proposed method on the videos of the DEAP dataset shows its superiority over both decision-level and non-latent space fusion methods. Furthermore, the results reveal that employing Marginal Fisher Analysis (MFA) for feature-level audio-visual fusion results in higher improvement in comparison to cross-modal factor analysis (CFA) and canonical correlation analysis (CCA). Also, the implementation results show that exploiting textual users' comments with the audiovisual content of movies improves the performance of the system.

INDEX TERMS Affective computing, emotion recognition, latent space model, multimodal fusion.

I. INTRODUCTION

Emotion recognition is the process of specifying the affective state of people. It plays an important role in affective computing and human-computer interaction (HCI) applications [1]. Due to exciting open challenges it poses and remarkable opportunities it provides, emotion recognition has raised increasing interest in both academia and business world [2]. Different applications benefit from emotion recognition, including video games [3], military healthcare [4], tutoring systems [5], predicting customer satisfaction [6], and Twitter analysis [7].

Recently, multimodal emotion recognition has attracted an increasing attention of researchers as it can overcome the limitations of monomodal systems [8]–[10]. Multimodal emotion recognition fuses complementary information of different modalities at different fusion levels. These levels can be classified into two categories: prior to matching and after matching fusion [11]. Sensor-level and feature-level fusion methods belong to the first category, while score-level, rank-level, and decision-level methods are known as the fusion after matching methods [12], [13]. The fusion level is determined with respect to the synchronization of modalities. When two modalities are synchronized, for each unit the classification should be performed (e.g., frame, video clip, or event), there are information from both modalities. For example, in the current study, audio and visual modalities are synchronized because for each frame, both audio and visual information are extracted. However, textual modality is not synchronized with audio and visual modalities, because there is no frame-level textual comment and each comment is usually written for the whole clip. If the modalities are synchronized, prior to matching fusion is possible, otherwise after mathcing fusion methodsl should be considered [14], [15].

Different combinations of modalities are proposed in the literature, including the fusion of facial expression with speech [1], [8], audiovisual and physiological signals [12], and the EEG response with textual modality [16]. In addition

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

to these combinations, recently, the fusion of textual modality with audiovisual modalities has appeared in two different lines of research, namely, emotion recognition [14], [15] and multimodal sentiment analysis [10], [17], [18]. The main difference between these research lines is that the goal of the former is to detect the emotion of the users in response to an affective video clip or music video, while the aim of the latter is to detect the emotion and sentiment expressed in video clips uploaded on social media platforms. Another difference is that the former improves the viewers' content selection and consumption [12], while the latter reveals the affective states of opinion holders who record their opinions on products, events, or services [2], [19].

Textual modality may be in the form of either the movie transcript [20], [21] or the user comments in social networks [14], [15], [22]. An obvious distinction between these two types of textual modalities is their nature. In fact, the movie transcripts are directly derived from the audiovisual contents and hence their sources are identical, while the users' comments in social networks are the crowd's sentiment toward an audiovisual content. Accordingly, the users' comments may be seen as the result of aggregating the conscious behavior of people's cognitive and affective system [23], while other modalities are resulted from unconscious responses. This makes the role of the users' comments more important in multimodal emotion recognition.

For emotion recognition of the users in response to multimedia content, both internal affective clues such as audiovisual content of video clips [1], [17] and external affective cues such as bodily and physiological changes [12], and the users comments are usable. Exploiting internal affective content and external users' comments is preferred over using physiological changes as it does not require any physiological sensors [24]. Therefore, in the current study, the fusion of audiovisual content with the users' comments for multimodal emotion recognition is considered.

Improving the multimodal fusion mechanism is a decisive factor for enhancing the performance of emotion recognition systems [18]. The majority of existing multimodal emotion recognition systems simply concatenate the extracted features from different modalities [9], [14], [15]. One of the main issues faced in this approach when used in traditional classification algorithms is the problem of redundant and conflicting information carried by different modalities [18]. Moreover, concatenating feature vectors pertaining to different modalities and forming a high dimensional feature vector lead to ignoring implicit correlation among modalities [25]. One exception to this drawback of feature concatenation is when deep neural models such as auto-encoders are used for feature representation. This neural models automatically extract relevant features from the data, prevent the problems mentioned for traditional classification models. The first motivation of the current is to consider user comments as an easy to obtain external affective cue for classifying viewers' emotional responses. The second motivation is to minimize the destructive effect of redundant and conflicting information in audio and visual modalities on the multimodal emotion recognition system.

To address the above problems, a new hybrid fusion method for the fusion of audiovisual content with textual users' comment is devised in this study. A latent space feature-level fusion method is employed to fuse the audio and visual modalities. By preserving the statistical correlation among the two modalities, this feature-level fusion estimates redundant features [25]. Having fused the audio and visual modalities, late fusion is employed to fuse the audiovisual and textual modalities. Here, decision-level fusion is inevitable because the textual and audiovisual contents are asynchronous.

The hybrid fusion method proposed here deals with the problem of redundant and conflicting information in audio and visual modalities. This is achieved by projecting audio and visual modalities into correlation space using three latent space fusion algorithms, namely, cross-modal factor analysis (CFA) [26], canonical correlation analysis (CCA) [27], and marginal Fisher analysis (MFA) [28] algorithms. Experimental evidence will be presented to show the better performance of the our method over simple concatenation of audio and visual feature vectors. Also, following the state-of-the-art in the fusion of audiovisual and textual modalities [9], [14], [15], we employ a Dempster-Shafer (DS)-based decision-level fusion method [29]–[31] to fuse the resulted audiovisual correlation space with the textual modality.

Although there are some similar studies that address the problem of redundant features in feature-level fusion by preserving the statistical correlation among the modalities, they are not used to complement the decision-level fusion. In other word, the existing methods either use a feature-level latent space fusion method [25] or use an evidential decision-level method to fuse audiovisual and textual modalities [9], [14], [15]. In the current study, we take the advantages of both methods by combining them in a hierarchical manner.

In summary, the current study provides the following contributions:

- We propose a hybrid fusion method for multimodal emotion recognition which benefits from both feature- and decision-level fusion.
- We utilize latent space fusion methods to find a common latent space by preserving the statistical correlation among the modalities.
- We exploit the social media comments besides the internal audiovisual content of movie clips to enhance the emotion recognition system.
- Using the proposed multimodal fusion method, better recommendations can be given to users who search for multimedia content on the Web.

The paper continues as follows. Section II reviews some important studies in multimodal emotion recognition. In Section III, our proposed system is described. Section IV, presents the results obtained during experiments and

discusses their implications. Finally, the main conclusions are given in Section V.

II. RELATED WORK

A brief review of some related work in the literature is presented in this section. To this aim, we discuss both multimodal emotion recognition and latent space fusion methods in the following.

A. MULTIMODAL EMOTION RECOGNITION

It has been shown that multimodal emotion recognition usually outperforms unimodal systems [17], [19], [32]. Although the use of multimodal affective systems has some advantages, it rises a number of challenges too [32]. For example, choosing modalities for the fusion, dealing with missing data, handling synchronization problem, and fusing different modalities are some of the most important challenges of multimodal systems [14], [17]. Among these challenges, selecting the best combination of modalities has been the focus of several studies in recent years [9], [12], [17], [33].

There is a long-standing research on the audio and visual content fusion [34]–[37]. The main drawback of earlier studies in audiovisual emotion recognition is the lack of evaluation on a standard data set [13]. For example, in [34], [35], and [38], the authors evaluated their system on their own recorded videos. To address this problem, later research studies such as [12] and [36] assessed their systems performance on standard multimodal databases such as eNTERFACE'05 [39] and RML [36]. Most of bimodal studies extract facial expression from visual channel and combine it with common audio features [13]. A comprehensive review of bimodal facial expression-based emotion recognition and audio modality may be found in [1], [40]–[42].

The fusion of EEG signals with other modalities is also a prevalent combination [43], [44]. Chanel *et al.* [45] fused peripheral physiological channel and EEG for adaptation of game difficulty according to the players emotions, and Wang *et al.* [46] combined EEG signal and video content for emotional video tagging. EEG signals have also been fused with facial expression [43], audio signals [47], and eye tracking data [48]. Alarcao and Fonseca [49] provided a good review of EEG-based emotion recognition systems.

Deep neural networks (DNN) models have been recently used in emotion recognition and show promising results. For example, Noroozi *et al.* [50], proposed a convolutional neural network (CNN) model for summarizing key-frame videos which are then combined with audio modality in a late fusion/stacking fashion for emotion prediction. They evaluated their method on the SAVEE, eNTERFACE'05, and RML databases. Avots *et al.* [1], proposed a CNN-based model for facial image classification. Specifically, they train an AlexNet CNN on the facial expression labeled according to their emotion. Then, they employed well-known audio and spectral features and combined their classification results obtained by SVM classifier using a decision-level fusion. Kulkarni *et al.* [51], proposed a CNN model trained for learning a static representation from images and enhancing it through Fisher vector encoding for visual feature extraction. Then, they used an SVM classifier for final emotion classification of congruent or incongruent facial expressions. They reported an improvement on CK+ and OULU-CASIA datasets for video emotion recognition.

The first reported study that addressed the fusion of external textual modality with audiovisual content was presented by Nemati and Naghsh-Nilchi [14]. The authors combined social media comments and low-level audiovisual features via a decision-level fusion method and applied their method on the affective video retrieval problem. They improved their DS-based fusion method using a weighting schema and showed that the fusion of the three modalities outperforms both unimodal and bimodal audiovisual affective video retrieval systems [15]. A feature-level fusion of audiovisual features and their decision-level combination with textual comments is also presented recently [9]. Movie transcripts has also been fused with other modalities [10], specifically with audio and visual modalities [2], [10], [18]. A comprehensive review of multimodal affect recognition and description of different combinations of modalities can be found in [32].

B. LATENT SPACE FUSION METHODS

Treating different modalities equally by concatenating their feature vector is shown to have low performance in data mining tasks [52]. Therefore, different strategies are proposed in the literature to consider the implicit correlation among modalities as well as to address the problem with redundant and conflicting information carried by different modalities [18], [25]. Latent space fusion methods are among the most powerful methods for addressing these issues. For example, CCA is used to elicit latent variables (i.e. common features) from two feature vectors by projecting them into correlation (latent) space [25].

An early study on audiovisual fusion using CCA was proposed for open-set speaker identification [53]. In this study, CCA was used to fuse lip and texture features. Nicolaou *et al.* [54], used dynamic probabilistic CCA (DPCCA) to fuse continuous annotations for analyzing affective behavior. Recently, the original CCA method is used to fuse different modalities for multimodal emotion recognition [55]. In this study, CCA is used for combining audiovisual features. More recently, labeled multiple canonical correlation analysis (LMCCA) is proposed for bimodal human emotion recognition [56]. They used class labels of training data to preserve the discriminative characteristics of different modalities.

CFA is a statistical linear method used for obtaining latent representation from two modalities [25]. CFA and CCA are different in that the former uses a criterion to minimize the Frobenius norm between two modalities while the latter maximizes the sets cross-correlation in the latent space [28]. CFA is used to fuse audiovisual content for multimodal emotion recognition in [36]. This study demonstrates the superiority of CFA to CCA for multimodal emotion recognition.

Latent Dirichlet Analysis (LDA) obtains the feature space by assuming the Gaussian distribution of the data in each class, which is not the case in many real-world problems, making the method ineffective [57]. Without taking into account this assumption, MFA creates a special space so that different classes' points are placed as far apart as possible and the data within a class are compressed as far as possible [25]. Xu *et al.* [58], applied MFA for content-based image retrieval (CBR). They showed that MFA and its extensions outperform similar methods on both HGR and CBR. Recently, Puthenputhussery *et al.* [57], applied MFA to several challenging visual recognition tasks and showed the feasibility of applying their marginal Fisher analysis (CMFA) method on different recognition applications.

C. THE IMPORTANCE OF THIS STUDY

As mentioned in previous section (see section I), this study attempts to consider all the available resources to predict users' affective responses to multimedia contents. For this reason and unlike most of previous studies (in which they considered only one or the combination of two such resources), we use a hybrid fusion method for having all of these important emotion recognition resources. Moreover, we would like to tackle the problem of redundant and conflicting information carried by existing multimodal emotion recognition systems. In this regard, we propose a latent space feature-level fusion method to fuse the audio-visual content, and then combining the classification results obtained by exploiting the audio-visual features with the textual modality using an evidential Dempster-Shafer theory-based fusion method. In the following, we give more details about the proposed method.

III. METHOD

The overall view of the proposed hybrid fusion method is shown in Fig. 1. The first part of the proposed framework consists of the feature extraction modules. These modules extract related features from the three modalities and pass them to the next stage modules. These features are extracted according to the previous research investigating audio, visual, and textual features and their effect on the emotional state of people who watch multimedia content [9], [12], [14], [43], [59]. The next subsections describe the process of extracting emotion-related features from the three modalities.

A. VISUAL FEATURE EXTRACTION

Shot length, lighting key, color, and motion are the four visual features extracted in the visual feature extraction module.

1) Shot length: The shot refers to sequential frames captured by a camera with no discernible color change in the content of the successive images. From the viewpoint of the viewer, the rapid changes in the scene indicate the dynamic and exciting nature of the scene [60]. For shot recognition, color histogram based techniques have been shown to be very effective and have been used in various studies [60]–[63]. To obtain the shot length, frames are shown in the HSV color space, and F^t is the frame at time t. Then, from each frame, the color histogram of the three channels of hue, h_H, saturation, h_s, and value, h_V, are calculated. For frames at time $t > N$, the feature matrix of X^t is constructed as:

$$X^t = \begin{bmatrix} x^t \\ x_{t-1} \\ \vdots \\ x^{t-N+1} \end{bmatrix} \tag{1}$$

where, $t = N, \ldots, T$, and N and T are window length and number of all frames, respectively. Then, the X^t matrix is decomposed using Singular Value Decomposition (SVD) and $s_1, s_2, \ldots, s_N$ denotes its eigenvalues with s_1 as its maximum value. The rank of X^t is determined by the number of eigenvalues greater than the threshold value, τ, multiplied by s_1. In other words, the r^t rank for X^t is the count of those s_i with $\frac{s_i}{s_1} > \tau$. This calculated rank has two important properties; First, if $r^t > r^{(t-1)}$, then the image contents of the current and previous frames are sufficiently different. Second, if $r^t < r^{(t-1)}$, the image content is stable enough to override the previous shot. Therefore, it can be concluded that the frame with the highest rank is the beginning of the frame of a shot. Also, a frame where $r^t > r^{(t-1)}$ and $r^{(t-1)} = 1$ is the last frame of the shot and hence, the difference between these two frames shows the shot length [62], [64].

2) lighting key: It has been shown that the amount and distribution of light in relation to the shadow and darkness of the scene are the primary visual aids for regulating emotional states [61]. There are two main lighting methods to relate the concept of scene and viewer's mode: high-key and low-key. In the high-key scene, there is usually less contrast and the difference between the brightest light and the darkest light is low. It has been shown that high key scenes with low contrast are commonly used to produce comedic and joyful scenes [65]. In low-key, the background is mostly dark and the contrast is high. Low-key with dark scenes is used to induce unpleasant emotions in violent and scary movies [65]. The brightness of a pixel in the image is proportional to the amount of light and the observed surface of the object. Therefore, high-light scenes that are brighter than low-light scenes have more pixels with high light. On the other hand, a low-light frame contains more pixels with low light. This simple property is used to distinguish between the two categories. For a frame, f, with $m \times n$ pixels:

$$LK_f = \delta_f \times \sigma_f \tag{2}$$

where, LK, δ_f, and σ_f are the average and variance of the value of the frame.

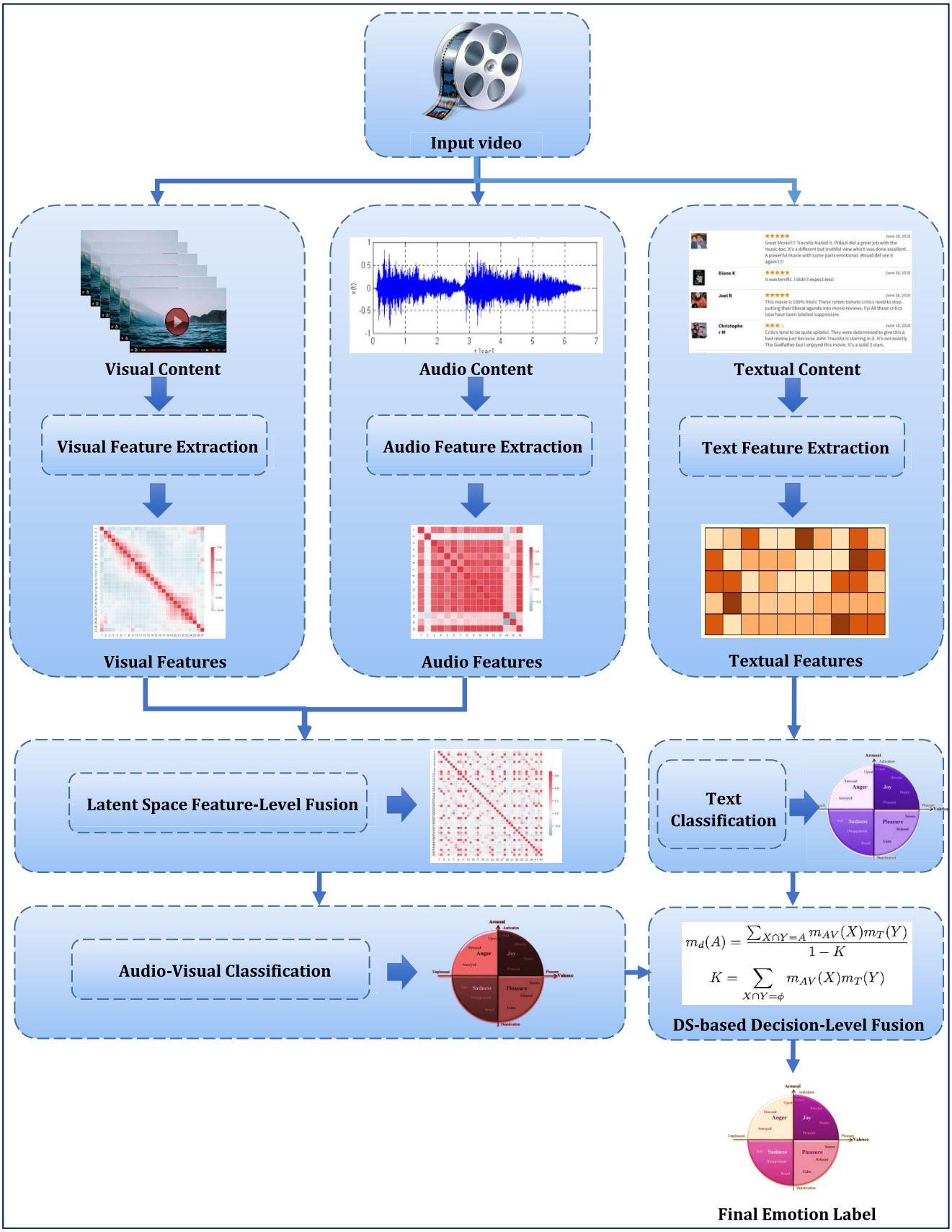


FIGURE 1. The proposed multimodal emotion recognition framework.

3) Color: Color and its related features play an important role in stimulating emotions [66]. For example, it has been shown that yellow, orange, and red are associated with feelings of fear and discomfort, while blue, purple, and green can evoke feelings of high valence and low arousal in the viewer [63], [65], [67]. Physiological studies on color have also shown that valence is associated with bright colors and arousal with color saturation [60]. It has been shown in [63] that the amount of color in the video is related to the color energy. The color energy is calculated by:

$$ColorEnergy^k = \sum_{j=1}^{PixelNum} \frac{s_j \times v_j}{std_{Hist_H} \times PixelNum} \quad (3)$$

where, $PixelNum$ is the total number of pixels in the frame k, s_j and v_j are the values of saturation and value of pixel j in the HSV space, respectively. std_{Hist_H} is the value of standard deviation of the hue histogram. According to the previous studies [59], [62], [66], [68], [69], In addition to the color energy, for each $n \times m$ frame, the 16 histogram of hue in the HSV space is considered as color features.

4) Motion: It has been shown that there is a relationship between camera movement and emotions such as happiness, sadness, and fear [70]. Psychological-physiological studies have also shown that there is a relationship between the intensity of the aroused emotion in the viewer and the observation of movement in the video [60]. In this study, according to [62], [68], the motion vector is used by the Block Matching Algorithm (BMA), which is designed for finding sequential blocks in a video sequence. The basic idea in estimating motion in this algorithm is that the object and background patterns in a video frame move in sequential frames to form relevant objects. To do this, the current frame is subdivided into a matrix of macroblock. Then, these blocks and their adjacent neighbors of the previous frame are compared. This, forms a vector for representing the movement of a macroblock. For all frame blocks, this motion vector is calculated. The search area for a macroblock is limited to search parameter, that is p pixels in four directions of the corresponding macroblock in the previous frame. More moves require larger p and this results in greater computational complexity [71]. In the current study, 16 pixels is considered as macroblock size and $p = 7$ pixels.

The output of the cost function is used for matching one macroblock against another. The least expensive macroblock is one that is the most compatible block to the current one. Mean Squared Error (MSE) and Mean Absolute Difference (MAD) are the most common cost functions used in the algorithm:

$$MSE = \frac{1}{S^2} \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} (C_{ij} - R_{ij})^2 \quad (4)$$

$$MAD = \frac{1}{S^2} \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} |C_{ij} - R_{ij}| \quad (5)$$

where S is the macroblock size, C_{ij}, and R_{ij} are pixels compared in the current macroblock and the reference macroblock [68]. The Peak-Signal-to-Noise-Ratio (PSNR) indicates the properties of the moving frame generated by the motion vector and the reference frame macroblocks:

$$PSNR = 10 \log \left[\frac{MAX_I^2}{MSE} \right] \quad (6)$$

where, MAX_I is the maximum of values for pixels in the image. In the current study, the Exhaustive Search (EX) algorithm is used to obtain the motion vector. This algorithm locates the best match with the highest PSNR among the pattern matching algorithms.

B. AUDIO FEATURE EXTRACTION

Previous studies have shown that features such as Mel Frequency Cepstrum Coefficient (MFCC), Zero Crossing Rate (ZCR), energy, and pitch can be used to extract emotions from the audio channel [72]–[74]. Therefore, these features are used in the current study.

1) Zero crossing rate: The number of times the audio signal crosses the x-axis, the zero line, or signal cahnge per unit time is known as zero crossing rate [62]:

$$Z_t = \frac{1}{2N} \sum_{n=1}^{N} |sign(x_t(n)) - sign(x_t(n-1))| \quad (7)$$

$$sign(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

where, $t \in [t_1, t_2]$, $x(t)$ is the audio signal divided to segments using a sliding window of length T, $n \in [0, N]$, and $x_t(n)$ is the time sequence of the t^{th} segment.

2) MFCC coefficient: These coefficients represent the spectral shape of the audio signal via the nonlinear Mel scale. To analyze the Cepstral coefficients with the Mel frequency we followed the following steps:
   - Split the audio signal into frames with fixed window size and fixed shift;
   - for each frame:
     - Calcuate the Fast Fourier Transform (FFT),
     - Pack frequencies based on the Mel scale,
     - Calculate the logarithm,
     - Calculate the Discrete Cosine Transform (DCT).

   MFCC are widely used to detect emotion because they exhibit acoustic tube characteristics, and therefore

contain a great deal of emotional information [75]. Many studies have also identified these coefficients as one of the most distinctive audio features. Low coefficients of this feature have been commonly used in past studies [62], [66], [72], [76]–[78]. The first 13 MFCC coefficients are used in the current study.

3) Pitch: It shows the basic frequency of the signal [63]. From the emotional viewpoint, the rhythm and the average pitch of the audio signal is related to the valence. For example, "sadness" is in accordance with a low standard deviation, while "discomfort" is associated with a higher amount of pitch [40]. It has also been shown that happiness and discomfort usually have a higher average pitch and higher talk rate. However, sadness is represented by lower average pitch and lower talk rate [79]. To calculate the pitch, autocorrelation is used. let $x[n]$ be a sinusoidal stochastic process function:

$$x[n] = cos(\omega_0 n + \phi) \tag{9}$$

then, the autocorrelation of $x[n]$ is:

$$R[t] = E\left\{x^*[n]x[n+t]\right\} = \frac{1}{2}cos(\omega_0 t) \tag{10}$$

The pitch can be calculated by finding the maximum of autocorrelation. In practice, only S samples are used to find an estimate of $R[t]$ as:

$$\hat{R}[t] = \frac{1}{S}\sum_{s=0}^{S-|t|}(w[s]x[s]w[s+|t|]x[s+|t|]) \tag{11}$$

where, $w[s]$ is a window of length s and the expected value of $\hat{R}[t]$ is:

$$E\left\{\hat{R}[t]\right\} = \left(1 - \frac{|t|}{S}\right)\frac{cos(w_0 s)}{2}, \quad |t| < S \tag{12}$$

where its maximum is on the pitch [15], [75].

4) Energy: It shows the signal strength or total energy of signal. From the emotional point of view, the energy associated with an audio signal of exciting emotions (e.g., discomfort or happiness) is higher than that of an audio signal containing sadness or fatigue emotions [12]. The total signal energy of $x_t(n)$ is usually obtained as [20]:

$$Energy = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(x_t(n)^2\right)} \tag{13}$$

C. FEATURE-LEVEL FUSION

There is different information in each modality and only a portion of them are related to emotion. For example, there are different information in audio and video signals for representing gender, age, personality, etc. Such features may reduce the quality of emotion recognition systems if they are used in training the model [25]. Feature-level latent space fusion methods may be used to find the common emotion-related features by mapping the features extracted from separate modalities into latent space. There are two main approaches to generate transform features into the latent space; 1) Maximizing cross-correlation of features, 2) Minimizing the distance between features or their norm. In this study, we applied both methods in the feature-level fusion stage (Fig. 1).

For the feature-level fusion module, we applied both supervised and unsupervised latent space fusion methods. Specifically, we applied CCA and CFA as unsupervised methods and MFA as supervised method on features of audio and video modalities. To apply the CCA, suppose $(\mathbf{a}, \mathbf{v})$ is zero mean feature vectors of audio and video signals as follows:

$$(\mathbf{a}, \mathbf{v}) = \{(a_1, v_1), (a_2, v_2), \ldots, (a_n, v_n)\}. \tag{14}$$

with n being the number of samples (i.e. frames), a_i and v_i are the original features extracted from modalities of s and t dimensions, respectively. The objective of CCA is to form two transformation matrices $\boldsymbol{W}_a$ and $\boldsymbol{W}_v$ with dimensions of $s \times r$ and $t \times r$ where $r \leq min(s, t)$ [25]. The original features of audio and video modalities are mapped into the latent subspace by $\boldsymbol{W}_a$ and $\boldsymbol{W}_y$ so that the correlation between $\hat{a} = \mathbf{a}\boldsymbol{W}_a$ and $\hat{v} = \mathbf{v}\boldsymbol{W}_v$ is maximized. This corresponds to maximizing *correlation coefficient* ρ between the projected feature vectors $\hat{a}$ and $\hat{v}$ as follows:

$$\begin{aligned}\rho &= \max_{\boldsymbol{W}_a, \boldsymbol{W}_v}\frac{E[\hat{a}^{\mathbf{T}}\hat{v}]}{\sqrt{E[\hat{a}^2]E[\hat{v}^2]}} \\ &= \max_{\boldsymbol{W}_a, \boldsymbol{W}_v}\frac{E[\boldsymbol{W_a}^{\mathbf{T}}\boldsymbol{a}^{\mathbf{T}}\boldsymbol{v}\boldsymbol{W_v}]}{\sqrt{E[\boldsymbol{W_a}^{\mathbf{T}}\boldsymbol{a}^{\mathbf{T}}\boldsymbol{a}\boldsymbol{W_a}]E[\boldsymbol{W_v}^{\mathbf{T}}\boldsymbol{v}^{\mathbf{T}}\boldsymbol{v}\boldsymbol{W_v}]}} \\ &= \max_{\boldsymbol{W}_a, \boldsymbol{W}_v}\frac{\boldsymbol{W}_a^{\mathbf{T}}C_{av}\boldsymbol{W}_v}{\sqrt{\boldsymbol{W}_a^{\mathbf{T}}C_{aa}\boldsymbol{W}_a\boldsymbol{W}_v^{\mathbf{T}}C_{vv}\boldsymbol{W}_v}}\end{aligned} \tag{15}$$

where C_{av}, C_{aa}, and C_{vv} are the cross-covariance matrix of $(\mathbf{a}, \mathbf{v})$, the covariance matrix of $\mathbf{a}$, and the covariance matrix of $\mathbf{v}$, respectively. Equation (15) is simiar to an Eigen-value problem and can be solved as [25]:

$$C_{aa}^{-1}C_{av}C_{vv}^{-1}C_{va}\boldsymbol{W}_a = \rho^2\boldsymbol{W}_a \tag{16}$$

$$C_{vv}^{-1}C_{va}C_{aa}^{-1}C_{av}\boldsymbol{W}_v = \rho^2\boldsymbol{W}_v. \tag{17}$$

For the feature-level fusion module of the proposed framework, CFA may also be used. When the CFA method is applied to $(\mathbf{a}, \mathbf{v})$ using two transformation matrices $\boldsymbol{W}_a$ and $\boldsymbol{W}_v$, the following criterion should be minimized:

$$\min_{\boldsymbol{W}_a, \boldsymbol{W}_v} = ||\mathbf{a}\boldsymbol{W}_a - \mathbf{v}\boldsymbol{W}_v||_F^2 \tag{18}$$

where, $\boldsymbol{W}_a^T\boldsymbol{W}_a$ and $\boldsymbol{W}_v^T\boldsymbol{W}_v$ are unit matrices. Frobenius norm F is the calculated as:

$$||\boldsymbol{W}||_F = \sqrt{\sum_{ij} w_{ij}^2}. \tag{19}$$

To obtain the optimal transformation matrices $\boldsymbol{W}_a$ and $\boldsymbol{W}_v$, (19) must be solved. Then, Singular Value Decomposition (SVD) is used to decompose the cross-covariance matrix

C_{av} as [25]:

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

Therefore, we have:

$$\boldsymbol{W}_a = S_{av}, \quad \boldsymbol{W}_v = D_{av}. \tag{21}$$

Class labels can be used to generate a shared latent space more effectively [57]. The use of class labels enables the feature-level fusion method to create a shared space in such a way that the resulted feature vectors have higher intra-class compactness and inter-class separability. To improve the performance of the feature-level fusion method, we employed the MFA algorithm. MFA uses class label information in the process of generating latent space. In MFA, the intra-class compactness is specified by:

$$\begin{aligned} S_C &= \sum_i \sum_{i \in N_{k_1}^+(j) \text{ or } j \in N_{k_1}^+(i)} ||W^T x_i - W^T x_j||^2 \\ &= 2w^T X(D-S)X^T w \end{aligned} \tag{22}$$

where, $X = [x_1, x_2, .., x_N]$ is the set of samples (i.e. frames), N is the numberf of samples, $N_{k_1}^+$ is k_1 neighbors of x_i in the same class, and S and D are calculated as:

$$S_{ij} = \begin{cases} 1, & \text{if } i \in N_{k_1}^+(j) \text{ or } j \in N_{k_1}^+(i) \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

$$D_{ij} = \sum_j S_{ij} \tag{24}$$

Moreover, the inter-class separability is specified by:

$$\begin{aligned} S_P &= \sum_i \sum_{(i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j)} ||W^T x_i - W^T x_j||^2 \\ &= 2w^T X(D^P - S^P)X^T w \end{aligned} \tag{25}$$

where, c_i is the i^{th} emotion class, $P_{k_2}(c)$ is a set of k_2 nearest pairs and S is calculated as:

$$S_{ij}^P = \begin{cases} 1, & \text{if } (i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j) \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Now the objective function can be formulated as:

$$\hat{w} = \arg\min_w \frac{W^T X(D-S)X^T w}{W^T X(D^P - S^P)X^T w} \tag{27}$$

By solving the generalized eigenvalue problem, the optimal solution, $y = X^T w$, can be calculated:

$$Ly = \lambda L^p y \tag{28}$$

where $L = D - S$ and $L^P = D^P - S^P$ are the Laplacian matrices of W and W^p, respectively.

D. TEXTUAL FEATURE EXTRACTION

To classify the textual viewers' comments using a supervised method, each comment should be first converted into a feature vector. It has been shown that n-gram features, TF-IDF, part of speech (POS), and lexical features are effective features for emotion recognition and sentiment analysis [22], [80]. The input textual comments are first preprocessed, then features are extracted from the preprocessed texts. In the preprocessing step, stop-words, non-English characters, and Web addresses are removed. Then, stemming is applied to the text to convert all forms of words to their stem. In the current study, unigram, bigram, and TF-IDF features are employed. Low-order N-grams including unigram and bigram are used broadly in NLP tasks to capture textual context [80]–[82]. It has been shown that the combination of unigram and bigram features are effective for sentiment analysis [22]. TF-IDF is also a popular feature extraction scheme in NLP and text mining tasks. It is used to calculate the importance of a word in a document. TF-IDF has two main parts; (1) term frequency (TF), showing the number of times that a term occurs in a document; (2) inversed document frequency (IDF), representing the amount of information a word provides in a document collection. TF-IDF is calculated as:

$$TF - IDF(w,d,D) = tf(w,d).idf(w,D), \tag{29}$$

$$tf(w,d) = f_{w,d}, \tag{30}$$

$$idf(w,D) = log\frac{|D|}{1 + |d \in D : w \in d|} \tag{31}$$

where, $f_{w,d}$ is the raw count of the word w in the document d, $|D|$ is the size of corpus D, and $|d \in D : w \in d|$ is the count of documents with $tf(w,d) > 0$. To prevent a division-by-zero caused by those terms with zero frequencies in the corpus, a 1 was added in the denominator.

E. DECISION-LEVEL FUSION

The features mapped into the latent space by CCA, CFA, or MFA methods are fed into a classifier to find the best emotion category of the corresponding frame based on audio and visual modalities. In the current study, the Naive Bayes and Support Vector Machines (SVM) are employed for classification. The reasons for selecting these methods are as follows. First, these methods are effective for both textual and audio-visual content classification [14], [83]. Second, most previous studies on multimodal emotion recognition on the DEAP dataset [14], [14], [15] have employed these methods and selecting the same classifier makes the comparison between our fusion method and theirs fair. A decision-level fusion method is needed for combining the result of these modalities with that of the textual modality [83], [84]. In most emotion-related researches, majority voting, maximum of scores, and average are used for desicion-level fusion [9], [14], [85]. In the current study, following the promissing results reported in [14], [14], [15], we used the Dempster-Shafer (DS) [29] fusion method for

decision-level fusion of textual and audio-visual results. The DS-based fusion method has several advantages over simple averaging-based fusion methods including taking into account all pieces of available evidence and preserving evidence's maximal agreements [30], [31].

To appliy the DS aggregation method for decision-level fusion, the evidence should be first defined. In this study, the output of feature-level fusion and textual feature extraction modules are considered as evidence supporting the final emotion category of each video. After defining the evidence, we should define the mass function, a basic probability assignment (BPA) that must satisfy the following properties [30], [84]:

$$m_d(\phi) = 0 \quad \text{and} \sum_{A \in 2^\theta} m_d(A) = 1 \tag{32}$$

where, θ, frame of discernment, is a finite set of mutually exclusive hypotheses and $d \in \{AV, T\}$. AV and T are the audio-visual and textual modalities, respectively [29]. When classes are disjoint, A has only one element and is called singleton. This is the case in the current study, because we assumed that each movie only belongs to one emotion category.

The mass function should be defined in a way that it shows the amount of support of evidence for each subset $A \subseteq \theta$ [9], [14]. Therefore, we define the probability of a feature vector, x_d, belonging to each emotion category computed by the classification modules as the mass function:

$$m_d(c_i) = \frac{P(c_i|x_d)}{\sum_{j=1}^{C} P(c_i|x_d)} \tag{33}$$

where $P(c_i|x_d)$ shows the probability of a clip belonging to class c_i based on the feature vector x_d and C is the total number of emotion categories.

Using the Dempster's rule of combination for decision-level fusion is the next step for using the DS theory in the proposed framework:

$$m_d(A) = \frac{\sum_{X \cap Y = A} m_{AV}(X) m_T(Y)}{1 - K} \tag{34}$$

$$K = \sum_{X \cap Y = \phi} m_{AV}(X) m_T(Y) \tag{35}$$

where m_{AV} and m_T correspond to the audio-visual and textual evidence, and K is a normalization factor ensuring the BPA properties of $M(A)$.

Finally, the final emotion category is calculated as:

$$c = \arg\max_i m_d(c_i) \tag{36}$$

IV. RESULTS

In this section, we discuss the the performance of the proposed method and the obtained results.

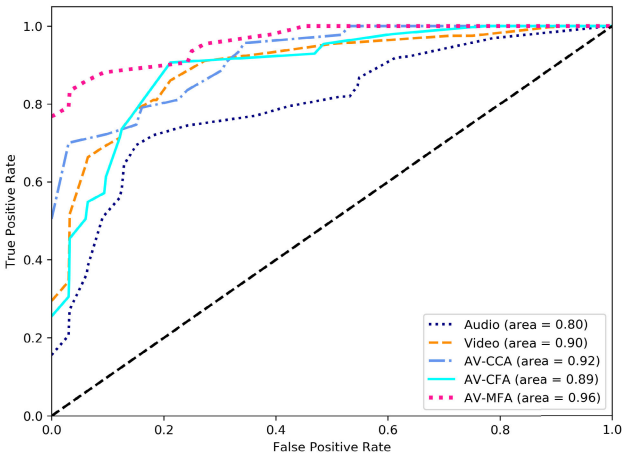


FIGURE 3. Comparison of the ROC curves using audio, video, and their fusion using CCA, CFA, and MFA with SVM classifier.

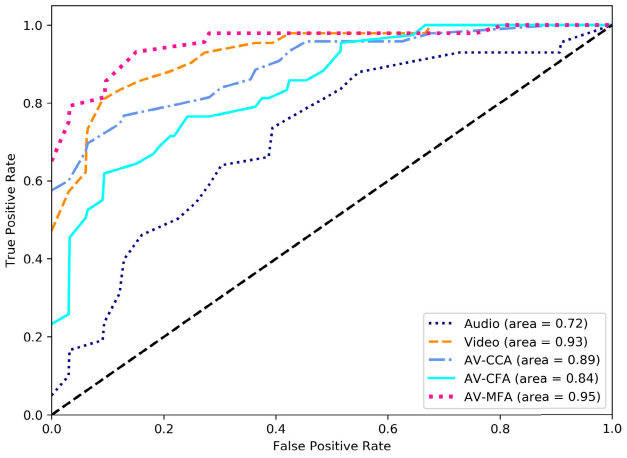


FIGURE 4. Comparison of the ROC curves using audio, video, and their fusion using CCA, CFA, and MFA with Naive Bayes classifier.

A. DATASET AND EVALUATION METRICS

To evaluate the proposed hybrid latent space fusion method, the experiments were conducted on the extended version of the DEAP dataset, a multimodal dataset for analyzing affective states [4]. This version of the dataset contains viewers' comments as textual modality and was previously used in [9], [14], [14], [15], [55]. We used the four emotion labels "PH", "NH", "PL", and "NL" corresponding to "Positive High", "Negative High", "Positive Low", and "Negative Low" quarters of the dimensional affect model.

In the experiments, the following five evaluation criteria are used; Precision (π), recall (ρ), F-Measure, accuracy, and specifity [31], [55], [75], [83]:

$$\pi = \frac{TP}{TP+FP}, \tag{37}$$

$$\rho = \frac{TP}{TP+FN}, \tag{38}$$

$$F - Measure = \frac{2 \times \pi \times \rho}{\pi + \rho}, \tag{39}$$

$$specifity = \frac{TN}{TN+FP}, \tag{40}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}. \tag{41}$$

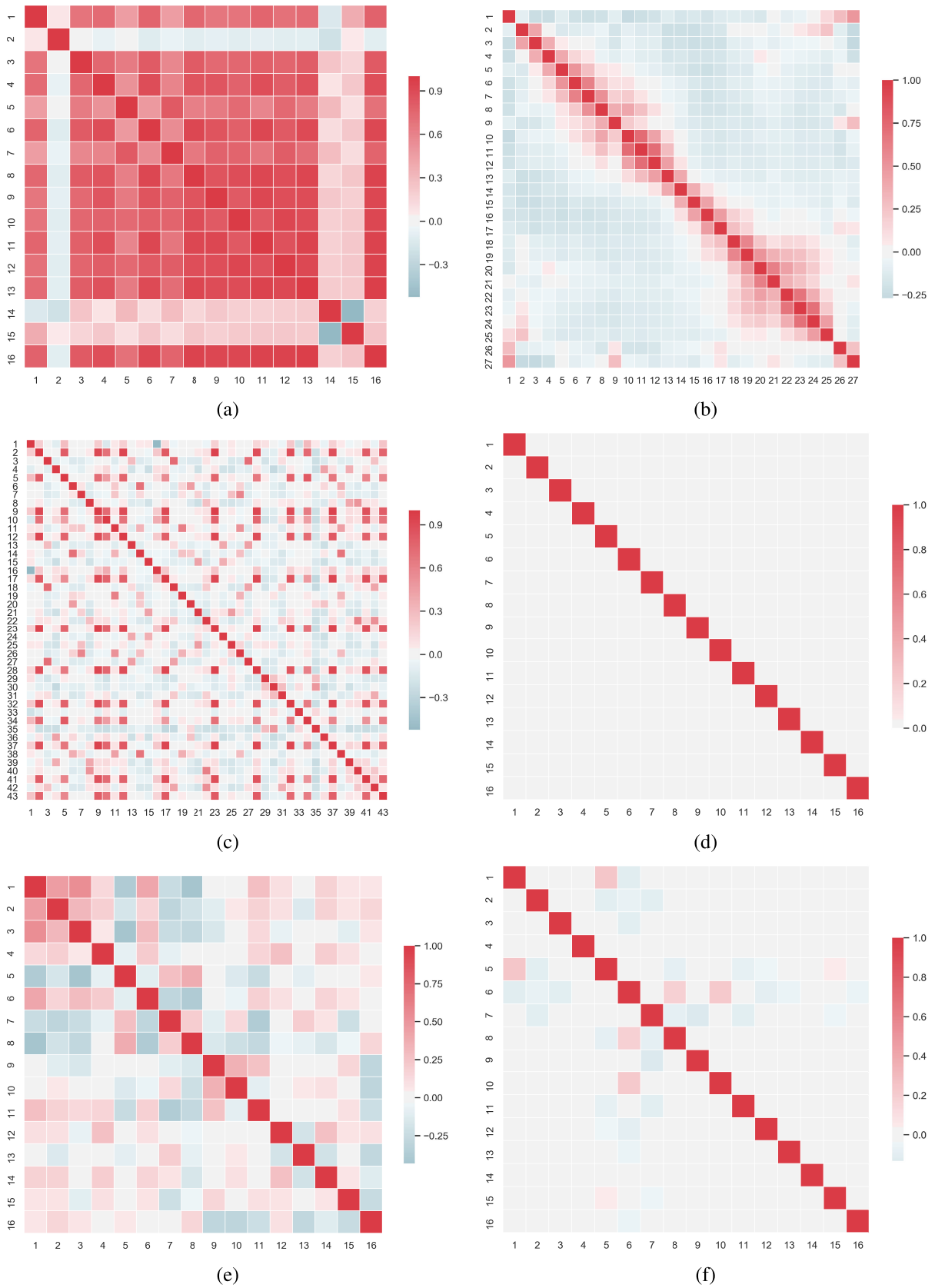


FIGURE 2. Comparison of the correlation between (a) audio features (b) video features, (c) concatenation of audio and video features (d) fusion of audio and video using CCA, (e) fusion of audio and video using CFA, and (f) fusion of audio and video using MFA.

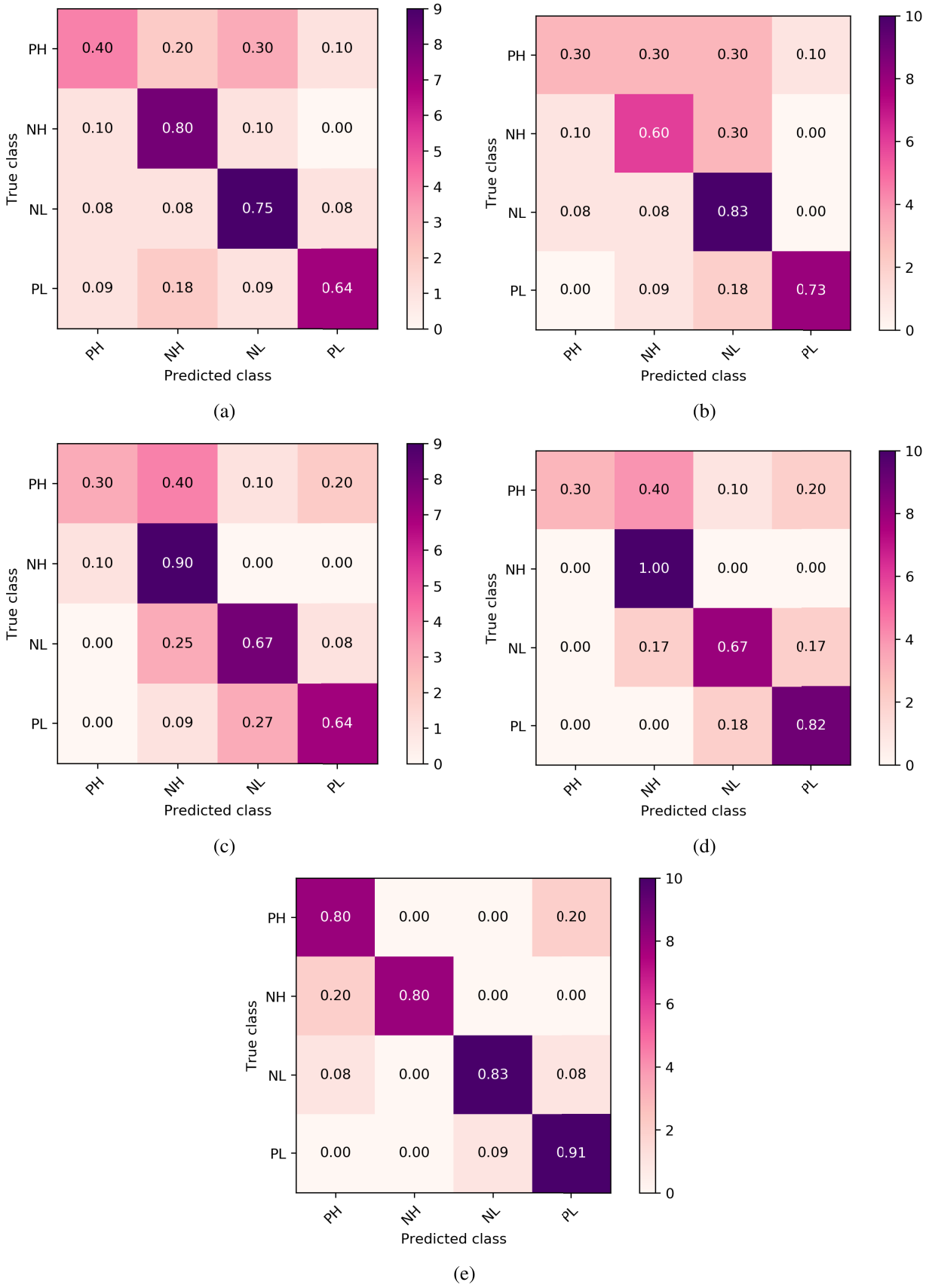


FIGURE 5. Comparison of the confusion matrices using: (a) audio features (b) video features, (c) CCA, (d) CFA, and (e) MFA using SVM classifiers.

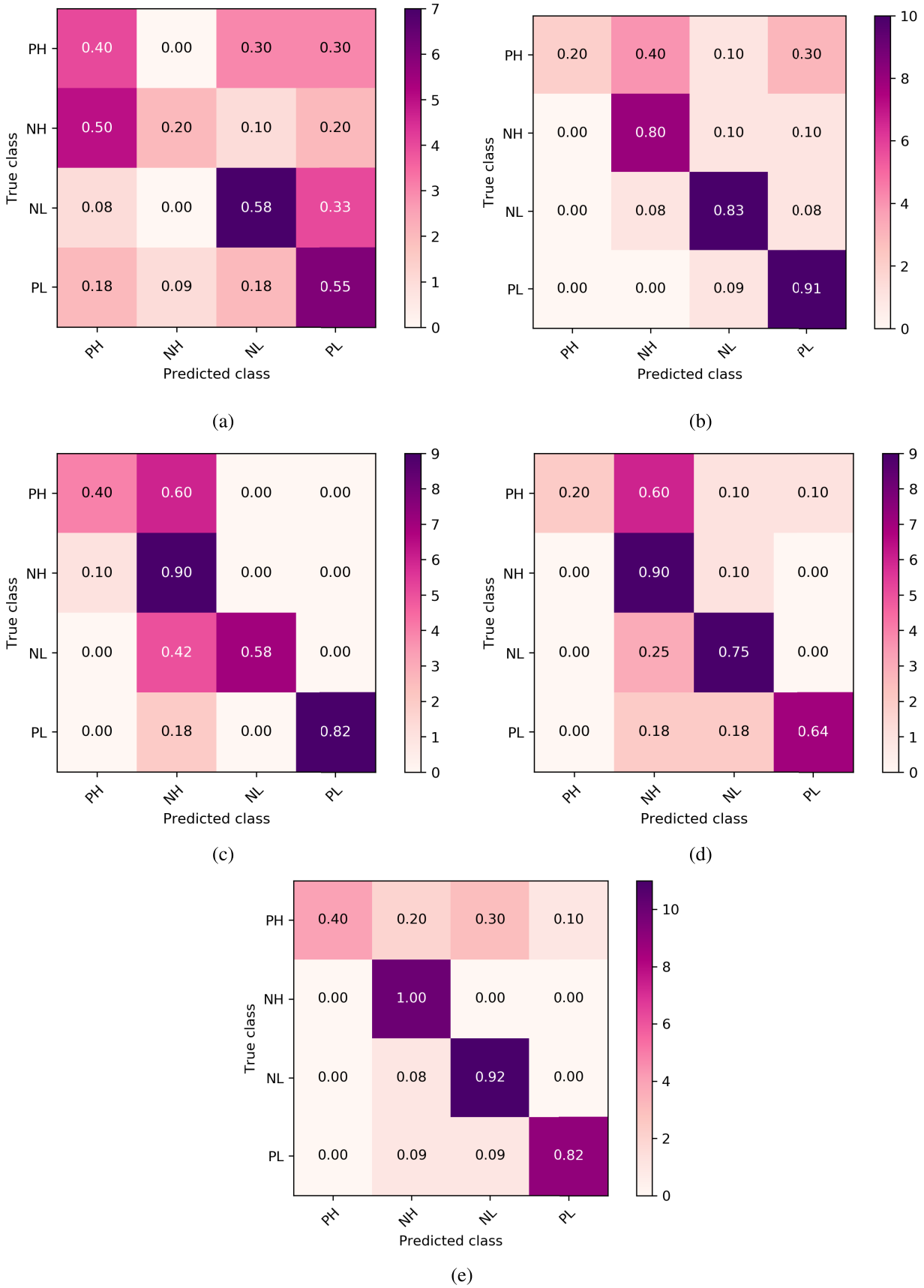


FIGURE 6. Comparison of the confusion matrices using: (a) audio features, (b) video features, (c) CCA, (d) CFA, and (e) MFA using Naive Bayes classifiers.

TABLE 1. Comparison of the performance of using audio and visual modalities with their feature-level fusion using CCA, CFA, and MFA fusion methods.

Modality	Classifier	Precision	Recall	F1-measure	Specifity	Accuracy
Audio	Naive Bayes	0.48	0.44	0.46	0.81	0.72
	SVM	0.65	0.65	0.65	0.88	0.82
Video	Naive Bayes	0.76	0.69	0.72	0.90	0.85
	SVM	0.62	0.65	0.63	0.87	0.81
AV-CCA	Naive Bayes	0.82	0.67	0.74	0.90	0.84
	SVM	0.66	0.63	0.64	0.88	0.81
AV-CFA	Naive Bayes	0.75	0.63	0.68	0.87	0.82
	SVM	0.76	0.70	0.73	0.90	0.85
AV-MFA	Naive Bayes	0.83	0.79	0.81	0.93	0.89
	SVM	0.85	0.84	0.84	0.94	0.92

TABLE 2. Comparison of the performance of using textual modality (T), feature-level fusion of audio and visual modalities (AV) with their decision-level fusion (AVT) using CCA, CFA, and MFA fusion methods.

Modality	Classifier	Precision	Recall	F1-measure	Specifity	Accuracy
T	Naive Bayes	0.79	0.53	0.64	0.84	0.70
	SVM	0.82	0.61	0.70	0.87	0.76
AV-CCA	Naive Bayes	0.82	0.67	0.74	0.90	0.84
	SVM	0.66	0.63	0.64	0.88	0.81
AVT-CCA	Naive Bayes	0.84	0.77	0.80	0.92	0.88
	SVM	0.84	0.81	0.82	0.94	0.91
AV-CFA	Naive Bayes	0.75	0.63	0.68	0.87	0.82
	SVM	0.76	0.70	0.73	0.90	0.85
AVT-CFA	Naive Bayes	0.74	0.66	0.70	0.89	0.84
	SVM	0.85	0.80	0.83	0.94	0.91
AV-MFA	Naive Bayes	0.83	0.79	0.81	0.93	0.89
	SVM	0.85	0.84	0.84	0.94	0.92
AVT-MFA	Naive Bayes	0.78	0.77	0.78	0.92	0.88
	SVM	0.88	0.86	0.87	0.95	0.93

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively [31].

B. THE EFFECT OF FEATURE-LEVEL FUSION

To compare the three feature-level latent space fusion methods described in the previous section, Fig. 2 compares the correlation heat maps of audio, visual, concatenation of audio and visual, and the fusion of audio and video using CCA, CFA, and MFA methods.

As shown in the figure, most of audio features have high correlation with each other, while video features and their concatenation with audio features have relatively lower correlation. Among the latent space fusion methods, the correlation between the features mapped into the shared latent space using CCA and MFA is quite lower than that of CFA method. This shows the higher ability of CCA and MFA in finding the shared latent space.

To compare the effect of applying the feature-level fusion on the classification of emotions, Figs. 3 and 4 show the macro-averaged ROC curves for classification using audio, visual, and fusion of audio and visual features with Naive Bayes and SVM classifiers, respectively.

As shown in the Figs. 3 and 4, the overall performance of the SVM classifier is higher than that of the Naive Bayes classifier. Also, the lowest performance macro-averaged area under curve (AUC) is obtained using audio features in isolation, while the highest AUC is achieved when the MFA is used to map audio-visual features into the shared latent space. This may be the result of considering class labels in characterizing the intra-class compactness and the inter-class seperability by the MFA algorithm.

For some applications like stress detection and therapy, depression detection and therapy, and emotion-aware recommender systems, the detection rate of different emotions is of different importance. To show which emotion classes are categorized more accurately, the confusion matrices of classification using audio, visual, and the fusion of audio and visual features are shown in Fig. 5 and Fig. 6.

As shown in Figs. 5 and 6, the lowest detection rate in terms of true positives is that of the ''PH'' emotion class, while, in average, the ''NL'' and ''NH'' emotion classes are detected more accurately. This shows that, negative emotions are detected more accurately using the feature-level fusion.The '' PH'' emotion class contains emotions such as joy, pleasant, and excited, while the ''PL'' class, for example, consists of calm, pleasure, and relaxed emotions [9], [14]. On the contrast, the ''NH'' emotion class contains emotions such as anger, annoyed, and upset, while the ''NL'' class includes sadness, depressed, and bored emotions [14], [15].

To compare the overall performance of the feature-level fusion methods, the results are shown in Table 1 according to the five performance measures described earlier.

As shown in Table 1, the use of only the audio modality, results in the lowest performance among the combinations

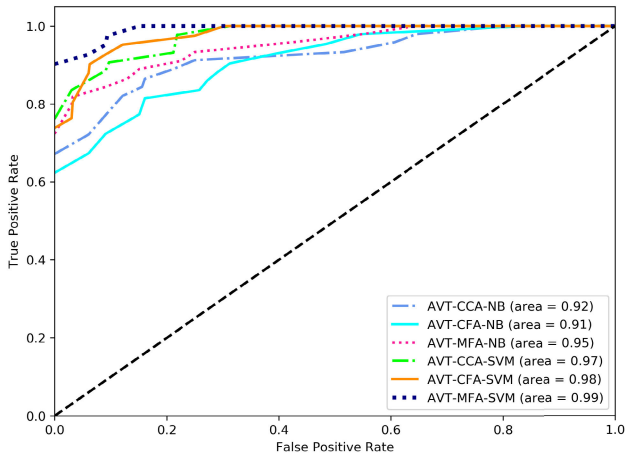


FIGURE 7. Comparison of the ROC curves of the fusion of audio, video, and textual modalities using CCA, CFA, and MFA with SVM and Naive Bayes (NB) classifiers.

shown in the Table. This may be due to the fact that there is not enough information in the audio content of video clips of the dataset to correctly classify their emotional class. Moreover, as shown in Fig. 2, the correlation between the audio features is very high, making some of the features low informative. The visual features ranked before the audio features and have relatively higher performance in classifying emotional content of video clips. This shows that there is more emotional information in the visual content of video clips in comparison to their audio content. Among the three combinations of audio and visual modalities, the MFA method outperforms other two feature-level methods in terms of all performance measures. This may be the results of exploiting class labels in the process of mapping audio and visual features into the shared latent space.

C. THE EFFECT DECISION-LEVEL FUSION

To show the effect of adding the textual modality to the result of audio and visual modalities, we fused the results obtained using the textual classifier with the results obtained using the audio-visual classifier. In other word, the decision-level fusion is used to combine the audio-visual and textual modalities. This is inevitable, because the audio and visual modalities are synchronized, while the textual modality is not synchronized with them. In fact, the audio and visual modalities belong to the frame-level (i.e., each record for these modalities corresponds to a frame in the dataset) while the textual modality belongs to the movie-level (i.e., each record for textual modality corresponds to a movie in the dataset). This is due to the fact that users usually write comments to describe a video clip or a part of it and do not provide frame-level comments.

As shown in Table 2, exploiting the results of classifying the textual modality using the SVM classifier in addition to audio and visual modalities, improves the performance of the emotion recognition system. However, as shown in Tables 1 and 2, this is not the case with the

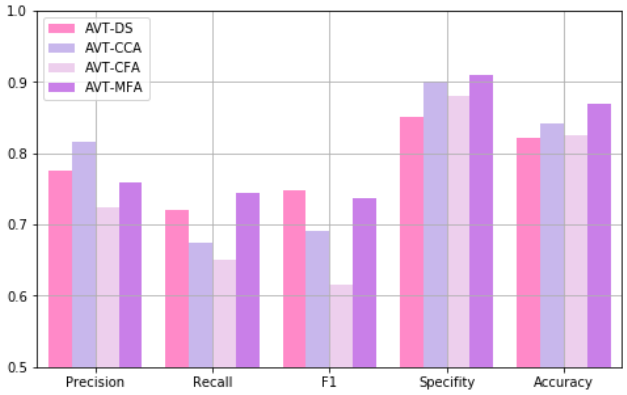


FIGURE 8. Comparison of the decision-level fusion of audio, visual, and textual modalities using the DS method (AVT-DS) with our proposed hybrid fusion of the three modalities using CCA (AVT-CCA), CFA (AVT-CFA), and MFA (AVT-MFA) methods by exploiting the Naive Bayes classifier.

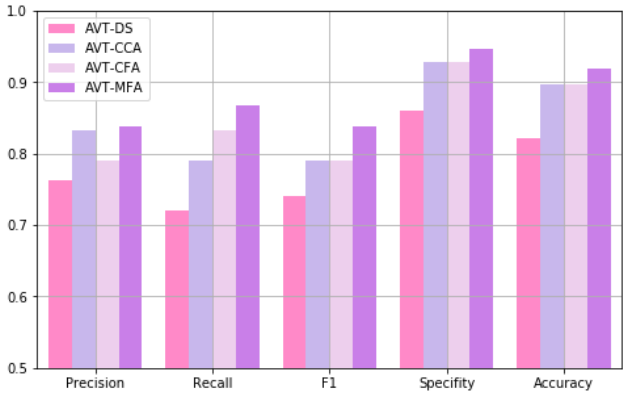


FIGURE 9. Comparison of the decision-level fusion of audio, visual, and textual modalities using the DS method (AVT-DS) with our proposed hybrid fusion of the three modalities using CCA (AVT-CCA), CFA (AVT-CFA), and MFA (AVT-MFA) methods by exploiting the SVM classifier.

Naive Bayes classifier. This may be the result of the lower performance of the textual classification module when the Naive Bayes classifier is used. In other words, the performance of the decision-level fusion of the results depends on the performance of the individual classifiers. This has been shown in Fig.7.

To view the effect of using the proposed hybrid fusion method, Fig. 8 and Fig. 9 compare the proposed system using the three latent space feature-level fusion methods with the decision-level fusion of audio, visual, and textual modalities proposed by Nemati and Naghsh-Nilchi [14].

As shown in Fig. 8 and Fig. 9, the proposed hybrid fusion method using the MFA for the feature-level fusion component outperforms other hybrid methods and the decision-level fusion method. This may be the result of considering the common latent space features for audio and visual modalities and then combining them with the textual modality.

V. CONCLUSION

In this study, a hybrid fusion method for multimodal emotion recognition is proposed. In the proposed method, the latent

space feature level fusion is used to map the audio and visual modalities into a shared latent space. These mapped latent features are employed to classify the video clips of the DEAP dataset into emotion categories. In addition to these audio-visual features, unigram, bigram, TF-IDF, and lexicon-based features are extracted from the viewers' textual comments on video clips. Then, textual features are used to train supervised classification methods. Finally, a decision-level Dempster-Shafer-based fusion method is used to fuse the textual and audio-visual classification results. The implementation results show that the feature-level fusion of the audio and visual modalities improve the performance of the system in comparison to simple concatenation of the audio and visual features. Moreover, among the three latent-space fusion methods, CCA, CFA, and MFA methods, the MFA achieves a higher accuracy. This result is probably due to employing class labels when generating the shared latent space. Also, the Dempster-Shafer-based decision-level combination of audio-visual and textual modalities outperform both the bi-modal fusion of audio and visual content and decision-level combination of audio, video, and textual modalities.

There remain some open questions, the most important ones being the investigation of the effect of more complex sentiment analysis methods on the overall performance of the emotion recognition system. In fact, more powerful methods for sentiment analysis that recently have been proposed, such as deep neural networks may be used to enhance the results obtained from the textual modality. This may be considered a promising direction for the future work. Addressing the intrinsic problems of the evidential DS fusion method employed for fusing the textual modality with the audio-visual modalities. One of the most important such limitations is the conflicting evidence problem of the DS fusion method. Addressing this problem may not only improve the performance of the decision level fusion, but also may enhance the overall quality of the multimodal system. In the future, we also plan to investigate other latent-space fusion methods for detecting emotion in multimodal contents. Also, exploiting other types of modalities may be another direction for the future research. Finally, exploiting score-level fusion methods in the proposed hybrid fusion model will be investigated as a future work.

DATA AVAILABILITY

The experimental data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 975–985, 2019.

[2] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.

[3] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Proc. Image Process. Commun. Challenges*. Cham, Switzerland: Springer, 2015, pp. 227–236.

[4] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Proc. Defense Sci. Res. Conf. Expo (DSR)*, Aug. 2011, pp. 1–5.

[5] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Comput. Sci.*, vol. 104, pp. 437–444, Jan. 2017.

[6] K. P. Seng and L.-M. Ang, "Video analytics for customer emotion and satisfaction at contact centers," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 3, pp. 266–278, May 2017.

[7] N. Colneriĉ and J. Demsar, "Emotion recognition on Twitter: Comparative study and training a unison model," *IEEE Trans. Affective Comput.*, to be published.

[8] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1319–1329, Jul. 2016.

[9] S. Nemati and A. R. Naghsh-Nilchi, "Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video retrieval," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 222–228.

[10] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[11] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. New York, NY, USA: Springer-Verlag, 2007.

[12] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[13] K. P. Seng, L.-M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.

[14] S. Nemati and A. R. Naghsh-Nilchi, "Incorporating social media comments in affective video retrieval," *J. Inf. Sci.*, vol. 42, no. 4, pp. 524–538, 2016.

[15] S. Nemati and A. R. Naghsh-Nilchi, "An evidential data fusion method for affective music video retrieval," *Intell. Data Anal.*, vol. 21, no. 2, pp. 427–441, 2017.

[16] S. Kumar, M. Yadava, and P. P. Roy, "Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction," *Inf. Fusion*, vol. 52, pp. 41–52, Dec. 2019.

[17] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[18] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[19] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[20] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[21] D.-A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Trans. Affective Comput.*, to be published.

[22] M. E. Basiri and A. Kabiri, "HOMPer: A new hybrid system for opinion mining in the persian language," *J. Inf. Sci.*, Feb. 2019.

[23] W. F. Morris, *Emotion and Anxiety: A Philosophic Inquiry*. Bloomington, IN, USA: Xlibris, 2006.

[24] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affective Comput.*, vol. 9, no. 2, pp. 147–160, Apr./Jun. 2018.

[25] R. R. Sarvestani and R. Boostani, "FF-SKPCCA: Kernel probabilistic canonical correlation analysis," *Appl. Intell.*, vol. 46, no. 2, pp. 438–454, 2017.

[26] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.

[27] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 39–50, Jun. 2010.

[28] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.

[29] A. P. Dempster, "A generalization of Bayesian inference," *J. Roy. Stat. Soc. B, Methodol.*, vol. 30, no. 2, pp. 205–232, 1968.

[30] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghasem-Aghaee, "Sentiment prediction based on Dempster-Shafer theory of evidence," *Math. Problems Eng.*, vol. 2014, Apr. 2014, Art. no. 361201.

[31] M. E. Basiri, N. Ghasem-Aghaee, and A. R. Naghsh-Nilchi, "Exploiting reviewers' comment histories for sentiment analysis," *J. Inf. Sci.*, vol. 40, no. 3, pp. 313–328, 2014.

[32] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, 2015, Art. no. 43.

[33] F. A. Salim, F. Haider, O. Conlan, and S. Luz, "An approach for exploring a video via multimodal feature extraction and user interactions," *J. Multimodal User Interfaces*, vol. 12, no. 4, pp. 285–296, 2018.

[34] C.-Y. Chen, Y.-K. Huang, and P. Cook, "Visual/acoustic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, pp. 1468–1471.

[35] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, Aug. 2006, pp. 1136–1139.

[36] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[37] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[38] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 332–335.

[39] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 audiovisual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.

[40] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[41] S. Poria, E. Cambria, A. Hussain, and G. B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Netw.*, vol. 63, pp. 104–116, Mar. 2015.

[42] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[43] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using EEG signals and facial expressions," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[44] Y. Shu and S. Wang, "Emotion recognition through integrating EEG and peripheral signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2871–2875.

[45] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 6, pp. 1052–1063, Nov. 2011.

[46] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users' EEG and video content," *Multimedia Tools Appl.*, vol. 72, no. 2, pp. 1257–1283, 2014.

[47] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Comput. Hum. Behav.*, vol. 65, pp. 267–275, Dec. 2016.

[48] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5040–5043.

[49] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 374–393, Jul./Sep. 2019.

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

[51] K. Kulkarni, C. Corneanu, I. Ofodile, S. Escalera, X. Baró, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of facial displays of unfelt emotions," *IEEE Trans. Affective Comput.*, to be published.

[52] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Mar. 2015.

[53] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, Nov. 2007.

[54] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1299–1311, Jul. 2014.

[55] S. Nemati, "Canonical correlation analysis for data fusion in multimodal emotion recognition," in *Proc. 9th Int. Symp. Telecommun. (IST)*, Dec. 2018, pp. 676–681.

[56] L. Gao, R. Zhang, L. Qi, E. Chen, and L. Guan, "The labeled multiple canonical correlation analysis for information fusion," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 375–387, Feb. 2018.

[57] A. Puthenputhussery, Q. Liu, and C. Liu, "A sparse representation model using the complete marginal Fisher analysis framework and its applications to visual recognition," *IEEE Trans. Multimed.*, vol. 19, no. 8, pp. 1757–1770, Aug. 2017.

[58] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2811–2821, Nov. 2007.

[59] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Oct./Mar. 2012.

[60] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[61] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[62] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, p. 26, 2013.

[63] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[64] W. Abd-Almageed, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 3200–3203.

[65] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Process.*, vol. 93, no. 8, pp. 2140–2150, 2013.

[66] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2014, pp. 303–314.

[67] H.-B. Kang, "Affective content detection using HMMs," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 259–262.

[68] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proc. 1st Int. ACM Workshop Music Inf. Retr. User-Centered Multimodal Strategies*, 2011, pp. 7–12.

[69] F. Eyben, F. Weninger, N. Lehment, B. Schuller, and G. Rigoll, "Affective video retrieval: Violence detection in hollywood movies by large-scale segmental feature extraction," *PLoS ONE*, vol. 8, no. 12, 2013, Art. no. e78506.

[70] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden Markov models," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Berlin, Germany: Springer, 2007, pp. 594–605.

[71] A. Barjatya, "Block matching algorithms for motion estimation," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 225–239, Apr. 2004.

[72] M. Schröder, "Speech and emotion research : An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Phonus 7, Res. Rep. Inst. Phonetics, Saarland Univ., Saarbrücken, Germany, 2004.

[73] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
[74] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.
[75] S. Nemati and M. E. Basiri, "Particle swarm optimization for feature selection in speaker verification," in *Proc. Eur. Conf. Appl. Evol. Comput.* Berlin, Germany: Springer, 2010, pp. 371–380.
[76] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening," *Frontiers Neurosci.*, vol. 8, p. 94, May 2014.
[77] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image Vis. Comput.*, vol. 31, no. 2, pp. 164–174, Feb. 2013.
[78] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2376–2379.
[79] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Netw.*, vol. 18, no. 4, pp. 389–405, 2015.
[80] M. E. Basiri and A. Kabiri, "Words are important: Improving sentiment analysis in the persian language by lexicon refining," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 17, no. 4, 2018, Art. no. 26.
[81] M. E. Basiri and A. Kabiri, "Sentence-level sentiment analysis in Persian," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 84–89.
[82] M. E. Basiri and A. Kabiri, "Translation is not enough: Comparing lexicon-based methods for sentiment analysis in Persian," in *Proc. Int. Symp. Comput. Sci. Softw. Eng. Conf. (CSSE)*, Oct. 2017, pp. 36–41.
[83] M. E. Basiri and A. Kabiri, "Uninorm operators for sentence-level score aggregation in sentiment analysis," in *Proc. 4th Int. Conf. Web Res. (ICWR)*, Apr. 2018, pp. 97–102.
[84] P. J. Khiabani, M. E. Basiri, and H. Rastegari, "An improved evidence-based aggregation method for sentiment analysis," *J. Inf. Sci.*, Mar. 2019.
[85] S. Nemati, "OWA operators for the fusion of social networks' comments with audio-visual content," in *Proc. 5th Int. Conf. Web Res. (ICWR)*, Apr. 2019, pp. 90–95.

SHAHLA NEMATI was born in Shiraz, Iran, in 1982. She received the B.S. degree in hardware engineering from Shiraz University, Shiraz, in 2005, the M.S. degree from the Isfahan University of Technology, Isfahan, Iran, in 2008, and the Ph.D. degree in computer engineering from Isfahan University, Isfahan, in 2016.

Since 2017, she has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. She has written several articles in the fields of data fusion, emotion recognition, affective computing, and audio processing. Her research interests include data fusion, affective computing, and data mining.

REZA ROHANI was born in Shiraz, Iran, in 1985. He received the B.S. degree in computer science from Shahid Bahonar Kerman, Iran, in 2007, and the M.S. and Ph.D. degrees in artificial intelligence from Shiraz University, Iran, in 2010 and 2016, respectively.

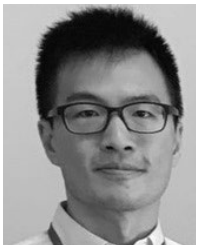
Since 2016, he has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. He has written several articles in the fields of emotion recognition, data fusion, and image processing. His research interests include data fusion, emotion recognition, and action recognition.

MOHAMMAD EHSAN BASIRI received the B.S. degree in software engineering from Shiraz University, Shiraz, Iran, in 2006, and the M.S. and Ph.D. degrees in artificial intelligence from Isfahan University, Isfahan, Iran, in 2009 and 2014, respectively.

Since 2014, he has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. He has authored three books and more than 35 articles. His research interests include sentiment analysis, natural language processing, machine learning, and data mining.

MOLOUD ABDAR received the bachelor's degree in computer engineering from Damghan University, Iran, in 2015, and the master's degree in computer science and engineering from the University of Aizu, Aizu, Japan, in 2018. He is currently pursuing the Ph.D. degree with the Université du Québec à Montréal, Montréal, QC, Canada.

He has written several articles in the fields of data mining, machine learning, and user modeling in some refereed international journals and conferences. His research interests include data mining, machine learning, ensemble learning, evolutionary algorithms, and user modeling. He was a recipient of the Fonds de Recherche du Quebec—Nature et Technologies Award (ranked 5th among 20 candidates in the second round of selection process), in 2019.



NEIL Y. YEN received the Ph.D. degree in human sciences from Waseda University, Japan, and the degree in engineering from Tamkang University, New Taipei, Taiwan, in 2012. His Ph.D. degree at Waseda University was funded by the Japan Society for the Promotion of Science under the RONPAKU Program.

He has been with the University of Aizu, Aizu, Japan, as an Associate Professor, since 2012. He has been involved extensively in an interdisciplinary field of research, where the themes are in the scope of big data science, computational intelligence, and human-centered computing. He is a member of the IEEE Computer Society, the IEEE System, Man, and Cybernetics Society, and the Technical Committee of Awareness Computing. He has been actively involved in the research community by serving as a Guest Editor, an Associate Editor, and a Reviewer for international refereed journals and as the Organizer/Chair for the ACM/IEEE-sponsored conferences, workshops, and special sessions.

VLADIMIR MAKARENKOV is currently a Full Professor and the Director of the Graduate Bioinformatics Program at the Department of Computer Science, Université du Québec à Montréal, Montréal, Canada. His research interests include bioinformatics, data mining, and software engineering.