Journal of Ambient Intelligence and Humanized Computing (2023) 14:1903–1917
https://doi.org/10.1007/s12652-021-03407-2

ORIGINAL RESEARCH

Multimodal emotion recognition based on feature selection and extreme learning machine in video clips

Bei Pan[1] · Kaoru Hirota[1] · Zhiyang Jia[1] · Linhui Zhao[2,3] · Xiaoming Jin[2,3] · Yaping Dai[1]

Received: 3 February 2021 / Accepted: 13 July 2021 / Published online: 27 July 2021

Abstract

Multimodal fusion-based emotion recognition has attracted increasing attention in affective computing because different modalities can achieve information complementation. One of the main challenges for reliable and effective model design is to define and extract appropriate emotional features from different modalities. In this paper, we present a novel multimodal emotion recognition framework to estimate categorical emotions, where visual and audio signals are utilized as multimodal input. The model learns neural appearance and key emotion frame using a statistical geometric method, which acts as a preprocesser for saving computation power. Discriminative emotion features expressed from visual and audio modalities are extracted through evolutionary optimization, and then fed to the optimized extreme learning machine (ELM) classifiers for unimodal emotion recognition. Finally, a decision-level fusion strategy is applied to integrate the results of predicted emotions by the different classifiers to enhance the overall performance. The effectiveness of the proposed method is demonstrated through three public datasets, i.e., the acted CK+ dataset, the acted Enterface05 dataset, and the spontaneous BAUM-1s dataset. An average recognition rate of 93.53% on CK+, 91.62% on Enterface05, and 60.77% on BAUM-1s are obtained. The emotion recognition results acquired by fusing visual and audio predicted emotions are superior to both recognition of unimodality and concatenation of individual features.

Keywords Emotion recognition · Multimodal fusion · Evolutionary optimization · Feature selection · Extreme learning machine

1 Introduction

Emotion is a significant part of our daily life that conveys the intention, mental state, and physical state of human beings (Zeng et al. 2008). With the fast development of the artificial intelligence, enabling the computer to recognize the human emotional state is with great importance to obtain more natural and better experience in human-computer interaction (Krithika and Priya 2020; Mendoza-Palechor et al. 2019). In general, emotions are conveyed mainly through facial expression and speech voice. As a result, a considerable amount of efforts have been made on emotion recognition based on individual facial expression or speech voice, and moreover, the combination of visual and audio modalities (Wang and Guan 2008). Among most of the existing studies, six principal emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise, are the major concerns.

Facial expression is primary and common signal for emotion recognition. An effective definition of visual features is a prerequisite for precise emotion recognition. Ekman and

✉ Zhiyang Jia
zhiyang.jia@bit.edu.cn

✉ Linhui Zhao
jdtlinhui@buu.edu.cn

Bei Pan
panbei@bit.edu.cn

Kaoru Hirota
hirota@jsps.org.cn

Xiaoming Jin
jinxm@buu.edu.cn

Yaping Dai
daiyaping@bit.edu.cn

[1] School of Automation, Beijing Institute of Technology, Beijing 100081, China

[2] College of Robotics, Beijing Union University, Beijing 100020, China

[3] Beijing Engineering Research Center of Smart Mechanical Innovation Design Service, Beijing 100020, China

Friesen (1978) developed the Facial Action Coding System (FACS) to reconstruct the facial expressions in terms of Action Units (AU), which is a foundation of facial expression feature extraction. Generally, visual feature extraction methods can be summarized into two categories, i.e., appearance feature and geometric feature (Chu 2017; Shan et al. 2009). The appearance feature is obtained through image filters and one of the most representative approaches is the local binary pattern (LBP) (Zhao and Pietikainen 2007). Geometric feature-based methods often exploit the geometric relationships among different facial components to describe an expressive face. Due to the eminent learning ability of deep learning (DL) (LeCun et al. 2015), researches studying DL algorithms in facial expression recognition are springing up (Jain et al. 2019; Liu et al. 2018). Chen et al. (2018a) proposed a SR-based deep sparse autoencoder network to recognize the facial expression, which uses a layered approach for extracting different levels of data features. In Xie et al. (2019), a deep attentive multi-path convolutional neural network (DAM-CNN) using the VGG-Face network is proposed to extract the advanced emotional features. Moreover, to learn high-level expression semantic features, Wu et al. (2019) proposed a weight-adapted CNN (WACNN) framework for facial emotion recognition.

In the view of audio emotion recognition, the prosodic, spectral, and voice features are often explored for emotion recognition (El Ayadi et al. 2011). Specifically, the prosodic features contain pitch period, energy, intensity, and duration time. In spectral features, Mel-Frequency Cepstral Coefficient (MFCC) is most commonly used to model the audio emotion recognition system. Formats, spectral energy distribution, and harmonics-to-noise-ration are the representative voice features. Those hand-crafted features are mostly considered as low-level features. To develop automatic feature learning techniques, researchers are paying much attention to utilize DL algorithms to obtain high-level features for speech emotion recognition (Akçay and Oğuz 2020). Han et al. (2014) proposed to employ a deep neural network (DNN) and extreme learning machine (ELM) to extract high-level features from low-level ones. Zhang et al. (2019) presented a multiscale deep convolutional long short-term memory (LSTM) framework for spontaneous speech emotion recognition where a deep CNN model is used to learn deep segment-level features.

For unimodal emotion recognition, it is required to extract appropriate visual or audio features and train the emotion classification model by using effective machine learning technology. Commonly used classifiers include hidden Markov model (HMM), support vector machine (SVM), artificial neural network (ANN), etc (Chen et al. 2018b). Considering that unimodality is sometimes insufficient to precisely recognize emotions, some other modalities that can offer supplementary information are also adopted to increase the recognition accuracy. In practice, researchers have made significant progress on multimodal emotion recognition (Busso et al. 2004; Pons and Masip 2020; Poria et al. 2016; Hossain and Muhammad 2019; Chen et al. 2016). It is worth noting that the fusion approaches on multimodal emotion recognition can be divided into three categories, i.e., feature-level, decision-level, and hybrid multimodal fusion (Poria et al. 2017). The key to feature-level fusion is to cascade the features extracted from different modalities as the input and send it into emotion classifiers (Zhang et al. 2017; Kansizoglou et al. 2019). For the decision-level fusion, the visual and audio modalities emotion classifiers are trained separately and the results of two classifiers are fused to further obtain the final emotion estimation (Bejani et al. 2014). While the hybrid multimodal fusion methods integrate the feature-level and decision-level fusion (Wöllmer et al. 2013).

As above mentioned, to achieve higher emotion recognition accuracy, it is important to extract appropriate features and exploit the emotion information, from different modalities and thus, finally integrate the complementary information. In this work, we propose a multimodal fusion framework for emotion recognition in video clips, which exploits the emotion information of visual and audio modalities. To address the visual modality, we firstly extract keyframes from a consecutive image sequence and define a geometric feature representation to detect the transformation of keyframes. Then the appropriate and informative facial emotion features are selected through evolutionary optimization to reduce feature dimension as well as improve learning speed. Once the critical facial features have been obtained, they are fed to the optimized ELM classifier for visual emotion recognition. Similar to the visual modality, key acoustic features are selected by evolutionary optimization and sent to the optimized ELM model for audio emotion recognition. Finally, a weighted fusion strategy is employed to integrate the visual and audio modalities for high recognition performance. The major contributions of this work are summarised as follows:

1. A framework for effectively extracting and fusing the information of visual and audio modalities to recognize emotion in the video clips is developed, which is applicable for both the acted and spontaneous emotion recognition.
2. The emotion feature and ELM model structures in visual and audio modalities are optimized simultaneously by balancing the emotion recognition accuracy and model complexity.
3. We demonstrate the performance of emotion recognition of the proposed multimodal fusion framework is superior than that based on the individual visual modality or audio modality through two video datasets.

The remainder of the paper is organized as follows: Sect. 2 briefly explains the principles of GA and ELM methods. The details of the proposed method are present in Sect. 3. Then, three emotional datasets are investigated to demonstrate the effectiveness and superiority of the proposed method in Sect. 4. Finally, conclusions and future work are given in Sect. 5.

2 Preliminaries

2.1 Genetic algorithms

Genetic algorithms (GA) (Whitley 1994) is one of the most influential evolutionary optimization algorithms inspired by the idea of Darwinian evolution. As a kind of stochastic search algorithms, GA enables the individual to act like a chromosome, and then execute iteration operation to search the optimal solution. During each iteration, candidate solutions undergo three main operations: selection, crossover, and mutation are executed orderly. The aim of selection is to directly inherit the optimized individuals or to produce new individuals through crossover to the next generation. The selection operation is based on the assessment of the fitness of the individuals in the population. During the crossover process, the offspring population is generated by crossing pairs of chromosomes in the current population. Mutation randomly changes some parts of the chromosomes to ensure the diversity of the new population. The basic operation process of genetic algorithm is summarized as follows:

1. Set the number of maximum evolution generation N_{gen} and randomly generate L individuals as the initial population.
2. Compute the fitness f of each individual and sort the population with descending order according to f.
3. Perform genetic operations of selection, crossover, and mutation to produce a new population.
4. Repeat (2) and (3) until the N_{gen} is reached and the best chromosome is obtained.

2.2 Extreme learning machine

Extreme learning machine (ELM) (Huang et al. 2006; Zhang et al. 2020a) is an efficient learning algorithm proposed for training single-hidden layer feedback networks (SLFNs). Different from traditional gradient-based iterative learning, ELM randomly chooses the weights of the hidden nodes and analytically determines the output weights of SLFNs. Since the iterative learning of parameters in the hidden layer is avoided, ELM model possesses an extremely fast training speed (Xiao et al. 2017; Zhang et al. 2020b). Thus, ELM classifier has gained much attention in emotion recognition applications because of the high computational efficiency and outstanding capability of generalization. Therefore, ELM is employed to classify emotions in this work. For a dataset $\{\mathbf{X}, \mathbf{Y}\}$ with N inputs and M output units, where $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right]$, $\mathbf{Y} = \left[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\right]$, the mathematical formulation of a basic ELM model can be described as:

$$\sum_{i=1}^{N_{node}} \theta_i g_i\left(\mathbf{x}_n\right) = \sum_{i=1}^{N_{node}} \theta_i g_i\left(\mathbf{w}_i \cdot \mathbf{x}_n + b_i\right) = \hat{\mathbf{y}}_n, \tag{1}$$

where N_{node} is the number of hidden nodes in the ELM, $\boldsymbol{\theta}_i = \left[\theta_{i1}, \theta_{i2}, \ldots, \theta_{im}\right]^T$ represents the output weight vector connecting the ith hidden node and output units. $g(\cdot)$ is the activation function. $\mathbf{w}_i = \left[w_{i1}, w_{i2}, \ldots, w_{in}\right]^T$ denotes the input weight vector connecting the input feature and hidden node, b_i is the bias of the ith hidden node, and $\hat{\mathbf{y}}_n$ is the predicted output vector.

An ideal ELM model is expected to find the least-squares solution to approximate the training data with zero error, i.e.

$$\sum_{i=1}^{N_{node}} \left\|\mathbf{y}_n - \hat{\mathbf{y}}_n\right\| = 0. \tag{2}$$

Thus, the parameters θ_i, $\mathbf{w}_i$, and b_i must be satisfied with

$$\sum_{i=1}^{N_{node}} \theta_i g_i\left(\mathbf{w}_i \cdot \mathbf{x}_n + b_i\right) = \mathbf{y}_n, \tag{3}$$

Then, ELM model can be rewritten as

$$\mathbf{H}\Theta = \mathbf{Y}, \tag{4}$$

$$\mathbf{H} = \begin{bmatrix} g\left(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1\right) & \cdots & g\left(\mathbf{w}_{N_{node}} \cdot \mathbf{x}_1 + b_{N_{node}}\right) \\ \vdots & \cdots & \vdots \\ g\left(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1\right) & \cdots & g\left(\mathbf{w}_{N_{node}} \cdot \mathbf{x}_N + b_{N_{node}}\right) \end{bmatrix}_{N \times N_{node}}, \tag{5}$$

$$\Theta = \left[\theta_1^T, \theta_2^T, \ldots, \theta_{N_{node}}^T\right]_{N_{node} \times M}, \tag{6}$$

where $\mathbf{H}$ stands for the hidden layer output matrix with N_{node} hidden nodes.

It is worth noting that the input weight and hidden bias are randomly generated and remain unchanged. Therefore, the key for training an ELM is to find a solution $\hat{\Theta}$ by minimizing the following cost function

$$\min_{\Theta} \|\mathbf{H}\Theta - \mathbf{Y}\|. \tag{7}$$

For many cases, the number of hidden nodes is much fewer than the number of training samples, which leads to a non-squared $\mathbf{H}$ matrix. Therefore, $\hat{\Theta}$ is estimated using the smallest-norm least-squares solution of the above linear system:

$$\hat{\Theta} = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{8}$$

where $\mathbf{H}^{\dagger}$ is the *Moore–Penrose generalized inverse* of $\mathbf{H}$. If the inverse of $\mathbf{H}^{T}\mathbf{H}$ exists, then $\mathbf{H}^{\dagger}$ can be estimated as:

$$\mathbf{H}^{\dagger} = \left(\mathbf{H}^{T}\mathbf{H}\right)^{-1}\mathbf{H}^{T}. \tag{9}$$

Finally, $\hat{\Theta}$ is given as follows:

$$\hat{\Theta} = \left(\mathbf{H}^{T}\mathbf{H}\right)^{-1}\mathbf{H}^{T}\mathbf{Y}. \tag{10}$$

3 Multimodal fusion-based method for emotion recognition

This section presents the details of our proposed method. The method to define and select visual and audio features are firstly introduced, and then the fusion strategy is described. The proposed multimodal fusion framework for emotion recognition is shown in Fig. 1. Specifically, to obtain high performance emotion recognition results in decision-level fusion, it is necessary to acquire eminent unimodal emotion recognition results. Therefore, high-performance visual and audio emotion recognition models are constructed. Then the individual classification results of the two modalities are integrated to enhance the final emotion recognition accuracy.

3.1 Visual feature

3.1.1 Keyframes extraction

Keyframes extraction is a fundamental procedure to pick out the representative frames from a consecutive image sequence for effectively emotion recognition (Noroozi et al. 2017). In general, the principle of the keyframe definition is that the number of keyframes should be small while the differences among keyframes should be large. To exploit the deformation of the different expressions in visual data, we focus on extracting the neutral frame and peak expression frame from an image sequence in this work. The frame without any expression of each subject is firstly picked out and fixed as the neutral frame for the reference of the expressive face.

Facial landmarks are widely used for facial expression recognition, which can describe the whole face by marking the eyebrows, eyes, nose, mouth, and chin regions. Given a set of facial landmarks $l = \left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_n, y_n\right)\right\}$, where $\left(x_i, y_i\right)$ denote the coordinates of the i-th facial landmark. In this study, the location information of 68 landmarks is used to compare the differences among successive frames in an image sequence. Initially, the face alignment is required to fix the orientations so that all the faces in the image sequence are straight. In our implementation, we use the open-source dlib face detection (Kazemi and Sullivan 2014) to locate the face bounding box and align face sequence. Then, facial landmark detector inside the dlib library is utilized to calculate the $\left(x_i, y_i\right)$-coordinates of 68 landmarks and the locations of facial landmarks in each frame are confirmed.

Generally, the facial region has structural symmetry in nature. If we coordinate the Y-axis with the bridge of the nose, the left part and right part of the face are almost identical when folding Y-axis. This property is found in both neutral and expressive faces. Therefore, we propose to compute the deformation on the left or right neutral and expressive faces to extract keyframes, which can dramatically decrease the computational cost compared with considering the whole face. Without loss of generality, the landmarks on the left

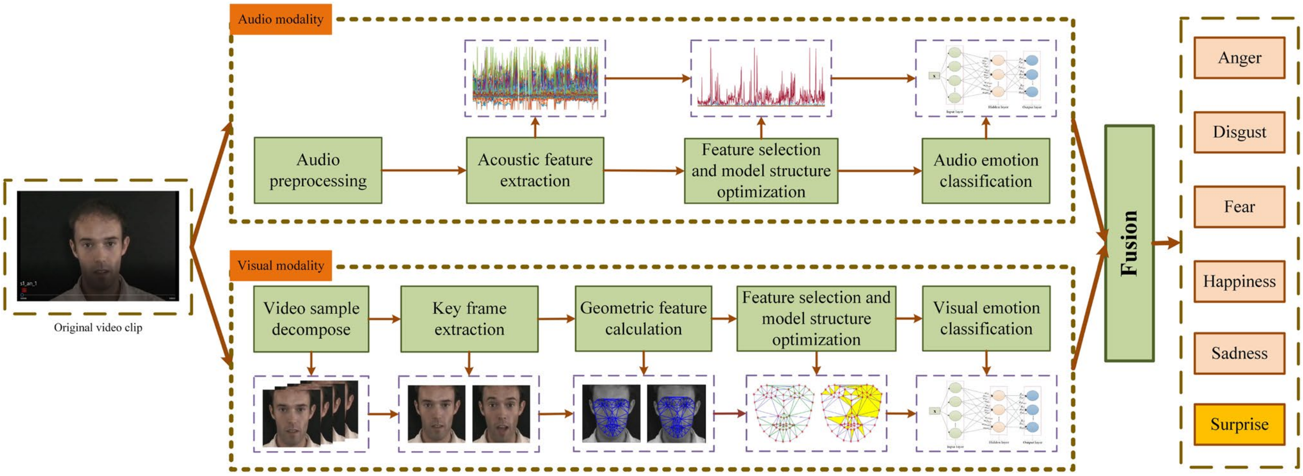

Audio modality
Audio preprocessing
Acoustic feature extraction
Feature selection and model structure optimization
Audio emotion classification
Original video clip
Visual modality
Video sample decompose
Key frame extraction
Geometric feature calculation
Feature selection and model structure optimization
Visual emotion classification
Fusion
Anger
Disgust
Fear
Happiness
Sadness
Surprise

Fig. 1 Proposed multimodal fusion framework for emotion recognition

part of the face are chosen as the specific landmarks in this study to extract keyframes.

Since the generation of an expression naturally yields geometrical transformation that can be reflected from the distance between two landmarks. It is feasible to evaluate the difference between two frames in one image sequence by computing the distance among specific landmarks in local regions. As shown in Fig. 2, to decrease the computational cost, we extract certain landmarks from five regions of each face image. In those regions, the changes in displacement between points are discriminative. The landmark pairs used for distance calculation are listed in Table 1 [For example, (18,19) indicates the index pair of landmark 18 and landmark 19]. The distance between two specific landmarks is calculated as:

$$d_{i,j} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{11}$$

Vector of differences calculated from specific landmarks for each frame is presented as $\mathbf{d} = \left[d_1, d_2, ..., d_K\right]$, as a result, 39 features obtained for keyframes extraction in this work. To select the peak frame for each video, the sum of differences between the neutral face and the expressive one is calculated, and the frame with the maximum is selected. It is formulated as:

$$d_{\underset{max}{o,i}} = \frac{1}{K}\sum_{k=1}^{K}\left(d_{o,k} - d_{i,k}\right), \tag{12}$$

where $d_{o,k}$ and $d_{i,k}$ are the differences of k-th pair of specific landmarks of neutral face and i-th frame, respectively.

Table 1 Landmark pairs used to calculate the distance for keyframe extraction

Region	Distance $d_i - d_j$
Eyebrow	(18,19),(19,20),(20,21),(21,22)
	(18,37),(20,38),(21,39),(22,40)
Eye	(37,38),(38,39),(39,40),(40,41)
	(41,42),(42,37)
Nose	(32,33),(33,34)
Mouth	(49,50),(50,51),(51,52),(58,59)
	(59,60),(60,49),(49,61),(50,61)
	(60,61),(61,62),(61,68),(51,62)
	(62,68),(68,59),(62,63),(68,67)
	(52,63),(63,67),(67,58),(52,58)
Chin	(6,7),(7,8),(8,9)

3.1.2 Geometric deformation features

Once keyframes are extracted from each image sequence, the visual features contained facial expression information should be investigated from images to classify different emotions. It is noteworthy that, compared to the appearance feature, the geometric feature is more efficient to track and is not restricted by the light. Therefore, we focus on investigating geometric features to capture the emotion information of visual modality. For each emotion, the coordinates of facial landmark change from the neutral face to the expressive face because of the facial muscle movement. Thus, the geometric features are defined by calculating the changes among landmarks. The face can be divided into numbers of non-overlap sub-regions by connecting landmarks according to Delaunay triangulation. The triangular mesh connected by landmarks in neutral and expressive faces is illustrated in Fig. 3, where the red dot and blue line stand for the landmark and the side of a triangle respectively.

It can be seen from Fig. 3 that for a pair of edges composed of the same two landmarks in the neutral and expressive faces, facial muscle movements induce the deformation of edges. Therefore, we firstly assume that the edges in the mesh are independent. The difference between the edge in the neutral face and the corresponding edge in the expressive face is estimated as one of the geometric deformation features. Additionally, considering the correlation of edges in a triangle, the triangles in the whole face are regarded as sub-blocks geometric features. As shown in Fig. 3, for a pair of green triangles composed of the same three landmarks in

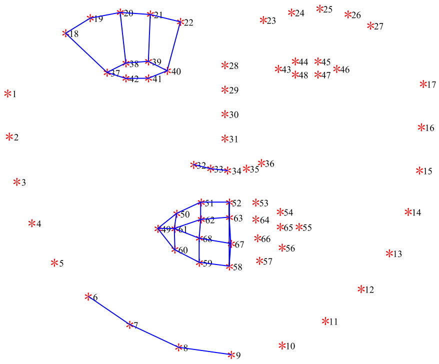


Fig. 2 Annotated facial landmarks and certain line segments used for deformation calculation

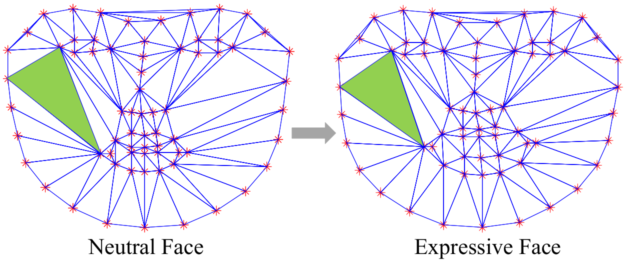


Fig. 3 Triangular mesh connected by facial landmarks

the neutral and corresponding expressive faces, the area of a triangle in the expressive face changes during expression generation. Specifically, the distance $E_{n,f}$ between two edges formed by two landmarks in the neutral face and corresponding expressive face is taken as an edge feature. Besides, the difference $A_{n,f}$ between the areas of triangles in the neutral face and the corresponding expressive face is calculated as the area feature. The two kinds of geometric features are calculated as:

$$E_{n,f} = d_{n,ij} - d_{f,ij}, \tag{13}$$

$$A_{n,f} = A_{n,ijk} - A_{f,ijk}, \tag{14}$$

$$A_{n,ijk} = \frac{1}{4}\sqrt{\left(d_{n,ij} + d_{n,ik} + d_{n,jk}\right)\left(d_{n,ij} + d_{n,ik} - d_{n,jk}\right)} \sqrt{\left(d_{n,ij} - d_{n,ik} + d_{n,jk}\right)\left(d_{n,ik} + d_{n,jk} - d_{n,ij}\right)}, \tag{15}$$

where $d_{n,ij}$ and $d_{f,ij}$ are the edge of a triangle in the neutral face and the corresponding edge in the expressive face, respectively. $A_{n,ijk}$ and $A_{f,ijk}$ denote the areas of triangles in the neutral face and the corresponding expressive face, respectively.

3.1.3 Emotion feature selection and model structure optimization by GA

Generally, the features defined from original data contain relevant features as well as irrelevant features. To construct a more accurate emotion recognition model, it is critical to execute feature selection so that extracting the appropriate features and removing the irrelevant ones. Although many researches have been conducted on features selection and dimensionality reduction for emotion recognition, the ignorance of correlation between the feature and actual emotion makes them suffer from low recognition accuracies. Therefore, in the current work, we aim to remove the useless features and extract discriminative features by considering the actual target in the learning algorithm for a better emotion recognition rate.

Actually, the selection of geometric features is essentially a binary optimization problem, where "0" and "1" denote the removal and involvement of features, respectively. As described in Sect. 2.1, genetic algorithms (GA) possess good performance in searching for the optimal solution and is expected to select the appropriate emotional features in this work. Thus, GA is carried out to select the appropriate features by searching the optimal emotional features that contribute to distinguish different emotions directly.

In addition to emotion feature, model structure is also an important factor that influence the model performance as well as complexity. To enhance the recognition accuracy and reduce the complexity of visual model, it is desirable to optimize the structure of emotion recognition model. With the advantages of extremely fast learning speed and prominent generalization performance of extreme learning machine (ELM), the multi-class ELM is applied for recognizing emotions in this study. By analyzing ELM structure, it is easy to see that the number of input features and nodes in hidden layer influence the model structure directly. Therefore, in this paper, we propose to optimize the emotion feature selection and model structure simultaneously through GA to ensure the high recognition performance of visual emotion model.

In GA optimization, the geometric features are encoded as a binary string, where each bit denotes whether the corresponding feature is selected or not. While the determination of the number of hidden nodes is an integer optimization problem. Consequently, the decision vector consists of the geometric features and the number of hidden nodes. Let $S = \left\{s_1, s_2, ..., s_N, N_{node}\right\}$ be the decision vector for feature selection and model structure optimization, where s_n is the geometric feature and N_{node} is the number of hidden nodes. Though the proposed emotion recognition method for visual modality can benefit from the strong adaptability and flexibility of GA optimization, there still exists a high risk of overfitting. GA may cause overfitting in the training set if the decision variables are over-optimized. To deal with this issue, an independent validation set is introduced to optimize parameters. Moreover, to simplify the calculation, we convert the multiobjective optimization problem into a single-objective optimization problem. Therefore, the objective function consists of two parts, i.e., the recognition error and model complexity measured by the connections. The objective function estimated from the validation set is formulated as:

$$f(S) = \alpha \cdot \frac{1}{\text{Accuracy}_{val}} + (1 - \alpha)\left(N_{node} \cdot N_{fea}\right), \tag{16}$$

where α is the weight for recognition error, the Accuracy_{val} stands for the recognition accuracy of the validation set. N_{fea} denotes the number of selected emotion features. The GA algorithm evolves the chromosomes to find the optimal subset of expression features by minimizing the fitness function.

To enable the GA optimization, the decision vector is first encoded as chromosome, and an initial population with N_{pop} individuals is generated randomly. Then, the objective function is calculated for each individual in the population, and an offspring population is generated by executing evolutionary operations, i.e., tournament selection, crossover, and mutation. Next, a new population is produced by merging the offspring and parent populations, and this is followed by non-dominated ranking and trimming of individuals. The above steps are repeated until the stopping conditions are satisfied. Finally, the best solution characterized by different geometric features and the optimal number of hidden nodes is obtained. Figure 4 shows the optimization diagram for emotional feature selection and ELM model structure using GA.

3.2 Audio feature

Audio modality, as another expression of emotion, can offer complementary and useful information in addition to the visual modality. As a result, audio modality contributes to dramatically increase the individual performance of emotion recognition in video clips. In this study, the visual modality and audio modality are processed in parallel to enhance the recognition efficiency of the model. To identify emotion through voice, 1582 acoustic features containing energy/spectral low-level descriptors, voice-related low-level descriptors are extracted from each video. The extracted acoustic features are given in Table 2.

Because of the existence of redundant and useless features, the efficiency of audio-based emotion recognition will be discounted. Therefore, to increase the accuracy of emotion recognition of audio modality and decrease the computation cost, we propose to optimize the acoustic features to select a set of significant features that are strongly related to the speech emotion. For audio modality, GA is utilized again in acoustic features' selection and ELM model structure optimization in this work. Also, the decision vector is composed of acoustic features and the number of hidden nodes. The fitness function is defined as the reciprocal of the accuracy and the number of connections in ELM of an individual validation set. To avoid redundancy, the detailed procedure is omitted.

Table 2 Acoustic feature: 38 low-level descriptors with regression coefficients and 21 functionals Schuller et al. 2010

Descriptors	Functionals
PCM loudness	Position max./min.
MFCC [0–14]	arith. mean, std. deviation
log Mel Freq. Band [0–7]	Skewness, kurtosis
LSP Frequency [0–7]	lin. regression coeff.
F0 by Sub-Harmonic Sum.	lin. regression error
F0 Envelope	Quartile
Voicing Probability	Quartile range
Jitter local	Percentile
Jitter DDP	Percentile range
Shimmer local	Up-level time

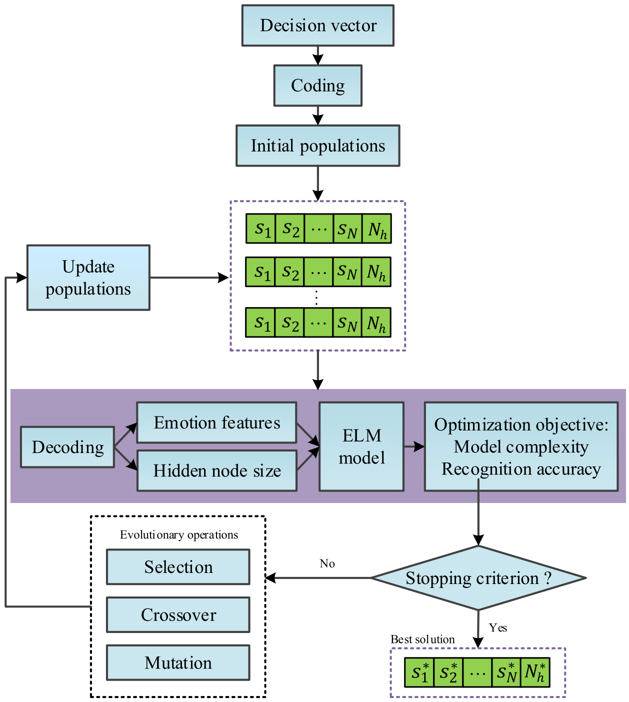


Fig. 4 Optimization diagram for emotional feature selection and ELM model structure using GA

3.3 Multimodal fusion

In 3.1 and 3.2, we present the approaches to select crucial features and optimize ELM models used to predict the emotions of visual and audio modalities. In order to obtain more accurate and robust recognition results, the knowledge from different modalities are combined in this work. It is noticeable that with decision-level fusion in emotion recognition, the facial expression features and acoustic features do not need to be synchronized compared with feature-level fusion. Therefore, the decision-level fusion strategy is carried out in this work. To realize decision-level fusion, the significant geometric and acoustic features need to be separately fed to the corresponding ELM classifiers with radial basis function kernel to recognize emotions. The outputs of each model indicate the probabilities that the emotion expressed by the subject belongs to different emotions.

Once the classification results of visual and audio modalities are obtained, the next critical task is to fuse the two modalities to build the final recognition model. In this study, a weighted fusion strategy is adopted to achieve the combination of expression related information obtained from the two modalities. The basic idea is to assign visual modality and audio modality different weights according to their importance for emotion recognition, which can be expressed as follows:

$$C\left(y_v, y_a\right) = \max_i \left(\beta p_v(i) + (1-\beta) p_a(i)\right), \tag{17}$$

where y_v and y_a denote the classification results of visual modality and audio modality, respectively. β is the weight that reflect the importance of visual modality. $p_v(i)$ and $p_a(i)$ are posterior probabilities of i-th class of visual modality and audio modality, respectively.

4 Experiments on three datasets

In this section, we present the results by using our proposed method on the three public databases, i.e., the Extended Cohn–Kanade (CK+) (Lucey et al. 2010), Enterface05 (Martin et al. 2006), and BAUM-1s (Zhalehpour et al. 2016). The unimodality results considering the defined geometric features and acoustic features without relevant feature selection and model structure optimization are shown separately. Then the results of emotion recognition with the selected features by GA of two modalities are shown respectively. Finally, the emotion recognition results of two modalities fusion are given.

4.1 Datasets and setup

CK+: The CK+ dataset contains 593 image sequences from 123 subjects. There are 327 image sequences with seven emotion labels, i.e., anger, contempt, disgust, fear, joy, sadness, and surprise. This dataset focuses on facial expression recognition and all the face emotions are lab-controlled. We aim to recognize six emotions in this dataset. The aligned and cropped facial images are shown in Fig. 5.

Enterface05: The Enterface05 dataset contains six emotions, i.e., anger, disgust, fear, joy, sadness, and surprise, which are posed by 43 subjects with 14 different nationalities. 1290 video samples are included in the dataset. Each audio sample rate is 48,000 Hz. Figure 6 presents samples of the aligned and cropped facial images from the Enterface05 dataset.

BAUM-1s: The BAUM-1s spontaneous dataset contains 1222 video clips from 31 Turkish subjects. The dataset is collected in real scenarios with spontaneous emotion expressions, which contains six basic emotions (anger, disgust, fear, happiness, sadness, surprise) as well as boredom and contempt. It also includes four mental states, i.e., unsure, thinking, concentrating, and bothered. Similar to the work(Hossain and Muhammad 2019), we focus on recognizing the six basic emotions, producing 521 video samples in total. Samples of the aligned and cropped facial images on the BAUM-1s dataset are shown in Fig. 7.

Fig. 5 Samples of the aligned and cropped facial images on CK+ dataset

For training our proposed model without feature selection, each dataset is trained and tested using Leave-One-Speakers-Group-Out (LOSGO) cross-validation. Moreover, to train the proposed feature selection model, we divide the three datasets into three sets: 50% for model training, 25% for parameter optimization, and 25% for model testing, respectively. In addition, appropriate value of the critical hyperparameter, i.e., the maximal number of hidden node N_{node}^{max}, the population size N_{pop} and maximal generation N_{gen} should be predetermined. Those hyperparameters are selected by trial and error as follows, $N_{node}^{max} = 30$, $N_{pop}^{visual} = 300$ for visual modality, $N_{pop}^{audio} = 2000$ for audio modality, $N_{gen} = 100$. To balance model structure and emotion recognition accuracy, α is equal to 0.5 in objective function.

4.2 Experimental results and snalysis

In this section, we present the experimental results on the three datasets, respectively. The results of emotion recognition based on facial expression three datasets are displayed separately. Then, the recognition results of audio modality on Enterface05 and BAUM-1s datasets are shown respectively. Finally, the fusion performance of two modalities on Enterface05 and BAUM-1s is given.

4.2.1 Visual modality performance

The defined geometric deformation features are firstly computed in three datasets individually. A total of 105 triangles and 169 non-redundant edges are used as two basic geometric features in each facial image. Concatenating the area

Fig. 6 Samples of the aligned and cropped facial images on Enterface05 dataset

Fig. 7 Samples of the aligned and cropped facial images on BAUM-1s dataset

Table 3 Average recognition rate (%) of visual and audio modalities with and without feature selection by GA on three datasets

Unimodality	Feature	CK+	Enterface05	BAUM-1s
Visual	V_{all}	79.27	31.15	48.94
	$V_{GA-Selection}$	93.53	86.65	55.38
Audio	A_{all}	–	61.85	48.75
	$A_{GA-Selection}$	–	74.85	53.08

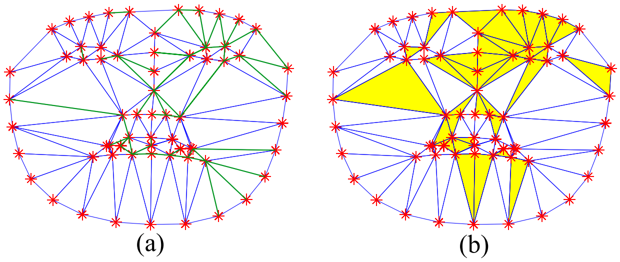


Fig. 8 Selected visual features through GA on the CK+ dataset

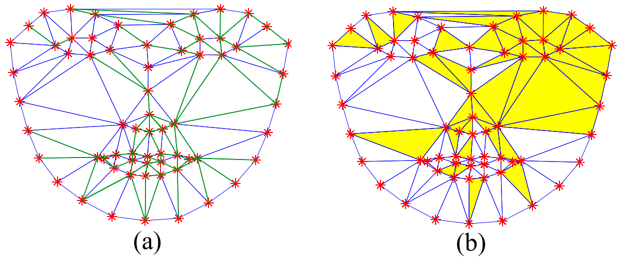


Fig. 9 Selected visual features through GA on the Enterface05 dataset

and edge features, 274 geometric features are obtained in total. Then GA is carried out to select the critical geometric features and optimal number of hidden nodes. Finally, the original defined geometric features and the selected geometric features by GA optimization are sent into the optimized multi-class ELM classifiers to recognize the six emotions separately.

As shown in Table 3, the recognition rates of facial repression classifiers without feature selection are 79.27%, 31.15%, and 48.94% on CK+, Enterface05, and BAUM-1s datasets, respectively. While 93.53 %, 86.65%, and 55.38% recognition rates are achieved on these datasets with GA feature selection. This table reveals that, as it was expected, feature selection by GA optimization contributes to improving the emotion recognition rates significantly.

Moreover, the selected visual features on the three datasets are given in Figs. 8, 9, and 10, respectively. It is noted that the green lines in Figs. 8a, 9a, and 10a denote the selected edge features on the three datasets. The yellow triangles in Figs. 8b, 9b, and 10b stand for the selected area features on the three datasets. For CK+ dataset, 94 visual features are selected through GA optimization. For the Enterface05 dataset, the numbers of visual features before and after GA optimization are 274 and 107. The number of visual features is reduced to 116 on the BAUM-1s dataset. These figures reveal that the local area in the five regions of a face gives valuable information for emotion classification. Moreover, it is convenient and intuitive to learn which facial areas give better emotion discrimination compared with the deep learning-based method. Besides, the number of hidden nodes optimized by GA is 12 for CK+, 10 for Enterface05, and 21 for BAUM-1s. The introduction of GA optimization for visual feature selection and ELM model optimization can dramatically decrease the complexity emotion recognition model and increase the classification accuracy.

The experiment results of visual emotion recognition on three datasets compared with previous methods are listed in Table 4. The bold values stand for the recognition results of our proposed method. The facial expression features and model structure are simultaneously optimized through GA in our method. It can be observed from Table 4 that for

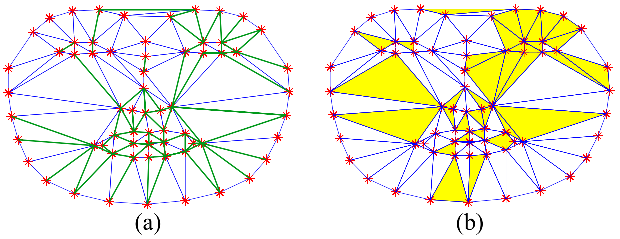


Fig. 10 Selected visual features through GA on the BAUM-1s dataset

Table 4 Average recognition rate of visual modality compared with previous works on three datasets

Dataset	Method	Accuracy (%)
CK+	Chen et al. (2016) (GWF)	89.00
	Jung et al. (2015) (DTAN)	91.44
	Jain et al. (2019)	93.24
	Ours	**93.53**
Enterface05	Zhang et al. (2017)	54.35
	Rahdari et al. (2019)	62.80
	Ma et al. (2019)	58.19
	Avots et al. (2019)	48.31
	Miyoshi et al. (2021)	49.26
	Ours	**86.65**
BAUM-1s	Zhang et al. (2017)	50.11
	Zhalehpour et al. (2016)	45.04
	Ma et al. (2019)	54.69
	Ours	**55.38**

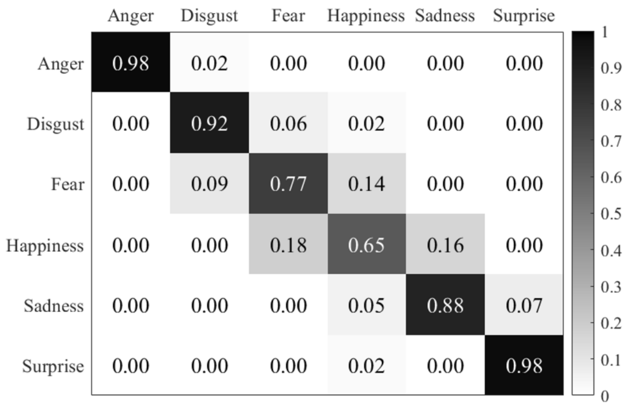
visual modality, our method outperforms the state-of-the-art methods, which include certain deep learning-based frameworks. The results demonstrate that the selected geometric deformation features are available to capture the vital emotion information during expression generation. Besides, the appropriate ELM model structure can help to enhance the recognition performance of visual modality.

Moreover, the classification confusion matrices of visual modality on the three datasets are shown in Figs. 11, 12, and 13, respectively. As can be observed, in the facial domain, all emotions are recognized with more than 65% accuracy on the CK+ and Enterface05 datasets. Anger, happiness, and sadness are much easier classified than the other three emotions on BAUM-1s. Notice that, it is failed to separate fear from anger, sadness, and disgust on the BAUM-1s dataset. Besides, comparing the three confusion matrices, it is clear that the acted emotions are much easier to recognize than the spontaneous ones.

4.2.2 Audio modality performance

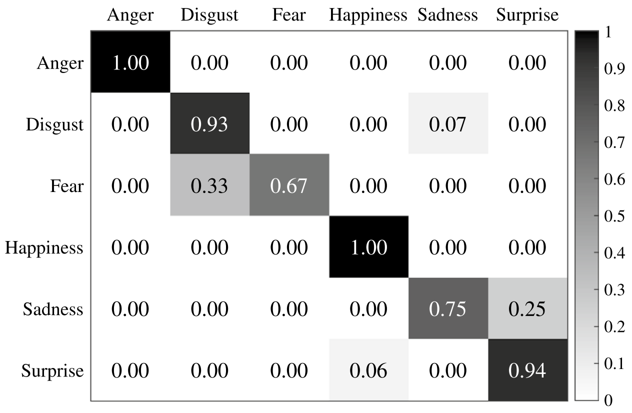
The acoustic features are firstly extracted by applying the OpenSMILE software and 1582 acoustic features are produced. Similar to the visual modality, the acoustic features with and without feature selection and model structure optimization by GA are used for emotion recognition individually and the results are presented in Table 3. In general, the results are similar to those in the visual modality. It is clear to see that the proposed method achieves an improvement from 61.85% to 74.85% for the Enterface05 dataset in audio modality. On the BAUM-1s dataset, the accuracy is improved from 48.75 to 53.08%. The results demonstrate that our proposed method contributes to improving emotion recognition accuracy for both visual and audio modalities. Besides, the dimensions of acoustic features are decreased from 1582 to 762 for the Enterface05 dataset. On the BAUM-1s dataset, the dimension of the acoustic feature is reduced to 680. From Table 3, it is suggested that the accuracies of the acoustic classifier are smaller than that of the facial expression classifier. Moreover, the optimized number of hidden nodes in ELM for Enterface05 and BAUM-1s are 12 and 20, respectively.

Also, the comparison results of our proposed audio feature selection method with previous audio emotion recognition methods are given in Table 5. On the Enterface05 and BAUM-1s datasets, our method with bold values in Table 5 achieves better recognition accuracy and it demonstrates that it is feasible to select crucial features and optimize model structure by applying the GA algorithm in audio modality. Besides, the classification confusion matrices of audio modality on the two datasets are presented in Figs. 14 and 15. Notice that, in the acoustic domain, anger, sadness, and surprise can be recognized with high accuracies, while disgust, fear, and happiness are slightly worse classified on the Enterface05 dataset. For the BAUM-1s dataset, happiness and sadness achieve



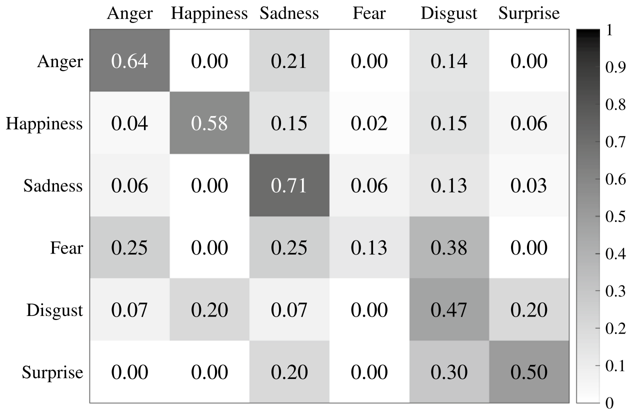
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.98	0.02	0.00	0.00	0.00	0.00
Disgust	0.00	0.92	0.06	0.02	0.00	0.00
Fear	0.00	0.09	0.77	0.14	0.00	0.00
Happiness	0.00	0.00	0.18	0.65	0.16	0.00
Sadness	0.00	0.00	0.00	0.05	0.88	0.07
Surprise	0.00	0.00	0.00	0.02	0.00	0.98

Fig. 12 Confusion matrix of visual modality with GA optimized feature selection on the Enterface05 dataset



	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	1.00	0.00	0.00	0.00	0.00	0.00
Disgust	0.00	0.93	0.00	0.00	0.07	0.00
Fear	0.00	0.33	0.67	0.00	0.00	0.00
Happiness	0.00	0.00	0.00	1.00	0.00	0.00
Sadness	0.00	0.00	0.00	0.00	0.75	0.25
Surprise	0.00	0.00	0.00	0.06	0.00	0.94

Fig. 11 Confusion matrix of visual modality with GA optimized feature selection on the CK+ dataset

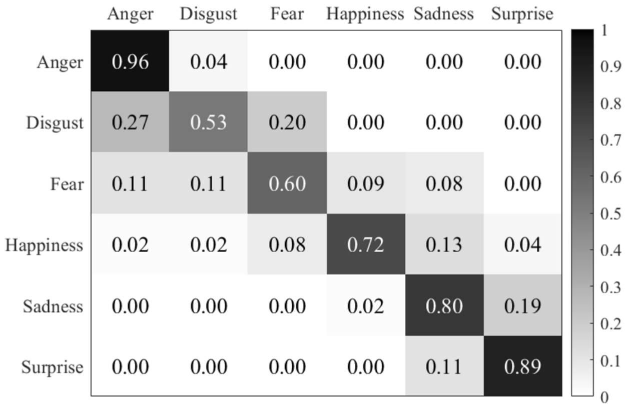


	Anger	Happiness	Sadness	Fear	Disgust	Surprise
Anger	0.64	0.00	0.21	0.00	0.14	0.00
Happiness	0.04	0.58	0.15	0.02	0.15	0.06
Sadness	0.06	0.00	0.71	0.06	0.13	0.03
Fear	0.25	0.00	0.25	0.13	0.38	0.00
Disgust	0.07	0.20	0.07	0.00	0.47	0.20
Surprise	0.00	0.00	0.20	0.00	0.30	0.50

Fig. 13 Confusion matrix of visual modality with GA optimized feature selection on the BAUM-1s dataset

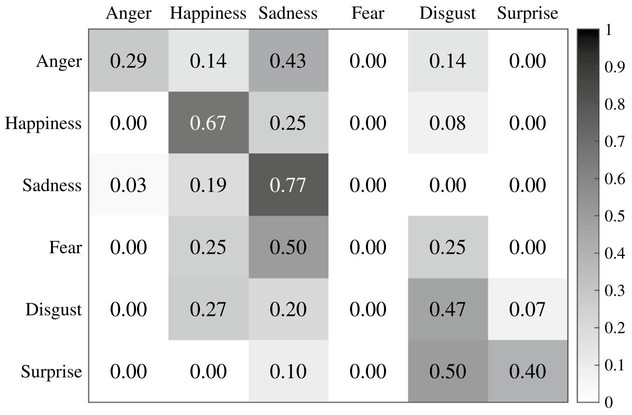
Table 5 Average recognition rate of audio modality compared with previous works on two datasets

Dataset	Method	Accuracy (%)
Enterface05	Zhalehpour et al. (2016)	72.95
	Avots et al. (2019)	50.22
	Ours	**74.85**
BAUM-1s	Zhang et al. (2017)	42.26
	Zhalehpour et al. (2016)	29.41
	Ma et al. (2019)	42.38
	Ours	**53.08**



	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.96	0.04	0.00	0.00	0.00	0.00
Disgust	0.27	0.53	0.20	0.00	0.00	0.00
Fear	0.11	0.11	0.60	0.09	0.08	0.00
Happiness	0.02	0.02	0.08	0.72	0.13	0.04
Sadness	0.00	0.00	0.00	0.02	0.80	0.19
Surprise	0.00	0.00	0.00	0.00	0.11	0.89

Fig. 14 Confusion matrix of audio modality emotion recognition results with features selection by GA on the Enterface05 dataset



	Anger	Happiness	Sadness	Fear	Disgust	Surprise
Anger	0.29	0.14	0.43	0.00	0.14	0.00
Happiness	0.00	0.67	0.25	0.00	0.08	0.00
Sadness	0.03	0.19	0.77	0.00	0.00	0.00
Fear	0.00	0.25	0.50	0.00	0.25	0.00
Disgust	0.00	0.27	0.20	0.00	0.47	0.07
Surprise	0.00	0.00	0.10	0.00	0.50	0.40

Fig. 15 Confusion matrix of audio modality emotion recognition results with features selection by GA on the BAUM-1s dataset

higher recognition rates than other emotions. “Fear” expressed through speech is failed to be classified because fear, sadness, disgust, and happiness have similar patterns observed in acoustic parameters. Therefore, it is expected that the recognition rate of fear emotion can be improved by fusing facial expression and acoustic information properly.

4.2.3 Multimodal performance

After training classifiers for visual and audio modality separately, the confidence values on the Enterface05 and BAUM-1s datasets are obtained. Next, the decision-level fusion method is applied to fuse the classification results of two modalities with GA optimization. To reduce the model complexity, a fusion strategy that weighs the two modalities according to their importance is used in this study. The weights of different modalities are defined as listed in Table 6. The performance of each pair of weights w_i with GA feature optimization is investigated and shown in Fig. 16. From the figure, it can be seen when the weights of visual modality and audio modality are 0.6 and 0.4 respectively on the Enterface05 dataset, the model possesses the best performance. While on the BAUM-1s dataset, the best performance is obtained when the weight equals to 0.2 or 0.4 for visual modality.

Besides, to compare the performance of decision-level fusion with feature-level fusion, the experiments of feature-level fusion are conducted on the Enterface05 and BAUM-1s datasets. Especially, two different feature-level fusion approaches are implemented in this work. One is to concatenate the raw geometric features and acoustic features directly and apply GA for feature selection as executed in visual and audio modalities. Then the selected emotion features are input to the ELM for classification. The other is to concatenate the selected visual feature and acoustic feature and use ELM for emotion recognition. The performance of emotion recognition based on two fusion methods is shown in Table 7. Notice that all these feature-level fusion methods give inferior performance than the combination of two classifiers in decision-level fusion. This table reveals that the best multimodal fusion approach in this work is the decision-level fusion because the synchronization of visual and acoustic features is free from the decision-level fusion.

The proposed method is compared with previous methods on the Enterface05 and BAUM-1s datasets separately. The results presented in Table 8 show that our multimodal fusion method is competitive with previous methods and the overall

Table 6 The weights of different modalities for fusion

Modality	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
Visual	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Audio	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

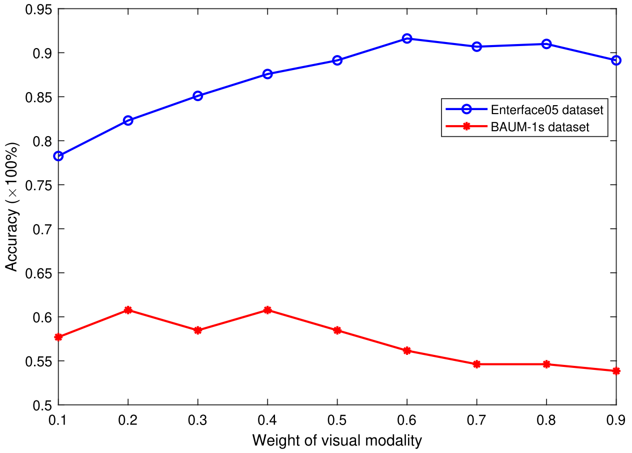


Fig. 16 Comparison of different weights of visual modality for multimodal fusion emotion recognition with GA optimized features selection on two datasets

Table 7 Multimodal emotion recognition performance (%) comparison at decision-level and feature-level fusion on two datasets

Fusion method	Enterface05	BAUM-1s
Feature-level1	49.15	45.38
Feature-level2	72.05	54.57
Decision-level	91.62	60.77

Feature-level1: concatenate the raw visual and acoustic features for GA selection. Feature-level2: concatenate the selected visual and acoustic features

Table 8 Average recognition rate of our proposed multimodal fusion method and previous works on two datasets

Dataset	Method	Accuracy (%)
Enterface05	Hossain and Muhammad (2019)	86.40
	Zhang et al. (2017)	54.57
	Bejani et al. (2014)	77.78
	Ma et al. (2019)	85.69
	Ours	**91.62**
BAUM-1s	Zhang et al. (2017)	54.57
	Zhalehpour et al. (2016)	51.29
	Ma et al. (2019)	59.17
	Ours	**60.77**

recognition rates are 91.62% and 60.77% on the Enterface05 and BAUM-1s datasets, respectively. Besides, the recognition result of the multimodal fusion is better than that obtained either from visual modality or audio modality in our framework. These promising results illustrate that even though audio modality gives inferior classification results than the visual modality, the acoustic features contain valuable information that cannot be extracted from visual modality. Also, it proves that visual and audio modalities are complementary to each other and the emotion classification results can be distinctly enhanced by integrating the two modalities. The improved emotion recognition rates also demonstrate that our proposed method is applicable to both acted and spontaneous emotions.

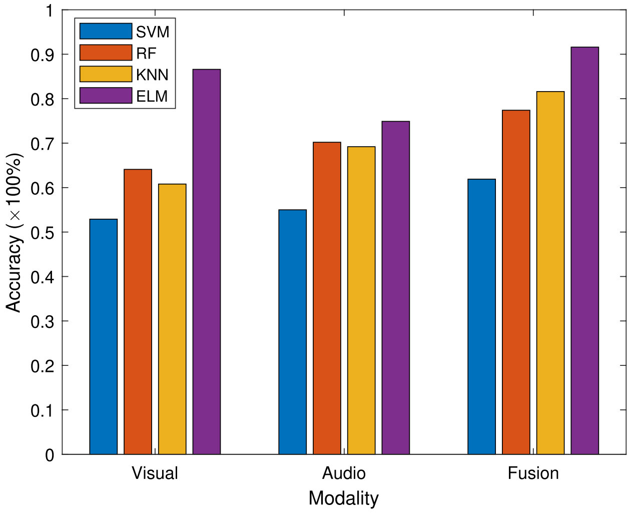


Fig. 17 Performance comparison of different classifiers on Enterface05

To compare the recognition performance of different classifiers, we conduct experiments using support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and ELM classifiers to train models on the same training and test sets. The recognition accuracies for unimodality and multimodality by different classifiers on the Enterface05 and BAUM-1s datasets are listed in Figs. 17 and 18. It is shown that, compared with other three classifiers for the proposed emotion recognition framework, ELM shows the best recognition performance on both unimodality and multimodality.

To further measure the multimodality emotion recognition performance, we compute precision, recall and F-score separately on the Enterface05 and BAUM-1s datasets. The experiment results are listed in Table 9 and Table10, respectively. The results indicate that "fear" emotion is difficult to recognize than other emotions. Comparing those two tables, it is clear to see that the posed emotions are easier to recognize than the spontaneous ones. Thus, there is still ample space to improve the performance of recognition rate of spontaneous emotions.

Moreover, it is worth noting that the proposed model is computationally efficient compared with the multimodal fusion frameworks using deep learning algorithms for emotion recognition. Specifically, it takes more than 18 h for model training using deep learning-based methods, on the computer with Intel(R) Core(TM) i7-8565U CPU, and 24GB RAM. On the same computer, the training time of

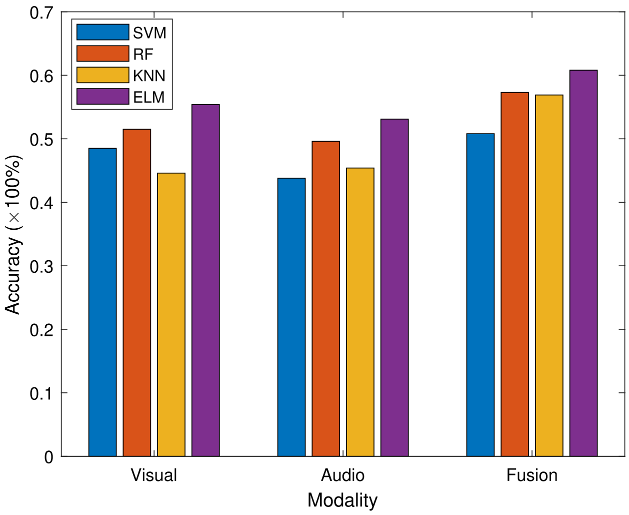


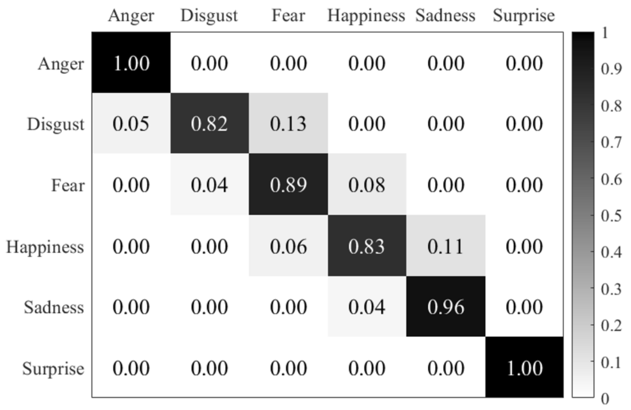
Fig. 18 Performance comparison of different classifiers on BAUM-1s

Table 9 Multimodal performance measure (%) for each emotion with the recognition accuracy of 91.62% on Enterface05

Emotion	Precision	Recall	F-score
Anger	94.64	100.0	97.25
Disgust	95.74	82.23	88.47
Fear	82.46	89.15	85.67
Happiness	88.00	83.04	85.45
Sadness	89.66	96.38	92.90
Surprise	93.05	100.0	96.40

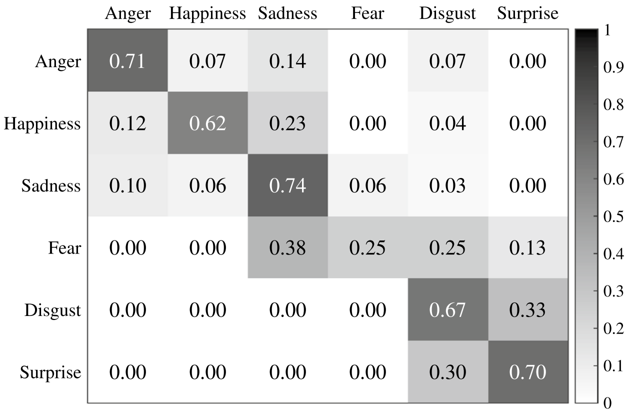
Table 10 Multimodal performance measure (%) for each emotion with the recognition accuracy of 60.77% on BAUM-1s

Emotion	Precision	Recall	F-score
Anger	75.00	71.48	73.20
Disgust	45.00	67.13	53.88
Fear	12.54	25.42	16.79
Happiness	76.92	62.37	68.89
Sadness	48.94	74.05	58.93
Surprise	66.67	70.21	68.39



	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	1.00	0.00	0.00	0.00	0.00	0.00
Disgust	0.05	0.82	0.13	0.00	0.00	0.00
Fear	0.00	0.04	0.89	0.08	0.00	0.00
Happiness	0.00	0.00	0.06	0.83	0.11	0.00
Sadness	0.00	0.00	0.00	0.04	0.96	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 19 Confusion matrix of the proposed multimodal fusion framework with 91.62% accuracy on the Enterface05 dataset



	Anger	Happiness	Sadness	Fear	Disgust	Surprise
Anger	0.71	0.07	0.14	0.00	0.07	0.00
Happiness	0.12	0.62	0.23	0.00	0.04	0.00
Sadness	0.10	0.06	0.74	0.06	0.03	0.00
Fear	0.00	0.00	0.38	0.25	0.25	0.13
Disgust	0.00	0.00	0.00	0.00	0.67	0.33
Surprise	0.00	0.00	0.00	0.00	0.30	0.70

Fig. 20 Confusion matrix of the proposed multimodal fusion framework with 60.77 % accuracy on the BAUM-1s dataset

the proposed multimodal fusion emotion recognition model takes only about 15 minutes while the testing time costs approximately 0.02s on the two datasets. Therefore, the real-time performance of the proposed model is satisfied with the emotion recognition system.

The confusion matrices of our proposed multimodal emotion recognition method on the Enterface05 and BAUM-1s datasets are shown in Figs. 19 and 20. It is observed that the performance of each emotion achieves more than 80% by fusing the facial expression and acoustic information on the Enterface05 dataset, which suggests that the recognition rate of each modality is increased by using the weighted strategy in decision-level fusion. Clearly, the holistic recognition accuracy on the Enterface05 dataset is higher than that of the BAUM-1s dataset. The reason is that the emotions expressed in the spontaneous emotion dataset are not deliberately exaggerated compared with the acted emotion dataset. Moreover, comparing Figs. 13, 15, and 20, it can be observed that the recognition accuracy of "fear" has been increased by using a multimodal fusion method instead of unimodal recognition on the BAUM-1s dataset, which agrees with our previous analysis that the redundant information from two modalities is valuable to improve the performance of emotion recognition.

5 Conclusions

A multimodal fusion emotion recognition method is proposed in this study. Visual and audio modality-based individual emotion recognition are investigated parallel. The decision-level fusion method is utilized to integrate the recognition results from two modalities. Specifically, in the visual modality, geometric deformation features are firstly computed from keyframes extracted through several facial components. Then, GA is utilized to select discriminative facial features and optimize model structure for performance improvement. Similar analytical procedure is performed in audio modality. Moreover, the ELM classifier is employed to identify the emotions for visual and audio modalities individually, and following by a decision-level fusion. The proposed framework is evaluated by applying the CK+, Enterface05, and BAUM-1s datasets. The results obtained from the three datasets show that

- The proposed keyframe extraction method and geometric deformation features are effective for facial emotion recognition.
- The optimized emotional features and model structure can not only significantly improve the accuracy of emotion recognition in visual and audio modalities, but also decrease the model complexity.
- The performance of the multimodal fusion method outperforms both emotion recognition methods that are individually used in visual and audio modalities.

Although the proposed method has made promising progress on performance improvement of emotion recognition in video clips, there still exist some topics that need to be discussed. For example, how to explore the high-level emotion features from different modalities; how to effectively fuse the features from different modalities, etc. In future work, deep learning technology will be investigated to extract high-level emotional features. In addition, developing strong fusion approaches to enhance the performance of multimodal fusion emotion recognition will also be pursued.

Acknowledgements This work was supported by the Open Foundation of Beijing Engineering Research Center of Smart Mechanical Innovation Design Service under Grant No. KF2019302, the General Projects of Science and Technology Plan of Beijing Municipal Commission of Education under Grant No. KM202011417005, and the National Talents Foundation under Grant No. WQ20141100198.

References

Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76

Avots E, Sapiński T, Bachmann M, Kamińska D (2019) Audiovisual emotion recognition in wild. Mach Vis Appl 30(5):975–985

Bejani M, Gharavian D, Charkari NM (2014) Audiovisual emotion recognition using anova feature selection method and multi-classifier neural networks. Neural Comput Appl 24(2):399–412

Busso C, Deng Z, Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on multimodal interfaces, pp 205–211. https://doi.org/10.1145/1027933.1027968

Chen J, Chen Z, Chi Z, Fu H (2016) Facial expression recognition in video with multiple feature fusion. IEEE Trans Affect Comput 9(1):38–50

Chen L, Zhou M, Su W, Wu M, She J, Hirota K (2018a) Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Inf Sci 428:49–61

Chen L, Zhou M, Wu M, She J, Liu Z, Dong F, Hirota K (2018b) Three-layer weighted fuzzy support vector regression for emotional intention understanding in human–robot interaction. IEEE Trans Fuzzy Syst 26(5):2524–2538

Chu WS (2017) Automatic analysis of facial actions: learning from transductive, supervised and unsupervised frameworks. PhD thesis, Carnegie Mellon University

Ekman P, Friesen WV (1978) Facial action coding system: investigators guide. Consulting Psychologists Press

El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44(3):572–587

Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth Annual Conference of the international speech communication association, pp 223–227

Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio-visual emotional big data. Inf Fusion 49:69–78

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3): 489–501

Jain DK, Shamsolmoali P, Sehdev P (2019) Extended deep neural network for facial emotion recognition. Pattern Recognit Lett 120:69–74

Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on computer vision, pp 2983–2991. https://doi.org/10.1109/ICCV.2015.341

Kansizoglou I, Bampis L, Gasteratos A (2019) An active learning paradigm for online audio-visual emotion recognition. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2019.2961089

Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 1867–1874. https://doi.org/10.1109/CVPR.2014.241

Krithika L, Priya GL (2020) Graph based feature extraction and hybrid classification approach for facial expression recognition. J Ambient Intell Human Comput 12:2131–2147. https://doi.org/10.1007/s12652-020-02311-5

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Liu Y, Yuan X, Gong X, Xie Z, Fang F, Luo Z (2018) Conditional convolution neural network enhanced random forest for facial expression recognition. Pattern Recognit 84:251–261

Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on computer vision and pattern recognition-workshops, IEEE, pp 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio-visual emotion fusion (avef): a deep efficient weighted approach. Inf Fusion 46:184–192

Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06), IEEE. https://doi.org/10.1109/ICDEW.2006.145

Mendoza-Palechor F, Menezes ML, Sant'Anna A, Ortiz-Barrios M, Samara A, Galway L (2019) Affective recognition from eeg signals: an integrated data-mining approach. J Ambient Intell Hum Comput 10(10):3955–3974

Miyoshi R, Nagata N, Hashimoto M (2021) Enhanced convolutional lstm with spatial and temporal skip connections and temporal gates for facial expression recognition from video. Neural Comput Appl 33:7381–7392. https://doi.org/10.1007/s00521-020-05557-4

Noroozi F, Marjanovic M, Njegus A, Escalera S, Anbarjafari G (2017) Audio-visual emotion recognition in video clips. IEEE Trans Affect Comput 10(1):60–75

Pons G, Masip D (2020) Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition. IEEE Trans Cybern 99:1–8. https://doi.org/10.1109/TCYB.2020.3036935

Poria S, Cambria E, Howard N, Huang GB, Hussain A (2016) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59

Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion .Inf Fusion 37:98–125

Rahdari F, Rashedi E, Eftekhari M (2019) A multimodal emotion recognition system using facial landmark analysis. Iran J Sci Technol Trans Electric Eng 43(1):171–189

Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan SS (2010) The interspeech 2010 paralinguistic challenge. In: Eleventh annual conference of the international speech communication association, pp 2794–2797

Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816

Wang Y, Guan L (2008) Recognizing human emotional state from audiovisual signals. IEEE Trans Multimed 10(5):936–946

Whitley D (1994) A genetic algorithm tutorial. Stat Comput 4(2):65–85

Wöllmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, Morency LP (2013) Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell Syst 28(3):46–53

Wu M, Su W, Chen L, Liu Z, Cao W, Hirota K (2019) Weight-adapted convolution neural network for facial expression recognition in human-robot interaction. IEEE Trans Syst Man Cybern Syst 51(3):1473–1484

Xiao W, Zhang J, Li Y, Zhang S, Yang W (2017) Class-specific cost regulation extreme learning machine for imbalanced classification. Neurocomputing 261:70–82

Xie S, Hu H, Wu Y (2019) Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognit 92:177–191

Zeng Z, Pantic M, Roisman GI, Huang TS (2008) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58

Zhalehpour S, Onder O, Akhtar Z, Erdem CE (2016) Baum-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans Affect Comput 8(3):300–313

Zhang S, Zhang S, Huang T, Gao W, Tian Q (2017) Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans Circ Syst Video Technol 28(10): 3030–3043

Zhang S, Zhao X, Tian Q (2019) Spontaneous speech emotion recognition using multiscale deep convolutional lstm. IEEE Trans Affect Comput 99:1–1. https://doi.org/10.1109/TAFFC.2019.2947464

Zhang J, Li Y, Xiao W, Zhang Z (2020a) Non-iterative and fast deep learning: multilayer extreme learning machines. J Frankl Inst 357(13):8925–8955

Zhang J, Li Y, Xiao W, Zhang Z (2020b) Robust extreme learning machine for modeling with unknown noise. J Frankl Inst 357(14):9885–9908

Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.