Speech Communication 127 (2021) 92–103

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition

Zeinab Farhoudi [a,*], Saeed Setayeshi [b]

[a] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[b] Department of Energy Engineering and Physics, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Audio-Visual emotion recognition
Brain emotional learning
Deep learning
Convolutional neural networks
Mixture of network
Multimodal fusion

ABSTRACT

Multimodal emotion recognition is a challenging task due to different modalities emotions expressed during a specific time in video clips. Considering the existed spatial-temporal correlation in the video, we propose an audio-visual fusion model of deep learning features with a Mixture of Brain Emotional Learning (MoBEL) model inspired by the brain limbic system. The proposed model is composed of two stages. First, deep learning methods, especially Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), are applied to represent highly abstract features. Second, the fusion model, namely MoBEL, is designed to learn the previously joined audio-visual features simultaneously. For the visual modality representation, the 3D-CNN model has been used to learn the spatial-temporal features of visual expression. On the other hand, for the auditory modality, the Mel-spectrograms of speech signals have been fed into CNN-RNN for the spatial-temporal feature extraction. The high-level feature fusion approach with the MoBEL network is presented to make use of a correlation between the visual and auditory modalities for improving the performance of emotion recognition. The experimental results on the eNterface'05 database have been demonstrated that the performance of the proposed method is better than the hand-crafted features and the other state-of-the-art information fusion models in video emotion recognition.

1. Introduction

Multimodal emotion recognition is one of the main challenging tasks due to the multimodality characteristic of human emotional expression. Indeed, human emotional expression is along spatial-temporal dimensions thus, it is hard to recognize emotion from fusion of multimodal cues. Since the last decade, multimodal emotion recognition especially, speech and facial expression has been attracting attention in several applications such as e-learning (Shen et al., 2020), human-computer interaction (Beale and Peter, 2008), health monitoring (Torous et al., 2014), mobile computing (Lv et al., 2015) and gaming (Szwoch and Szwoch, 2015). However, audio-visual emotion recognition is still an open challenge in machine learning and computer vision due to two reasons. First, it is hard to extract the best audio and visual features due to the variations of emotional expression for each person. Second, the essential problem is to find how to integrate the audio and visual modalities by considering the spatial-temporal correlation exists between them.

Recent studies are based on a combination of multiple modalities such as facial expressions and vocal expressions. Feature extraction and multimodal fusion of different modalities as an effective manner are still open problems.

The representation of facial expression has been attracted a great deal of attention in the past decade. Recently, the rapidly growing of CNN models as a feature extraction of face images or video frames has been considered and some face pre-trained models have been presented (Schroff et al., 2015). Since these models have a lack of temporal information, they are not suitable for video analysis. To deal with this issue, a variety of the 3D-CNN models have recently been proposed (Tran et al., 2015).

Adding audio effective information plays an important role in emotion recognition. Most of the researches employ low-level hand-crafted audio emotional features like Mel-frequency Cepstral Coefficient (MFCC) or spectrograms with either traditional approaches (Paleari et al., Jul. 2010) or deep learning approaches (Zhang et al., Oct. 2018). However, these audio features are not appropriate for video emotion recognition. In the proposed architecture a CNN is firstly applied to the two-dimensional Mel-spectrogram representation of the short-term

* Corresponding author.
E-mail addresses: zeinab.farhoudi@srbiau.ac.ir (Z. Farhoudi), setayesh@aut.ac.ir (S. Setayeshi).

https://doi.org/10.1016/j.specom.2020.12.001
Received 5 April 2019; Received in revised form 5 June 2020; Accepted 1 December 2020
Available online 3 December 2020
0167-6393/

segments of the audio signals. Then, the RNN method is applied to model the temporal information that is closely associated with target emotion.

Since emotion recognition from only audio or facial expression has some ambiguity, it is expected that emotion recognition based on the audio-visual fusion could perform a better accuracy. Despite the benefits of multimodal fusion, they still encounter the following challenges. First, signals are not synchronous at the same time. Second, it is difficult to construct a model that uses complementary relationship information rather than supplemental information. Third, each modality shows different types and levels of noise at different points in time. To reduce the above-mentioned problems, recent multimodal fusion researches have used artificial neural network (ANN) approaches to fuse multimodal information (Kim et al., May 2013; Zhang et al., Jun. 2016). The main advantages of ANN or Deep NN (DNN) methods are generalization capability at large scale and end-to-end learning strategies. Moreover, in the multimodal fusion task, these methods can jointly modeling non-linear correlations of multiple cues with different properties (Zhang et al., Oct. 2018). Hence, we have proposed a combined neural network model-based fusion method to find the correspondence between the audio and visual streams and classify video emotion.

Mixture of expert neural network (MoE) is basically inspired by the associative cortex of the brain, which can handle information integration from many sources (Stein et al., 2009). The MoE consists of a number of experts (learners) used to divide the problem space into homogeneous regions and a gating network decides which expert to use for each input region (Jacobs et al., 1991). In addition, in the human brain the limbic system is responsible to reply the emotional cues and a powerful computational bio-inspired model, namely the BEL model, has been firstly presented by Morén and Balkenius (Christian Balkenius, Sep. 2001; Balkenius and Morén, 1998). Therefore, it seems combining two computational bio-inspired models including the BEL model and the MoE is a good idea for multimodal fusion in emotion recognition. The proposed model, namely the MoBEL network is like a mixture of experts in which each expert and gating network is consists of the BEL model.

The BEL model consists of the amygdala and the Orbitofrontal Cortex (OFC) components inspired by the limbic system. The original model of the BEL is controlled by reinforcement signal i.e., reward and reinforcement learning (Babaie et al., Jul. 2008). The improved BEL model has been successfully adopted in many applications such as on the intelligent controller the Brain Emotional Learning Based Intelligent Controller (BELBIC) model (Lucas et al., Jan. 2004), (Lucas, 2011), pattern recognition (Asad et al., 2017), speech emotion recognition (Farhoudi et al., 2017). Lotfi et al. (Lotfi et al., 2018) proposed the Competitive BEL model as an emotional brain-inspired learning algorithm that is like an ensemble of neural networks. Although, there exists a nonlinear BEL model for solving complex problems (Jafari and Xu, 2019; Fang et al., 2019; Zhao et al., 2019), We propose an extended BEL model with fewer parameters and more accurate than other methods (Zhao et al., 2019) in the field of emotion recognition.

Learning audio-visual features for emotion recognition is one of the critical steps in finding the correspondence between multimodal cues. Previous works focus on using hand-crafted features, which have been verified not discriminative enough to human emotions. In contrast, this work aims at automatically learning a joint audio-visual feature representation from raw audio and visual signals using the MoBEL model as a fusion network. This work is organized as follows. In Section 2, we describe the proposed model i.e., multimodal fusion by using MoBEL from deep visual and audio representation and training MoBEL in detail. We explain the experimental results in Section 3. Also, this section contains a description of the audio-visual emotional database such as eNterface'05. Finally, the conclusions and future works are given in sections 4.

2. Related work

An audio-visual emotion recognition consists of two important steps: feature representation and multimodal fusion. In the following part, we review related works of these two steps.

2.1. Feature representation

Feature representation and multimodality fusion are two important steps for audio-visual emotion recognition. For facial expression recognition in static images, there have been several works (Dhall et al., 2011; Wang and Guan, 2008; Wang et al., 2012; Zhao and Pietikainen, 2007) focusing on extracting hand-crafted low-level features can be summarized into two methods, namely appearance-based and geometry-based. Appearance-based representation methods adopt the whole or specific region of a face image to explain the changes such as wrinkles and furrows, methods such as Gabor wavelet (Wang and Guan, 2008), (Wang et al., 2012), Local Binary Patterns (LBP) (Zhao and Pietikainen, 2007), Local Phase Quantization (LPQ) (Dhall et al., 2011). Geometry-based feature extraction methods represent facial motion parameters including eyes, eyebrows and mouth, such as Active Appearance Model (AAM) method (Lucey et al., 2007), (Chang et al., 2006). In recent years, CNN as visual features are used to extract facial features of static and dynamic images in a video (Trigeorgis et al., 2016). For dynamic image sequences, popular visual features represent facial muscle movement during the time (Mansoorizadeh and Moghaddam Charkari, 2010). For example, (Tran et al., 2015), (Fan et al., 2016) have used 3D-CNN or CNN-RNN models to extract visual features of facial expression in a video.

Generally, audio affective hand-crafted feature extraction methods are summarized into two categories including prosodic and spectral features. Prosodic features including Pitch, energy, intensity and zero-crossing while spectral features including formants, spectral energy distribution, Mel-frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), etc. Most of the multimodal approaches employ prosodic or hybrid features with traditional classifiers such as Hidden Markov Model (HMM) or Support Vector Machines (SVM). Zeng et al. (Zeng et al., 2008) had used prosody features with multi-stream HMM (MS-HMM) for speech emotion recognition in video. Ntalampiras et al. had extracted Pitch, wavelet domain and MFCC features then, employed fusion method, namely fusion HMM (F-HMM) for speech emotion recognition task (Ntalampiras and Fakotakis, 2012). Nowadays, deep learning methods are employed and achieved good results in this task. Zhang et al. (2018) first, split a video clip into a number of overlapping segments and then convert a one-dimensional audio signal to log Mel-spectrogram and second applied CNN on three channels of Mel-spectrogram.

2.2. Multimodal fusion

After feature extraction, multimodal fusion is used to integrate audio and visual multimodal cues for emotion recognition tasks in video. Previous works focus on four typical fusion strategies to handle these challenges, including feature-level fusion (Mansoorizadeh and Moghaddam Charkari, 2010), (Schuller et al., 2007), decision-level fusion (Wang et al., 2012), (Sahoo and Routray, 2016), model-level fusion (Gurban et al., Oct. 2008), (Chen and Jin, 2015) and hybrid fusion (Lan et al., Jul. 2014). Nowadays, in model-based fusion, some fusion networks based on fully connected layers have been suggested to improve video classification by capturing the mutual correlation among different modalities (Zhang et al., 2016). For example, in (Pini et al., 2017), a fusion network is trained to jointly extract static and dynamic features from different modalities. Although these methods demonstrate good performance on audio-visual emotion recognition tasks, they cannot take full advantage of the complimentary non-linear correlation between visual and auditory modality at high-level features. Therefore, it is needed to design a mixture of neural network fusion methods to improve the performance of the emotion recognition task.

Motivated from the algorithm in this work, we discuss a new

approach, namely MoBEL as a fusion network and classifier. This novel method utilizes high-level learning features and integrate them in a fusion network to find the correspondence between the audio and visual streams. In this work, we propose MoBEL as a fusion network. The MoBEL model is a modified and neuro-inspired version of the Mixture of Experts neural network (MoE). Since in the nervous brain system there are two types of inhibitory and excitatory synapses (Hall, 2015), in the MoBEL model, each expert and gating network consists of the BEL model, and all parts are trained jointly by back-propagation.

3. Proposed model

Adopting the power of deep learning models in feature extraction, we propose the audio-visual fusion model of deep learning features with the MoBEL model. The proposed model composed of two stages: First, deep learning methods, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are being applied to represent features at high-level. Second, the fusion model, MoBEL, is trained to jointly learn audio-visual learning features. For the visual modality representation, 3D-CNN models are used to learn the spatial-temporal features of visual expression (Zhang et al., 2018), while for auditory modality the Mel-spectrograms of the speech signals are fed into Convolutional Recurrent Neural Network (CRNN) for spatial-temporal audio feature extraction. Fig. 1 demonstrates the architecture of our proposed model. As shown in Fig. 1, this architecture consists of these steps:

1) Split video to two streams: visual and audio and then segment each visual and audio stream separately with a specific frame rate.
2) For multiple sequences of visual segment, we first preprocess each frame and then applied the 3D-CNN model to learn facial expression representation.
3) In the audio stream first, we convert the raw speech signal into a Mel-spectrogram image and then applied CRNN to learn speech emotion recognition.
4) We get the fully connected layer of each stream and concatenate them to create a fixed-length video features then the MoBEL network is trained to learn the correspondence between the audio and visual streams and finally to recognize the video emotion.

To deal with the multimodal emotion recognition task in video clips based on audio-visual information fusion, we build a mixture of neural network, MoBEL, as a fusion network that can integrate the spatial-temporal audio-visual information and learn their correlation to recognize the emotion of the video clip. Motivated by the success of MoE and the BEL neural network, we propose the MoBEL model that each expert and gating network consists of the BEL model. The reason to choose the BEL model is its capability in decision-making as its cognitive-based structure and the reason to choose the MoE model is basically inspired by the associative cortex of the brain, which can handle information integration from many sources. Based on (Stein et al., 2009), it is evident that the presence of the associative cortex is needed to improve the perception of the environment by the brain. Thus, The MoE would be able to extract much wealthier relationships between input streams and would be better to provide pattern recognition accurately. Using a gating scheme in the MoBEL model enables a network to better trained to understand under what conditions the weights of each modality should be increased.

Fig. 1 demonstrates a black box of the BEL model which we will explain in Section 3.3 in detail. The BEL model consists of four main subsystems: amygdala, OFC, Thalamus and sensory cortex. The high-level learning features extracted from various unimodal are integrated into the cortex and based on the interaction with the memory, the amygdala and OFC make a decision. Amygdala receives emotional stimuli from the sensory cortex and Thalamus as well as the external reward signal, and it interacts with the OFC. The OFC receives sensory input from the sensory cortex and evaluates the amygdala's response to prevent inappropriate learning connections. The system also receives a reinforcing signal (reward) which has been left unspecified at the original BEL model (Christian Balkenius, 2001). In the proposed model we have used a supervised version of the extended BEL model, where the reward value is the target value in each expert BEL network and for gating BEL network, the reward value is a posterior probability that each expert can generate the desired output.

Thus, according to these studies, we present a new bio-inspired model for multimodal emotion recognition, the MoBEL model, in which learn a joint nonlinear audio-visual representation for emotion recognition task. Indeed, as shown in the experimental results, the MoBEL model can improve the accuracy of the system. In other words, the advantages of the MOBEL methods are able to jointly learn feature representations and appropriate classifiers. Indeed, the main contribution of the MoBEL is to decide the weights of different audio-visual modalities for the best description.

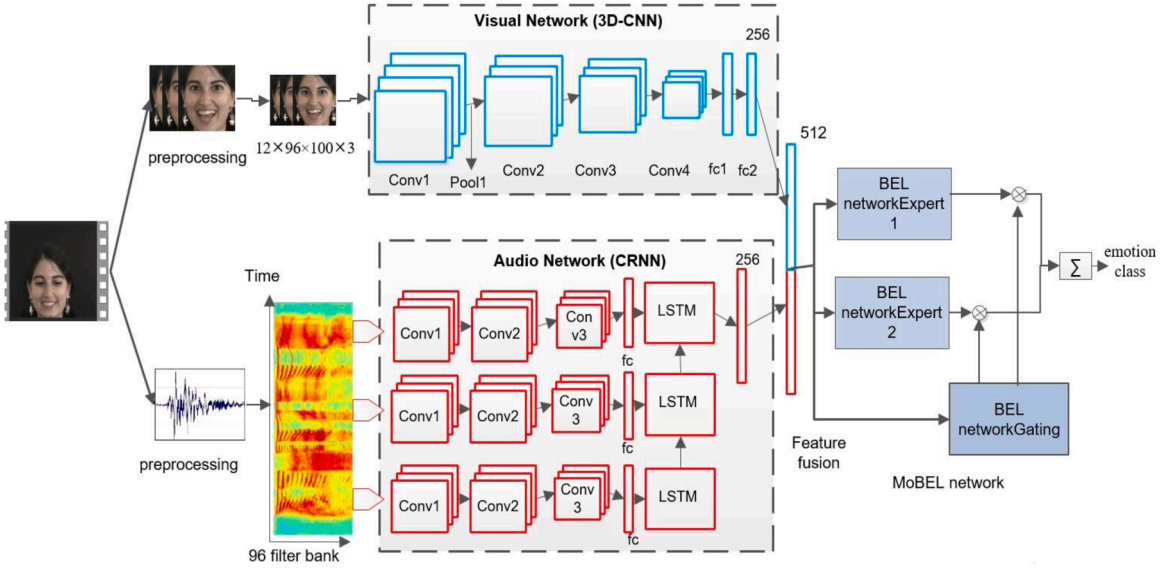

Fig. 1. The architecture of the proposed model of audio-visual emotion recognition task by using deep learning features such as 3D-CNN and CRNN network for visual and audio representation, respectively and then fusion audio-visual high-level features by MoBEL model.

In the following, we explain each step of the model in detail. In visual and audio representation, we describe how to prepare input and then training the 3D-CNN and CRNN models for the specific inputs, respectively, and how to train the MoBEL model as a multimodal fusion network.

3.1. Deep visual representation

Face emotion recognition system in video comprises of three steps: 1) extract key frames of the video 2) preprocess each frame, including face detection and alignment 3) apply 3D-CNN network model for a sequence of frames. In face detection, we use the Single Shot Detection Multibox detector (Mobilenet SSD) method based face detector with a pre-trained model provided (Liu et al., 2016). This method is faster (60 fps on Nvidia GTX1080 GPU) and more accurate and requires less memory than the other methods. The output of the SSD method is a coordinate of face bounding boxes. Using these bounding boxes, we crop a RGB face image and resize it to $96 \times 100 \times 3$. In our experiment, we have tested many face sizes and sequences of frames. Thus, the best parameters that suitable for our network and require less memory and more recognition accuracy were 12 sequences of key frames with $96 \times 100 \times 3$ face size of each frame. If a video sample has F frames more than 12 frames ($F > 12$), then we delete the first and the last $(F-12)/2$ overlapping frames, and for video samples with $F < 12$ we repeat the first and the last $(F-12)/2$ overlapping frames. We note that all video files of the eNterface'05 dataset used in this work have 25 fps and the duration of each video is more than 1 second. So, in this case, we do not need to deal with $F < 12$ frames.

Finally, after creating the input vector of size ($12 \times 96 \times 100 \times 3$) we fed the input to the 3D-CNN network for training. The 3D-CNN network can learn spatial-temporal features from the video directly. Indeed, the 3D-CNN is used for feature extraction and the last fully connected layer of the network is saved for the next operation.

In (Zhang et al., 2018), the authors had proposed a 3D-CNN network, namely the C3D-Sport-1M model that had been pre-trained on large video sport classification tasks and then fine-tuned the model with the labeled emotion data. Since, we have memory limitations of 8 GB GPU, we simplify the model as we hacked off some layers. Our customized 3D-CNN model consists of 3 convolutional layers (with 64 kernel size of $3 \times 3 \times 3$), 3 max-pooling layers followed by 3 fully connected layers. In the end, a softmax layer with 7 nodes, according to the number of our classes, is utilized for facial emotion recognition.

3.2. Deep audio representation

In the audio stream, we first convert the raw signal into log Mel-spectrogram images as a time-frequency map and then applied the CRNN model to extract speech emotional features. Since the CNN methods are commonly used for image and video processing, robust and discriminative features are learned by converting the 1-D audio signal into the 2-D spectrogram and using CNN to extract features of spectrogram images (Lim et al., 2016). The extracted log Mel-spectrogram is computed with the output of Mel-frequency filter banks that demonstrates more discriminant power than MFCC for speech emotion recognition (Busso et al., 2007). First, we convert an audio signal into 20ms frames with 10ms overlapping. Then, the log-Mel spectrogram of each overlapping hamming window is calculated by the Eq. (1):

$$f_{mel}(f) = 2595\log_{10}\left(1 + \frac{f_{HZ}}{700}\right) \tag{1}$$

typically, for a given audio we adopt 96 Mel-filter banks from 20 to 8000 Hz and a context window of 32 frames is used. Therefore, the size of the input image to fed into CNN is $96 \times 32 \times 1$ which represented as a gray image.

Once the input audio signal is represented as log-Mel spectrogram segments, the same property as a sequence of images become share, from which the CNN can be applied to extract spatial features. Then, the Long Short-term Memory (LSTM) model is used to learn temporal dependencies between different time-step local invariant features. Finally, the last embedding vector of speech emotion representation is obtained in the classification method. The detailed architecture of the speech emotion recognition stream with the CRNN model is demonstrated in Fig. 2.

The CNN network architecture, including the number of layers, number of filters in the convolutional layer and max-pooling filter sizes is the same as visual stream emotion recognition. In the LSTM part, we have a layer with 256 cells. The last FC layer with a size of 256 is extracted as a speech emotion representation.

Afterward, audio and visual learning emotional features of each stream are combined and fed to the MoBEL network model to train jointly spatial-temporal information of different modalities.

3.3. Training the MoBEL network

As mentioned, we used the structure of the BEL model in the experts and the gating networks. In Fig. 3, the structure of the proposed Mixture of BEL networks model is shown in the mid-level feature fusion for bimodal emotion recognition.

According to Fig. 3, there are two neural experts and a gating network. Assume that $p = \{p_1, p_2, ..., p_n\}$ is a vector of the input pattern, n is 512 and $E1$ and $E2$ are the output vectors of the expert 1 and expert 2, $g1$ and $g2$ are the weight assigned to the *BEL net Expert1* and the *BEL net Expert2* by the gating network, respectively. The g_i Estimates of the probability that the i-th network can correctly classify the p-model. The number of neurons in the output layer of the gating network is equal to the number of experts. The final output of the proposed model is y and is calculated as Eq. (2):

$$y = \sum_i E_i\, g_i \;.\; i = 1. \;....\; N \tag{2}$$

And the g_i output of the gating network as Eq. (3) is a function of the input patterns and learning weights.

$$g_i = \frac{\exp(O_{gi})}{\sum_{j=1}^{N}\exp(O_{gi})} \tag{3}$$

Where N is the number of expert systems. In our experiments, N is equal to 2. By defining the above equation for g_i, the sum of the weights assigned to each of the base classifications is equal to one. If the difference of the output of each of the expert classifiers with the desired output is lower, then the error is less and, therefore, it is assigned a greater g_i weight. In the gating network, a BEL model is consisting of two

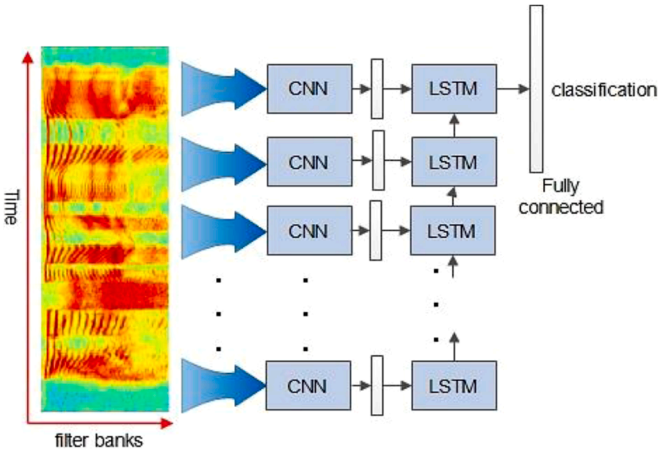

Fig. 2. The architecture of CRNN for speech emotion recognition.

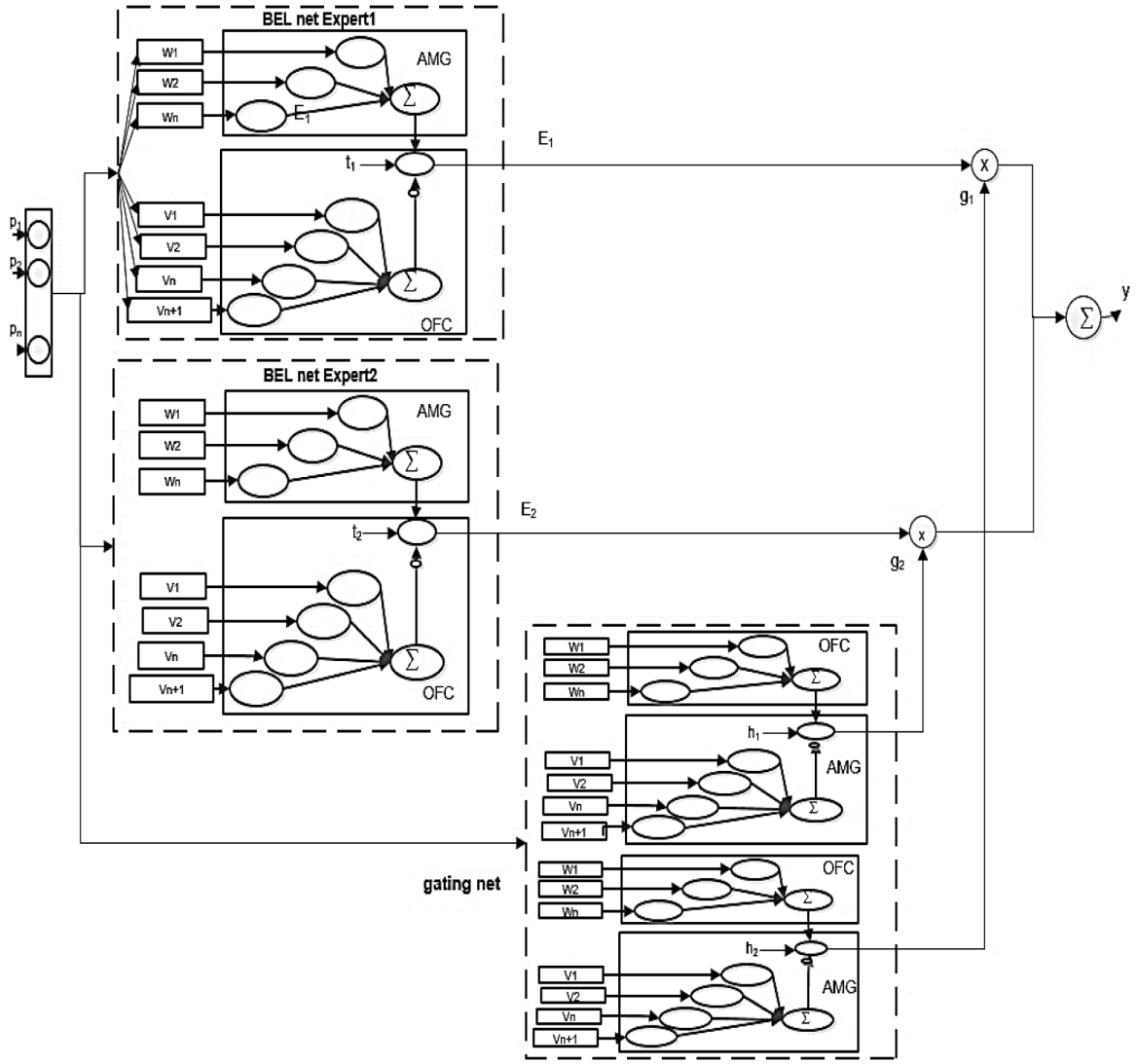

Fig. 3. The proposed MoBEL network. This mixture of experts of the BEL network consists of two experts and one gating network. Each BEL model consists of AMG and OFC components.

AMG and two OFC for two outputs and, therefore two output weights.

In this model, the value of the optimal output h_i is defined by Eq. (4) and can be considered softmax operations. h_i is a posterior probability that each expert can generate the desired output.

$$h_i = \frac{g_i \exp\left(-\frac{1}{2}(t-E_i)^T(t-E_i)\right)}{\sum_j g_i \exp(-\frac{1}{2}(t-E_i)^T(t-E_i)} \tag{4}$$

Where t is the desired output value, and E_i is the output of each of the base (expert) classifiers. In a mixture of the expert model with the definition of the loss function and employ the Gradient Descent (GD), a method is provided for the supervised training in each of the expert classifiers. Eq. (5) is the error function of the total system.

$$e = \sum_i g_i \| y - E_i \|^2 \tag{5}$$

Also, the outputs of each of the expert classifiers are calculated as Eqs. (8) and (9). Each expert consists of two component the Amygdala and OFC, so the output of each components e_a and e_o is as follows, respectively. (Eqs. (6) and (7))

$$e_a = \sum_{j=1}^{n} v_j\, p_j + v_{n+1} p_{th} \tag{6}$$

$$e_o = \sum_{j=1}^{n} w_j\, p_j \tag{7}$$

Where f is the activator function, v_j is the AMG weights, and w_j is the OFC weights, p_j is the input pattern and p_{th} is the maximum input value that acts as a bias.

$$p_{th} = \max(p_j)\ .\ j = 1.\ \ldots.\ n \tag{8}$$

Therefore, the summation output function of each expert is as defined as follow:

$$E_i = f_e\left(f_{amg}(e_a) - f_{OFC}(e_o)\right) \tag{9}$$

Where f_{amg} and f_{OFC} are the activation functions of the AMG and OFC, respectively. Based on the above equations and error-derived equations for each of the AMG and OFC weights, updating weights in the modified BEL model is calculated as Eqs. (10) and (11).

$$\Delta v_{ij} = -\gamma\, v_{ij} + \alpha_e\, p_j\, max\left(0,\ t - f_{amg}(e_a)\right) \tag{10}$$

$$\Delta w_{ij} = \beta_e \, p_j \, (E_i - t) \; for \; i = 1, 2 \; for \; j = 1, \ldots, n \tag{11}$$

Where α_e is the learning rate of the AMG, β_e is the learning rate of OFC and γ is the degradation rate for the momentum for each expert and i indicate the number of experts and j indicate the number of nodes. Also in the gating network, the summation output function g_i of the BEL model, updating weights of the AMG and OFC are calculated as Eqs. (12), (13) and (14), respectively.

$$g_i = f_g \left(f_{amg}(e_a) - f_{OFC}(e_o) \right) \tag{12}$$

$$\Delta v_{gj} = -\gamma \, v_{gj} + \alpha_g \, p_j \, max(0. \, h_i - f(e_a)) \tag{13}$$

$$\Delta w_{gj} = \beta_g \, p_j \, (g_i - h_i) \tag{14}$$

that α_g, β_g are the learning rate of the AMG and OFC in the gating network, respectively.

The convergence of the weights of the AMG (V_i, V_g) and the OFC (W_i, W_g) has been proved in Theorem 1 for the BEL model by Jafari et al. (Jafari and Xu, 2019; Jafari et al., 2019). According to theorem 1, by satisfying the tuning parameters α, β for each expert and gating network under these conditions:

$$\text{I} \quad |[1 - \alpha(p_j)^2]| \prec 1$$
$$\text{II} \quad |[1 - \beta(p_j)^2]| \prec 1$$

the estimated weights of the Amygdala and OFC converge to desired targets asymptotically. Also, as shown in Eqs. (10) and (12), the Amygdala weights v_{ij} and v_{gj} cannot be decreased, because the Amygdala cannot learn. So, it is the task of the OFC to inhibit this reaction when it is inappropriate. The OFC connection weight can both increase and decrease.

In this way, in the training process, expert classifiers compete with each other for each input pattern, and the gating network selects a winning expert based on the error of each expert classifier. After training the MoBEL fusion network, the output result indicates the video emotion class. In training the MoBEL model, the learning rate of the AMG and OFC α_g, β_g and α_e, β_e in the gating network and expert networks respectively are updated at the end of each iteration adaptively. According to this step, when the training error is increased, the learning rate is decreased. The pseudo-code of the algorithm is as follow:

// update learning rate α_e
if (current_performance/ previous_performance) > 1.04:
$\alpha = \alpha \times 0.7$
else:
$\alpha = \alpha \times 1.05$

The computational complexity of the original BEL model is O(n). According to Eqs. (8) and (9). the number of learning weights of the BEL model is (n+1) for each the Amygdala and OFC. However, in the MoBEL model, the computational complexity is O(3n+1). Thus, we have (n+1) connecting weights of each Amygdala and OFC for each expert, and (n) weights of gating network output. In the MoBEL model, each expert can be done in parallel. Therefore, increasing the number of expert networks does not affect the computational complexity of the model.

4. Experiments

To testify the effectiveness of our proposed neuro-inspired fusion networks for audio-visual emotion recognition, we employ our proposed video emotion recognition model on audio-visual emotional eNterface'05 dataset (Martin et al., 2006). To evaluate the performance, we first present the result of each modality, i.e., audio and visual emotion recognition separately, then we give the multimodal fusion and demonstrate the results of various integrating audio and visual stream for multimodal emotion recognition tasks.

The implementation details of the audio and visual deep learning models are as follows:

1 Mini-batch size equal 32
2 Adam optimizer with learning rate 0.0001 and a momentum of 0.00001.
3 The number of epochs is set to 500 for 3D-CNN, 400 for CRNN and 200 for MoBEL, respectively.
4 Implement under Tensorflow framework

The experiments of the proposed model perform on NVIDIA GTX TITAN X GPU with 8GB memory for training these deep models. For training the MoBEL model, we use two BEL models as an expert network and one BEL model as a gating network. The learning rate for each AMG and OFC is set to 0.009. The stochastic gradient descent with the stochastic momentum for each AMG and OFC is set to 0.0001. The activation function of each layer in AMG (f_{amg}) and OFC (f_{OFC}) is the '*tansig*' function. The *Reward* value of each expert BEL network is set to the target value in a supervised BEL learning network, as previously described in Section 3.3.

As suggested in (Schuller et al., 2010), To increase the training data while keeping an early stopping epochs, the validation set of the eNterface'05 dataset is split into five folders, and the models are trained five times by 5-fold cross-validation strategy. Also, in each fold, we adopt a subject-independent Leave-One-Subject-Out (LOSO) cross-validation technique. Later, the average accuracy after 5-fold runs is reported to evaluate the performance of the proposed models on the eNterface'05 dataset.

4.1. Multimodal datasets

There are many audio-visual emotional benchmark datasets such as acted RML (Wang and Guan, 2008), acted eNterface'05 (Martin et al., 2006), the spontaneous BAUS-1s (Zhalehpour et al., 2017) and AFEW (Dhall et al., 2012). We train and evaluate the proposed model on the eNterface'05 dataset.

The eNterface'05 audio-visual acted dataset includes six emotions, i.e., anger, disgust, fear, joy, sadness and surprise, from 43 subjects with 14 different nationalities speaking in English. It contains 1290 video samples. Each audio sample is recorded with a sampling rate of 48000 Hz with 16-bit resolution and mono channel. The video files are on

Fig. 4. A sequence of six facial expression images of one person in the eNterface'05 dataset.

average 3-4 seconds long. The size of the original video frame is $720 \times 576 \times 3$. In Fig. 4, a sequence of six facial expression images of one person in the eNterface dataset is shown.

4.2. Visual experimental results

In this section, we demonstrate the experimental results of facial expression recognition in video. First, we preprocess each frame following face detection by the SSD face method. Second, crop each RGB face image and resize it to $96 \times 100 \times 3$. At last, a sequence of 12 frames with a size of $12 \times 96 \times 100 \times 3$ is fed to the 3D-CNN network for training and testing. The number of cells in the last FC layer is set to 256 and the size of the softmax layer is set to 6.

Table 1 shows the accuracy rate of facial emotion recognition on the eNterface'05 dataset compare between our deep learning features by 3D-CNN and the corresponding hand-crafted features.

According to Table 1, it can be observed that our deep learning features by 3D-CNN which consider the spatial-temporal correlation of visual expression during time yield better performance than the compared hand-crafted features (Mansoorizadeh and Moghaddam Charkari, 2010; Sahoo and Routray, 2016; Zhalehpour et al., 2017; Bejani et al., 2014) such as Facial points, LBP, LPQ and Quantized Image Matrix (QIM). This indicated that our learning visual features have a powerful ability to extract more discriminative cues than compared manual low-level features.

Moreover, we compare the proposed 3D-CNN model for facial expression recognition with other deep neural networks and state-of-the-art methods. Table 2 shows this comparison on the eNterface'05 dataset.

However, another approach to consider the spatial-temporal correlations of facial expression in video frames is CNN-RNN method (Fan et al., 2016). We have compared the recognition rate of our visual features 3D-CNN with another deep learning feature CNN-RNN. As indicated in Table 2, Figs. 5 and 6, the recognition accuracy of the 3D-CNN model is better than the CNN-RNN model in facial expression recognition. But, the CNN-RNN needs less memory than the 3D-CNN model. In the CNN-RNN method, the 2D-CNN network is a VGG16 network that first is trained on every image of the FER-2013 facial face dataset (Vielzeuf et al., 2017). Then, fine-tuning is conducted for the CNN network with the label emotion data on the eNterface'05 dataset. Finally, the features of the last fully connected layer of the CNN network for 12 sequences of key frames are feed into the LSTM model with 4096 neurons to learn temporal correlation existed in the emotion representation.

In Fig. 5, the confusion matrix of employing the 3D-CNN model for facial expression recognition with subject-independent and cross-validation of average 5 runs in the eNterface'05 dataset is shown. According to Fig. 5, the "disgust", "sadness" and "happiness" have the highest recognition rate, and the "anger" and "surprise" emotion have the lowest one. Moreover, the total recognition rate of 6 emotional class using the 3D-CNN model for facial emotion recognition is 62%.

We report We report the recognition rate of the CNN-RNN model for facial emotion recognition in terms of the confusion matrix in Fig. 6. As shown in Fig. 6, the recognition rate of CNN-RNN is less than the 3D-CNN model.

Table 1
The comparison of the recognition rate of the 3D-CNN model with other hand-crafted features for facial emotion recognition on the eNterface'05 dataset.

Refs.	Visual features	Accuracy (%)
Mansoorizadeh et al., (Mansoorizadeh and Moghaddam Charkari, 2010)	Facial Points	37
Bejani et al., (Bejani et al., 2014)	QIM	39.27
Zhalehpour et al., (Zhalehpour et al., 2017)	LPQ	42.16
Sahoo et al., (Sahoo and Routray, 2016)	LBP	57
proposed model	**3D-CNN**	**62**

Table 2
The comparison of the recognition rate of the 3D-CNN model with other state-of-the-art methods for facial emotion recognition on the eNterface'05 dataset.

Refs.	Visual features	Accuracy (%)
Noroozi et al. (Noroozi et al., Jan. 2019)	AVER-Geometric	49.59
Noroozi et al. (Noroozi et al., Jan. 2019)	**AVER-CNN**	**62**
Wang et al. (Wang et al., Jun. 2012)	KCMFA	58
Zhang et al. (Zhang et al., Oct. 2018)	3D-CNN	54.35
Proposed model (2)	CNN-RNN	57.8
Proposed model (1)	**3D-CNN**	**62**

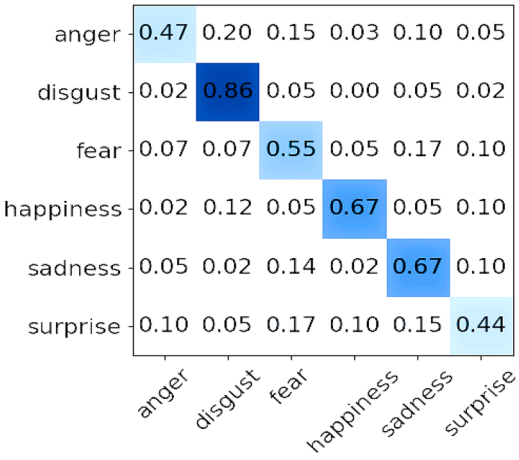

Fig. 5. Confusion matrix of facial expression recognition results by using 3D-CNN visual feature learning.

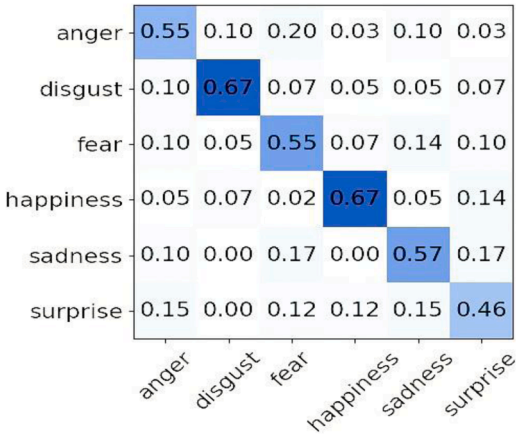

Fig. 6. Confusion matrix of facial expression recognition results by using CNN-RNN visual feature learning.

CNN model.

4.3. Audio experimental results

In this section, we describe the experimental results of the CRNN model for speech emotion recognition. For each speech sample, we create an image of a log Mel-spectrogram of the audio signals with a size of $96 \times 32 \times 1$ and 12 sequences of images. In other words, we adopt 96 Mel-filter banks and a context window of 32 frames which for a 4000ms of an audio sample in the eNterface'05 dataset create 12 sequences. It should be noted that, for videos with a duration of less than 4000ms, we have used zero-padding. Therefore, all video samples resize to $12 \times 96 \times 32 \times 1$, then fed to the CRNN for training, and the parameters set as same as Section 3.2. In Fig. 7, the confusion matrix of the CRNN

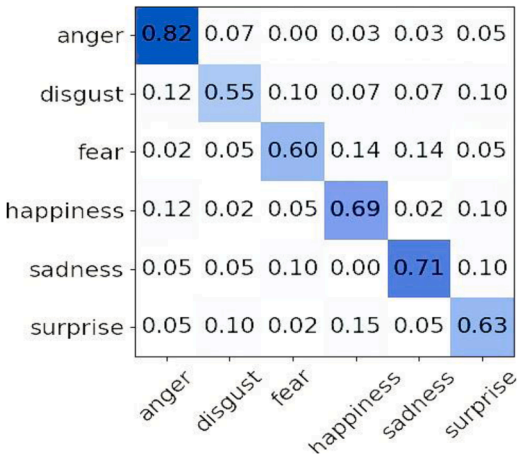

Fig. 7. Confusion matrix of speech emotion recognition by using CRNN.

model for speech emotion recognition for the 6 basic emotional classes of the eNterface'05 dataset is illustrated.

According to Fig. 7, the "anger" has the highest recognition rate and the "disgust" emotion has the lowest one. The experimental results of speech emotion recognition are almost in contrast with the facial emotion recognition ones.

The experimental results of speech emotion recognition are almost the opposite of the facial emotion recognition ones. Moreover, the total recognition rate for 6 emotional classes using the CRNN model for speech emotion recognition is 67.7%.

To present the advantages of the deep learning features, we compare the result of our model with the previous works using manually features on the eNterface'05 and Berlin datasets. Table 3, reports the performance comparison of audio emotion recognition between our CRNN model and the corresponding manually features and CNN deep features without considering temporal relationship exists in a video.

From Table 3, on the eNterface'05 dataset we compare our learning feature with hand-crafted features such as prosody features (Mansoorizadeh and Moghaddam Charkari, 2010), MFCC (Sahoo and Routray, 2016), hybrid prosody and spectral features (Badshah et al., 2017),

Table 3
Speech emotion recognition accuracy comparison with previous works on the eNterface'05 and Berlin dataset.

Datasets	Refs.	Audio features	Accuracy (%)
eNterface'05	Mansoorizadeh et al., (Mansoorizadeh and Moghaddam Charkari, Aug. 2010)	prosody, LDA	43
	Sahoo et al., (Sahoo and Routray, Sep. 2016)	MFCC	57
	Zhang et al., (Zhao et al., 2014)	Prosody+spectral	62.7
	Bejani et al., (Bejani et al., Feb. 2014)	Prosody+MFCC	54.9
	Zhalehpour et al., (Zhalehpour et al., Jul. 2017)	MFCC-RASTA-PLP	72.9
	Proposed model	**Melspectrogram+CRNN**	**67.7**
Berlin	Badshah et. al., (Badshah et al., Feb. 2017)	Spectrogram+CNN	65.5
	Mansoorizadeh et. al., (Mansoorizadeh and Charkari, 2020)	Prosody	71
	Farhoudi et. al., (Farhoudi et al., Sep. 2017)	Prosody+MFCC	66
	Proposed model	**Melspectrogram+CRNN**	**74.5**

MFCC and Relative Spectral Transform (RASTA) and Perceptual Linear Prediction (PLP) (Zhalehpour et al., 2017). According to these evaluations on the eNterface'05 dataset, our CRNN model has higher recognition accuracy than others except the one feature learning method of MFCC-RASTA-PLP (Zhalehpour et al., 2017). However, our audio emotion recognition model is more robust and general in comparison with other learning feature models.

Furthermore, we employ our CRNN model on the known Berlin audio dataset and compare our performance with the results of previous works such as using spectrogram and CNN feature learning (Badshah et al., 2017), prosody feature extraction method (Mansoorizadeh and Charkari, 2020), hybrid prosody and MFCC feature extraction and the modified BEL model for classification (Farhoudi et al., 2017). According to Table 3, our CRNN model for speech emotion recognition by considering spatial-temporal correlation information is more discriminative than the hand-crafted features. Moreover, the performance indicates that a log Mel-spectrogram is better than the spectrogram and the CRNN model is better than the CNN model.

4.4. Multimodal fusion results

In this work, we investigate the effectiveness of our fusion network model with two distinct fusion methods, i.e., feature-level fusion (early fusion) and decision-level fusion (late fusion). As previously described, our proposed model is a feature-level fusion that uses a high-level of learning features and integrate them to train in the MoBEL network. This model learns the correlation between the deep learning spatial-temporal information of unimodal and perform emotion classification on the global video features.

However, we experiment and compare two fusion methods, namely feature-level and decision-level fusion.

4.4.1. Decision-level fusion

Fig. 8 shows the structure of the decision-level fusion method. In this strategy, since each modality are complementary for emotion classification, different combinations are employed to make full use of all the multimodal features. As shown in Fig. 8, we obtain speech emotion classification results by using the CRNN model and in the visual stream, we perform the 3D-CNN model to get the facial emotion recognition results. Then we combine the results of two audio-visual stream and fed to the BEL network model for video emotion classification. It can be said that the BEL network of the final model plays a role in the stack generalization method in classifier combination techniques.

Also, to present the advantage of the BEL model, we employ other classifiers such as KNN, SVM, MLP models, and compare the experimental results for video emotion recognition in the decision-making level as shown in Table 4.

Moreover, we employ other ensemble rules in decision-level or score-level fusion such as Max, Min, Sum, Average and Product. Table 5 presents the recognition accuracy rate of different ensemble rules on decision-level fusion. According to Table 5, the "Product" rule yields the best performance. Since the "Product" rule is calculated by multiplying the score of all modalities and then output the class with the maximum score, it has the best performance in decision-level fusion.

4.4.2. Feature-level fusion

Fig. 9 presents the architecture of the feature-level fusion method that shows our proposed model. In our proposed model, we concatenate the latent features of each modality with 256 dimensions and create a 512-dimension feature vector. Then we fed the features of all video samples to the MoBEL model to get the video emotion classes. In our experiment, we have used two expert BEL networks and one gating BEL network in the MoBEL network model as describes in Section 3.3.

To present the advantages of the MoBEL model, we compare the MoBEL network with other classifiers such as Mixture of NN, MLP, BEL, SVM, RBF, and Weighted KNN. Table 6 presents the performance of

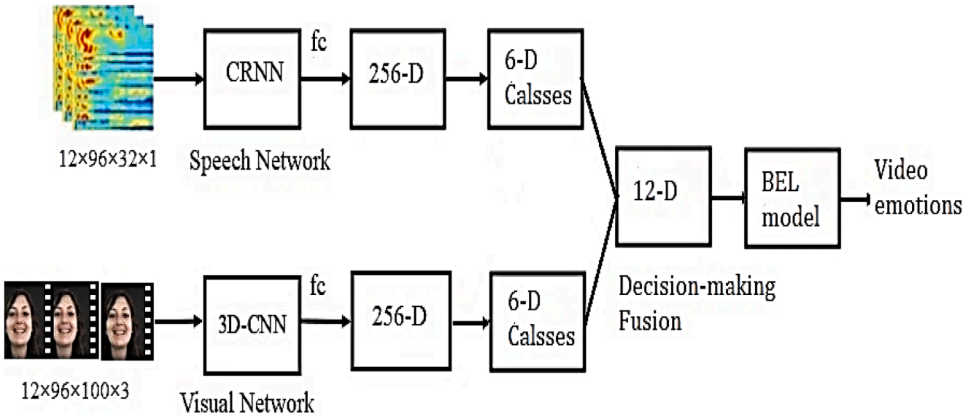

Fig. 8. The architecture of decision-level fusion with the BEL model as a classifier combination method in video emotion recognition.

Table 4
Accuracy rate comparison of different classifiers at decision-level for audio-visual emotion recognition.

Classifiers	Accuracy at decision-level (%)
BEL	**73.9**
MLP	71.5
SVM	72.5
KNN	72

Table 5
Comparison of the ensemble rules at the decision-level fusion for audio-visual emotion recognition.

Ensemble rules	Average	Sum	Min	Max	Product
Accuracy (%)	71.4	71.4	71	68.6	**74.3**

video emotion recognition of our model in comparison with other classifiers at feature-level fusion.

According to Table 6, the comparison clearly shows that the effectiveness of the MoBEL model in feature-level fusion, which shows its ability to learn joint audio-visual features from the output of deep models. Although, MoE neural network model is better than the other classifiers, using a bio-inspired BEL model in each expert of neural networks and gating network outperforms the other classifiers at feature-level fusion. The MoBEL model has such advantages as more efficient in terms of memory consumption, processing speed, and neuron numbers than the MOE ones. Furthermore, the MoBEL network can be used in the end-to-end audio-visual emotion recognition system and has the benefits of deep learning neural networks such as incremental learning. To testify the effectiveness of the MoBEL model, we show the MSE of the model in Fig. 10. According to this figure, the MoBEL, quickly learned and converged during the first 8 epochs for train, test, and validation data.

Also, compared with feature-level fusion and decision-level fusion methods, we found that the recognition accuracy of multimodal emotion recognition is significantly increased from 74% at decision-level fusion using product rule to 81.7% at the feature-level fusion using the MoBEL method. In our proposed model, the accuracy of video emotion recognition for all classes on average of 5-fold cross-validation is 81.7%.

Table 6
Video emotion recognition performance comparison between our proposed model and other classifiers at feature-level fusion

Classifiers	Accuracy (%)
MLP	78
BEL	78.7
SVM	77.9
Weighted KNN	78
RBF	71
Mixture of Expert NN	80
MoBEL	**81.7**

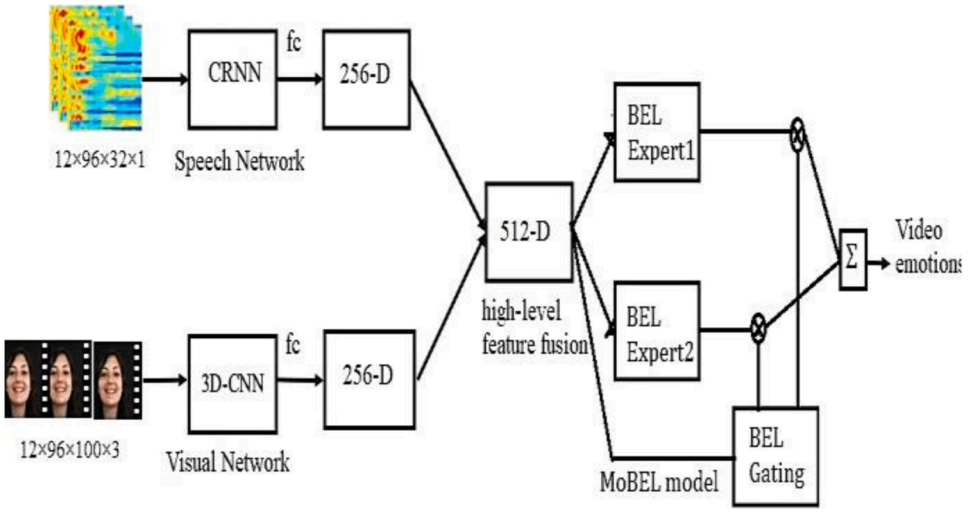

Fig. 9. The architecture of feature-level fusion with the MoBEL network in audio-visual emotion recognition.

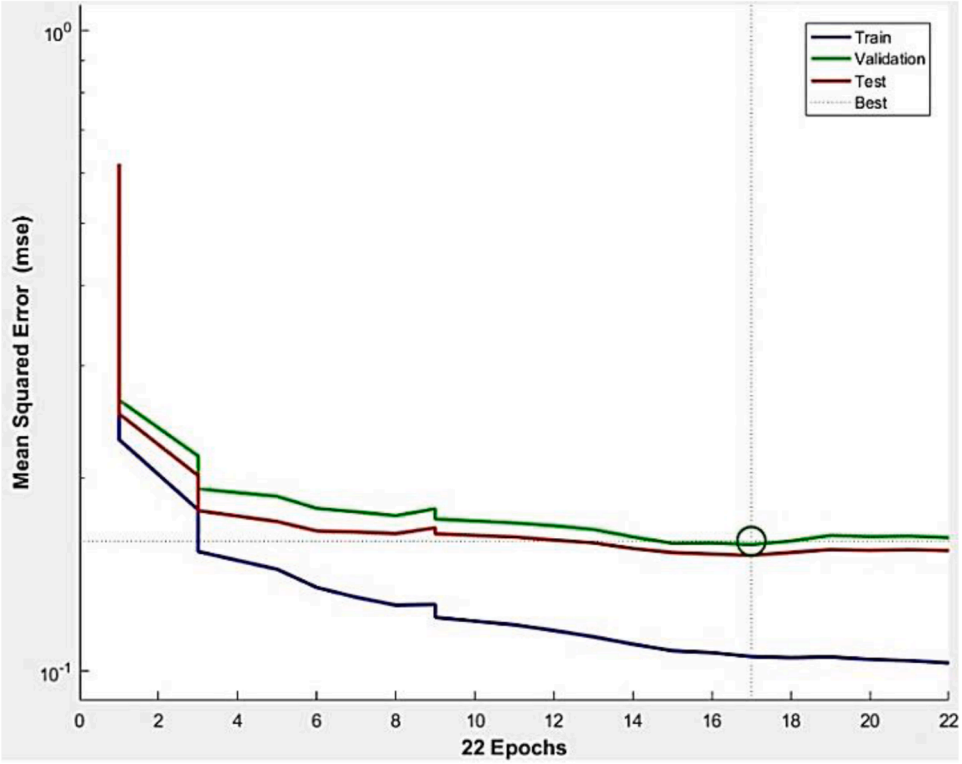

Fig. 10. The MSE error during the first 22 learning epochs of the MoBEL model for train, test and validation data.

Fig. 11 presents the confusion matrix of multimodal emotion recognition using the MoBEL model at feature-level fusion on the eNterface'05 dataset.

According to Fig. 11, it is interesting to find that in the eNterface'05 dataset, "disgust" and "anger" emotions have the higher recognition accuracy and "surprise" and "fear" are recognized with lower accuracy, *i.e.*, about less than 78%, while other emotions have a recognition accuracy higher than 80%. The reason might be that the emotion of "fear" and "sadness" are overlapping as much as 14% of the emotional examples of fear are in sadness. Therefore, it is difficult to identify "fear" in audio-visual emotion recognition. Additionally, it is indicated from this figure that the "disgust" is identified well with recognition accuracy of about 88.2%.

In Fig. 12 we demonstrate the recognition accuracy based on audio, visual, and audio-visual emotion recognition.

As shown in Fig. 11, the recognition accuracy of audio-visual modality is more than each unimodal one. Furthermore, the figure illustrates that the audio-visual emotion recognition of our proposed model could deal with low accuracy in each modality. For example, in facial emotion recognition, the "surprise" emotion has low accuracy, while in speech emotion recognition, the "disgust" emotion has low accuracy and audio-visual emotion recognition at feature-level fusion outperforms all unimodal methods. In addition to the confusion matrix, precision and recall rates are calculated to further compare the multimodal emotion recognition performance. The experimental results are shown in Table 7.

We compare our proposed model results with the start-of-art results on eNterface'05 dataset as well. Table 8 presents the comparison of the proposed model with the previous works that perform subject-independent cross-validation experiments in audio-visual emotion

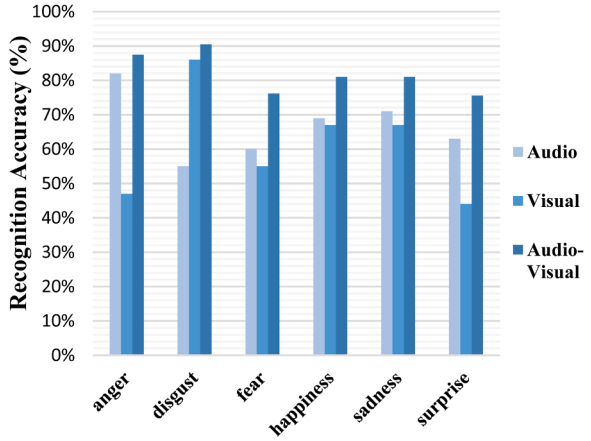

Fig. 11. Confusion matrix of audio-visual emotion recognition results using the MoBEL model on the eNterface'05 dataset.

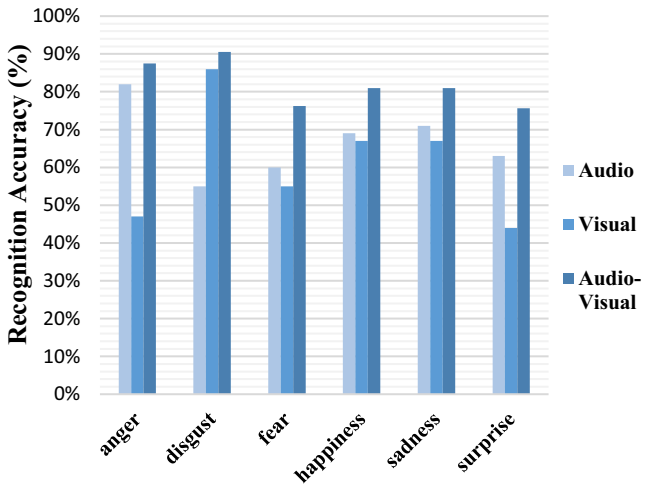

Fig. 12. Comparison between emotion recognition of audio, visual and audio-visual on the eNterface'05 dataset.

Table 7
Audio-visual emotion recognition performance in our proposed model on eNterface'05 dataset.

Emotion	Precision (%)	Recall (%)
Anger	85	86
Disgust	89.7	78
Fear	65.5	68
Happiness	80.1	71.3
Sadness	87.5	71.5
Surprise	83.21	86

Table 8
Audio-visual emotion recognition comparison with state-of-art works on the eNterface'05 dataset

Ref.	Fusion model	Audio accuracy (%)	Visual accuracy (%)	Final accuracy recognition (%)
Mansoorizadeh et. al., (Mansoorizadeh and Moghaddam Charkari, Aug. 2010)	Hybrid	43	37	71
Bejani et. al., (Bejani et al., Feb. 2014)	Hybrid	54.9	39.2	77.7
Zhalehpour et. al., (Zhalehpour et al., Jul. 2017)	Decision-making	72.9	42.1	77
Sahoo et. al., (Sahoo and Routray, Sep. 2016)	Rule-based	57	45	77.02
Zhange et. al., (Zhao et al., 2014)	Product decision-making	62.7	44.7	67.4
Proposed Model	**MoBEL**	**67**	**62**	**81.7**

recognition. In this evaluation, we show the fusion method, speech emotion recognition accuracy, visual emotion recognition accuracy, and final accuracy after employ fusion method for all classes on the eNterface'05 dataset.

According to Table 8, our proposed model outperforms previous works (Mansoorizadeh and Moghaddam Charkari, 2010; Sahoo and Routray, 2016; Zhalehpour et al., 2017; Bejani et al., 2014; Zhao et al., 2014) by more than about 4%. Also, it is indicated that our feature representation methods using 3D-CNN and CRNN for visual and audio emotion recognition have better results than hand-crafted feature methods in other works. Besides, Although, the speech emotion recognition accuracy rate of Zhalehpour et al. (2017) is more than our result of speech emotion recognition, the final accuracy follows by the MoBEL fusion model is much better than the final result of Zhalehpour et. al., after decision-making fusion. Thus, this indicates the advantage of our deep learning spatial-temporal feature extraction and feature-level fusion network, namely the MoBEL model.

5. Conclusion and future works

In this work, we have presented a new method for audio-visual emotion recognition with the MoBEL fusion network model that integrates high-level learning features using CRNN and 3D-CNN for audio and visual modalities, respectively. We find that deep learning spatial-temporal features have a more powerful ability in discriminating video emotions than the static and hand-crafted features. For the visual modality representation, the 3D-CNN model is used to learn the spatial-temporal features of visual expression, while for auditory modality the Mel-spectrograms of the speech signals are fed into the CRNN model for spatial-temporal audio feature extraction. To improve the performance of deep learning feature representations, as future work, we plan to optimize our framework by fine-tuning our CNN and 3D-CNN features in the training process.

The feature representations of audio and visual networks are fed into the fusion model, namely MoBEL model to jointly learn audio-visual discriminative feature representations. The MoBEL network is a mixture of BEL neural networks inspired by the brain limbic system and also the associative cortex of the brain, which can handle information integration from many sources. Experimental results on the eNterface'05 dataset show that our proposed model performs better than previous manual feature extraction and other fusion methods on emotion recognition. Also, the MoBEL model boosts the system's accuracy more than other state-of-the-art information fusion models in video emotion recognition. That is, because the MoBEL network has better ability in capturing the highly non-linear correlations between audio and visual modalities which in terms of memory consumption and speed processing is more efficient than the other classification methods. Since the BEL model has one hidden layer and, due to its structure in terms of memory consumption is more efficient. Furthermore, an end-to-end learning strategy would be more concise by employing a two-stage learning strategy, *i.e.*, the deep learning feature extraction and feature-level MoBEL fusion network, to train the audio-visual networks automatically. Thus, end-to-end learning strategy would be investigated in our future work.

One of the advantages of the proposed model is that it is a general model and can be used in other applications, such as driver drowsiness detection and video action recognition. Also, to improve the accuracy performance of the model, we need a variety of datasets.

Besides, in the proposed model feature representation of emotional speech and facial expression, and then integrating information to get video emotion classes were performed for the entire video. For long videos, the idea in our future work is first to divide the whole video into segments in about four seconds, and then the proposed model applies in each segment so that a jointly audio-visual learning feature is achieved. Therefore, as a future work by performing some preprocessing operations we will evaluate the proposed models on spontaneous audio-visual emotion datasets such as AFEW and BAUM-1s.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2020.12.001.

References

L. Shen, M. Wang, and R. Shen, "Affective e-learning: using 'emotional' data to improve learning in pervasive learning environment," p. 15, 2020.
Beale, R., Peter, C., 2008. The role of affect and emotion in HCI. Affect Emot. Hum.-Comput. Interact. 1–11.
Torous, J.B., Friedman, R.S., Keshavan, M.S., 2014. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. JMIR MHealth UHealth 2.
Lv, Z., Feng, S., Feng, L., Li, H., 2015. Extending touch-less interaction on vision based wearable device. In: 2015 IEEE Virtual Reality (VR), pp. 231–232. Mar.
Szwoch, M., Szwoch, W., 2015. Emotion recognition for affect aware video games. Image Process. Commun. Chall. 6 227–236.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. Jun.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks. In: Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497.
Paleari, M., Huet, B., Chellali, R., Jul. 2010. Towards multimodal emotion recognition: a new approach. In: Proceedings of the ACM International Conference on Image and Video Retrieval. Xi'an, China, pp. 174–181.

Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q., Oct. 2018. Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE Trans. Circuits Syst. Video Technol. 28 (10), 3030–3043.
Kim, Y., Lee, H., Provost, E.M., May 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687–3691.
Zhang, S., Zhang, S., Huang, T., Gao, W., Jun. 2016. Multimodal deep convolutional neural network for audio-visual emotion recognition. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. New York, New York, USA, pp. 281–284.
Stein, B.E., Stanford, T.R., Rowland, B.A., Dec. 2009. The neural basis of multisensory integration in the midbrain: Its organization and maturation. Hear. Res. 258 (1), 4–15.
Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., Mar. 1991. Adaptive mixtures of local experts. Neural Comput 3 (1), 79–87.
Christian Balkenius, J.M., Sep. 2001. Emotional learning: a computational model of the amygdala. Cybern. Syst. 32 (6), 611–636.
C. Balkenius and J. Morén, "A computational model of emotional conditioning in the brain," 1998.
Babaie, T., Karimizandi, R., Lucas, C., Jul. 2008. Learning based brain emotional intelligence as a new aspect for development of an alarm system. Soft Comput. 12 (9), 857–873.
Lucas, C., Shahmirzadi, D., Sheikholeslami, N., Jan. 2004. Introducing BELBIC: brain emotional learning based intelligent controller. Intell. Autom. Soft Comput. 10 (1), 11–21.
Lucas, C., 2011. BELBIC and Its industrial applications: towards embedded neuroemotional control codesign. In: Integrated Systems, Design and Technology 2010. Berlin, Heidelberg, pp. 203–214.
Asad, M.U., et al., 2017. Neo-fuzzy supported brain emotional learning based pattern recognizer for classification problems. IEEE Access 5, 6951–6968.
Farhoudi, Z., Setayeshi, S., Rabiee, A., Sep. 2017. Using learning automata in brain emotional learning for speech emotion recognition. Int. J. Speech Technol. 20 (3), 553–562.
Lotfi, E., Khazaei, O., Khazaei, F., Apr. 2018. Competitive brain emotional learning. Neural Process. Lett. 47 (2), 745–764.
Jafari, M., Xu, H., Mar. 2019. A biologically-inspired distributed fault tolerant flocking control for multi-agent system in presence of uncertain dynamics and unknown disturbance. Eng. Appl. Artif. Intell. 79, 1–12.
Fang, W., Chao, F., Lin, C.-M., Yang, L., Shang, C., Zhou, C., 2019. An improved fuzzy brain emotional learning model network controller for humanoid robots. Front. Neurorobotics 13.
Zhao, J., Lin, C.-M., Chao, F., 2019. Wavelet fuzzy brain emotional learning control system design for MIMO uncertain nonlinear systems. Front. Neurosci. 12.
Wang, Y., Guan, L., Aug. 2008. Recognizing human emotional state from audiovisual signals*. IEEE Trans. Multimed. 10 (5), 936–946.
Wang, Y., Guan, L., Venetsanopoulos, A.N., Jun. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Trans. Multimed. 14 (3), 597–607.
Zhao, G., Pietikainen, M., Jun. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 915–928.
Dhall, A., Asthana, A., Goecke, R., Gedeon, T., Mar. 2011. Emotion recognition using PHOG and LPQ features. In: Face and Gesture 2011, pp. 878–883.
Lucey, S., Ashraf, A.B., Cohn, J.F., 2007. Investigating spontaneous facial action recognition through AAM representations of the face. In: Face Recognition Book. Pro Literatur Verlag.
Chang, Y., Hu, C., Feris, R., Turk, M., Jun. 2006. Manifold based analysis of facial expression. Image Vis. Comput. 24 (6), 605–614.
Trigeorgis, G., et al., Mar. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204.
Mansoorizadeh, M., Moghaddam Charkari, N., Aug. 2010. Multimodal information fusion application to human emotion recognition from face and speech. Multimed. Tools Appl. 49 (2), 277–297.
Fan, Y., Lu, X., Li, D., Liu, Y., Oct. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan, pp. 445–450.
Zeng, Z., Tu, J., Pianfetti, B.M., Huang, T.S., Jun. 2008. Audio–visual affective expression recognition through multistream fused HMM. IEEE Trans. Multimed. 10 (4), 570–577.
Ntalampiras, S., Fakotakis, N., Jan. 2012. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. IEEE Trans. Affect. Comput. 3 (1), 116–125.
Schuller, B., Müeller, R., Höernler, B., Höethker, A., Konosu, H., Rigoll, G., Nov. 2007. Audiovisual recognition of spontaneous interest within conversations. In: Proceedings of the 9th international conference on Multimodal interfaces. Nagoya, Aichi, Japan, pp. 30–37.
Sahoo, S., Routray, A., Sep. 2016. Emotion recognition from audio-visual data using rule based decision level fusion. In: 2016 IEEE Students' Technology Symposium (TechSym), pp. 7–12.
Gurban, M., Thiran, J.-P., Drugman, T., Dutoit, T., Oct. 2008. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition. In: Proceedings of the 10th International Conference on Multimodal interfaces. Chania, Crete, Greece, pp. 237–240.
Chen, S., Jin, Q., 2015. Multi-modal dimensional emotion recognition using recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15. Brisbane, Australia, pp. 49–56.
Lan, Z., Bao, L., Yu, S.-I., Liu, W., Hauptmann, A.G., Jul. 2014. Multimedia classification and event detection using double fusion. Multimed. Tools Appl. 71 (1), 333–347.
Pini, S., Ahmed, O.B., Cornia, M., Baraldi, L., Cucchiara, R., Huet, B., Nov. 2017. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow, UK, pp. 536–543.
Hall, J.E., 2015. Guyton and Hall Textbook of Medical Physiology E-Book, 13 ed. Saunders.
Liu, W., et al., 2016. SSD: single shot MultiBox detector. In: Computer Vision – ECCV 2016. Cham, pp. 21–37.
Lim, W., Jang, D., Lee, T., Dec. 2016. Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4.
Busso, C., Lee, S., Narayanan, S.S., 2007. Using neutral speech models for emotional speech analysis. INTERSPEECH.
M. S. Jafari, H. Xu, and L. R. G. Carrillo, "A neurobiologically-inspired intelligent trajectory tracking control for unmanned aircraft systems with uncertain system dynamics and disturbance," 2019.
Martin, O., Kotsia, I., Macq, B., Pitas, I., Apr. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06), 8-8.
Schuller, B., et al., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: In Proc. Interspeech.
Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E., Jul. 2017. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. IEEE Trans. Affect. Comput. 8 (3), 300–313.
Dhall, A., Goecke, R., Lucey, S., Gedeon, T., Jul. 2012. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed 19 (3), 34–41.
Bejani, M., Gharavian, D., Charkari, N.M., Feb. 2014. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. Neural Comput. Appl. 24 (2), 399–412.
Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G., Jan. 2019. Audio-visual emotion recognition in video clips. IEEE Trans. Affect. Comput. 10 (1), 60–75.
Vielzeuf, V., Pateux, S., Jurie, F., Nov. 2017. Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow, UK, pp. 569–576.
Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W., Feb. 2017. Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–5.
X. Zhao, S. Zhang, X. Wang, and G. Zhang, "Multimodal emotion recognition integrating affective speech with facial expression," 2014.
M. Mansoorizadeh and N. M. Charkari, "Speech emotion recognition: comparison of speech segmentation approaches," 2020 p. 5.