Information Fusion 49 (2019) 69–78

Contents lists available at ScienceDirect

Information Fusion

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

Emotion recognition using deep learning approach from audio–visual emotional big data

M. Shamim Hossain[a,*], Ghulam Muhammad[b]

[a] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[b] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

A B S T R A C T

This paper proposes an emotion recognition system using a deep learning approach from emotional Big Data. The Big Data comprises of speech and video. In the proposed system, a speech signal is first processed in the frequency domain to obtain a Mel-spectrogram, which can be treated as an image. Then this Mel-spectrogram is fed to a convolutional neural network (CNN). For video signals, some representative frames from a video segment are extracted and fed to the CNN. The outputs of the two CNNs are fused using two consecutive extreme learning machines (ELMs). The output of the fusion is given to a support vector machine (SVM) for final classification of the emotions. The proposed system is evaluated using two audio–visual emotional databases, one of which is Big Data. Experimental results confirm the effectiveness of the proposed system involving the CNNs and the ELMs.

1. Introduction

The use of automatic emotion recognition has a great potential in various intelligent systems, including digital advertisement, online gaming, customers' feedback assessment, and healthcare. For example, in an online gaming system, if there is an emotion recognition component, the players can have more excitement, and the gaming display can be adjusted according to the emotion. In an online shopping system, if there is a live emotion recognition module, the selling company can get immediate emotional feedback from the customers, and thereby can present a new deal to the customers. In a healthcare system embedded with an emotion recognition module, patients' mental and physical states can be monitored, and appropriate medicine or therapy can be prescribed [1].

Recently, emotion-aware intelligent systems are in use in different applications. The applications include emotion-aware e-health systems, affect-aware learning systems, recommendation systems for tourism, affect-aware smart city, and intelligent conversational systems. Many of these systems are based on text or emoticons inputs. For example, an emotion-aware e-health systems were proposed in [2,3]. Various keywords were searched from textual feedback from the patients and emotions were recognized from these keywords. Therefore, the input to this system is text, not speech or video. An intelligent tutoring system integrating emotion-aware framework was described in [4]. In this system, students are allowed to express their satisfaction using texts or emoticons. A similar affect-aware learning technology was introduced in [5]. A recommendation system for tourism using context or emotion was presented in [6]. A healthcare recommender system called iDoctor was introduced in [7] using a text sentiment analysis based on emotions. To enhance the experience of smart city inhabitants, an affect-aware smart city was proposed using a detection and visualization of emotions [8]. The emotions were recognized using the keywords, hashtags, and emoticons. An interesting smart home system embedding botanical Internet of Things (IoT) and emotion detection was introduced in [9]. In this system, an effective communication between smart greeneries (in smart greenhouses) and home users was established. All of these systems were based on text or emoticons.

Emotions can be detected using different forms of inputs, such as speech, short phrases, facial expression, video, long text, short messages, and emoticons. These input forms vary across applications. In social media, the most common forms are short texts and emoticons; in the gaming system, the most common form is video. Recently, electroencephalogram (EEG) signal-based emotion recognition systems are also proposed [10,11]; however, the use of EEG cap is invasive and hence, uncomfortable to the users. Based on a review of the related available literature, we find that only one input modal does not provide the desired accuracy of emotion recognition [12,13]. Though there exist different input modalities for emotion recognition, the most common is a bimodal input with a combination of speech and video. These two are chosen because both can be captured in a non-invasive manner and more expressive than other input modalities.

Though there are several previous works on audio-visual emotion recognition in the literature, most of them suffer from low recognition accuracies. One of the main reasons behind that is the way to extract features from these two signals and the fusion between them [14]. In most of the cases, some handcrafted features are extracted, and the features from the two signals are combined using a weight.

This paper proposes an audio–visual emotion recognition system using a deep network to extract features and another deep network to fuse the features. These two networks ensure a fine non-linearity of fusing

* Corresponding author.
E-mail address: mshossain@ksu.edu.sa (M.S. Hossain).

https://doi.org/10.1016/j.inffus.2018.09.008
Received 24 November 2017; Received in revised form 10 August 2018; Accepted 9 September 2018
Available online 13 September 2018
1566-2535/

Table 1
Summary of previous work on emotion recognition from speech using deep learning approach.

Ref.	Method	Database	Accuracy (%)
[16]	Segment-level features and DNN; utterance-level features and ELM	IEMOCAP	54.3
[20]	Sparse autoencoder-based feature transfer learning	EMO-DB; eNTERFACE	Recall: 57.9; 59.1
[24]	Linear regression, DBN	MoodSwings Lite	Error rate: 5.41
[25]	Speech features; SVM; DBN	Chinese Academy of Sciences emotional speech database	94.6 (using DBN)
[26]	Spectrogram; DBN	IEMOCAP	64.78
[27]	Prosodic features, spectrum features; ELM	CASIA Chinese emotion corpus	89.6
[29]	Probabilistic echo-state network	WaSeP	96.69
[30]	Spectrogram; Deep Retinal Convolution Neural Networks (DRCNNs)	IEMOCAP	99.25

Table 2
Summary of previous work on emotion recognition from the image using deep learning approach.

Ref	Method	Database	Accuracy (%)
[31]	CNN	EmotiW 2015	55.6
[34]	IDP; ELM	eNTERFACE	84.12
[35]	HoG; Deep sparse autoencoders	CK+	96
[37]	DNN	CK+	93.2
[38]	FaceNet2ExpNet	CK+	96.8
[39]	Deep Neural Networks with Relativity Learning (DNNRL)	FER-2013	70.6

the features. The final classification is done using a support vector machine (SVM). The deep learning has been extensively used nowadays in different applications such as image processing, speech processing, and video processing. The accuracies in various applications using the deep learning approach vary due to the structure of the deep model and the availability of huge data [15]. The contributions of this paper are (i) the proposed system is trained using Big Data of emotion and, therefore, the deep networks are trained well, (ii) the use of layers, one layer for gender separation and another layer for emotion classification, of an extreme learning machine (ELM) during fusion; this increases the accuracy of the system, (iii) the use of a two dimensional convolutional neural network (CNN) for audio signals and a three dimensional CNN for video signals in the proposed system; a sophisticated technique to select a key frame is also proposed, and (iv) the use of the local binary pattern (LBP) image and the interlaced derivative pattern (IDP) image together with the gray-scale image of key frames in the three dimensional CNN; in this way, different informative patterns of key frames are given to the CNN for feature extraction.

The rest of the paper is structured as follows. Section 2 gives a related literature review. Section 3 presents the proposed emotion recognition system. Section 4 shows the experimental results and provides discussion. The paper is concluded in Section 5.

2. Related previous work

This section is divided into three parts. These parts give an overview of some exiting works of emotion recognition from speech signals, image or video signals, and both speech and video signals, respectively.

2.1. Emotion recognition from speech

Han et al. used both segment-level features such as Mel-frequency cepstral coefficients (MFCC), pitch period, and harmonic to noise ratio, and utterance-level features to detect emotions. Deep neural networks (DNNs) were utilized to create emotion probabilities in each speech segment [16]. These probabilities were used to generate the utterance-level features, which were fed to the ELM based classifier. The interactive emotional dyadic motion capture (IEMOCAP) database [17] was used in the experiments. 54.3% accuracy was obtained by the method. High-order statistical features and a particle swarm optimization-based feature selection method were used to recognize emotion from a speech signal in [18]. The obtained accuracy was between 90% and 99.5% in the Berlin Emotional Speech Database (EMO-DB) [19]. Deng et al. proposed a sparse autoencoder-based feature transfer learning method for emotion recognition from speech [20]. They used several databases including the EMO-DB and the eNTERFACE database [21]. Prosodic features together with paralinguistic features were used to detect emotions in [22]. An accuracy of around 95% was obtained using the EMO-DB database. A collaborative media framework using emotion from speech signals was proposed in [23]. Conventional features such as the MFCCs were used in the proposed framework.

A technique based on linear regression and the DBN was used to recognize musical emotion in [24]. An error rate of 5.41% was obtained by the technique in a music database named MoodSwings Lite. Deep belief networks (DBNs) and the SVM were investigated using the Chinese Academy of Sciences emotional speech database in [25]. The accuracy using the SVM was 84.54% and that using the DBNs was 94.6%. In [26], the authors proposed a deep learning framework in the form of convolutional neural networks (CNNs), where the input was the spectrogram of the speech signal. They achieved 64.78% accuracy in the IEMOCAP database. The ELM based decision tree was used to recognize emotions from a speech in [27]. This method achieved 89.6% accuracy using the CASIA Chinese emotion corpus [28]. A probabilistic echo-state network-based emotion recognition system was proposed in [29]. Using the WaSeP database, the system obtained 96.69% accuracy%. A more recent work as described in [30] introduced a deep retinal CNNs (DRCNNs), which was proved to be successful to recognize emotions from speech signals. It achieved an accuracy as high as 99.25% in the IEMOCAP database. Table 1 summarizes the previous works on emotion recognition from speech signals using the deep learning techniques.

2.2. Emotion recognition from image or video frames

Ng et al. used the CNN with the transfer learning from the ImageNet to recognize emotions from static images [31]. Using the 2015 Emotion Recognition sub-challenge dataset of static facial expression, the authors achieved 55.6% accuracy. A local binary pattern (LBP), Gaussian mixture model (GMM) and support vector machine (SVM) based emotion recognition system from images was proposed in [32]. The system achieved an accuracy of 99.9% using the Cohn-Kanade (CK) database [33]. An interlaced derivative pattern (IDP) and the ELM based emotion recognition system from images was introduced in [34]. Using the eNTERFACE database, the system obtained 84.12% accuracy.

Table 3
Summary of previous work on emotion recognition from audio–visual modality using deep learning approach.

Ref	Method	Database	Accuracy (%)
[41]	Feature selection and DBN	IEMOCAP; contains facial markers	70.46–73.78
[42]	CNN for video, DBN for audio, 'bag-of-mouth' model, and autoencoder	EmotiW 2014	47.67
[44]	Multidirectional regression, SVM	eNTERFACE	84
[45]	CDBN	MAHNOB-HCI	58.5
[46]	Mel-spectrogram; face images; CNN for audio, 3D CNN for video	eNTERFACE	85.97
[47]	MDR, ridgelet transform; ELM	eNTERFACE	83.06
[49]	audio features, facial features; triple stream DBN model	eNTERFACE	66.54 (correlation rate)
[50]	Audio features, dense features, CNN based features	EmotiW 2015; CK+	54.55; 98.47

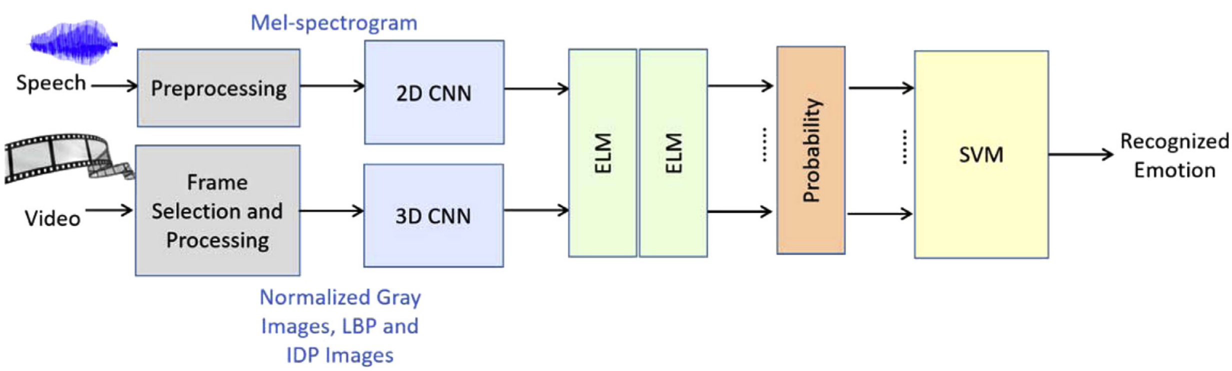

Fig. 1. An overall block diagram of the proposed emotion recognition system.

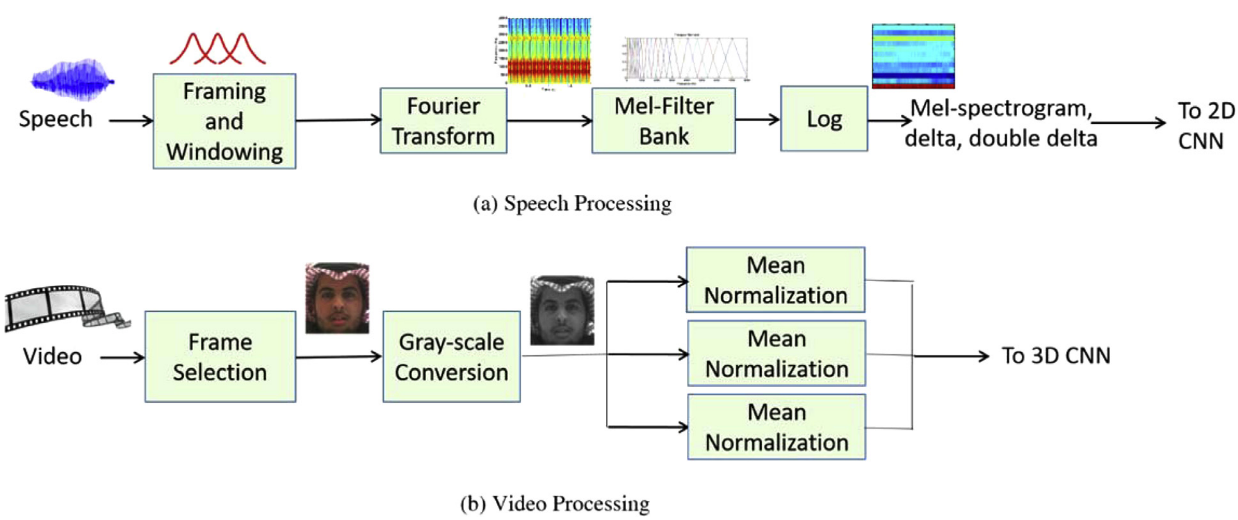

Fig. 2. Preprocessing steps of speech (top row) and video (bottom row) in the proposed system.

Zeng et al. proposed a histogram of oriented gradients (HoG) features and deep sparse autoencoder based emotion recognition system from images in [35]. Using the extended CK database (CK+), they got around 96% accuracy. A mobile application of emotion recognition from faces was developed in [36]. In the application, a bandlet transform and the LBP were used to extract facial features, and the GMM was used as the classifier. An accuracy of 99.7% was achieved using the CK database.

A deep neural network (DNN) based approach to recognize emotion was proposed in [37]. The input to the DNN was the raw face image. 93.2% accuracy was found using the CK+database. A deep network combining several deep models was introduced in [38]. The authors called the network as FaceNet2ExpNet, and the network achieved 96.8% accuracy with the CK+database. Deep Neural Networks with Relativity Learning (DNNRL) model was developed in [39] to recognize emotion from face images. 70.6% accuracy was obtained using the FER-2013 database. The HoG descriptors followed by a principal component analysis and a linear discriminant analysis were used in an emotion recognition system in [40]. The system got more than 99% accuracy with the CK+database. Table 2 summarizes the previous works on emotion recognition from face images using the deep learning techniques.

2.3. Emotion recognition from speech and video

Kim et al. proposed an emotion recognition system using both speech and video modalities [41]. A feature selection technique was used before feeding the features to a DBN. The IEMOCAP database was used; the database contains face images with facial markers. Accuracies between 70.46% and 73.78% were obtained by some variants of the system. A challenge audio-visual database was used for emotion recognition in [42]. The authors in [42] investigated different deep models to recognize emotions. Specifically, they used a CNN for video, the DBN for audio, a 'bag-of-mouth' model to extract features around the mouth re-

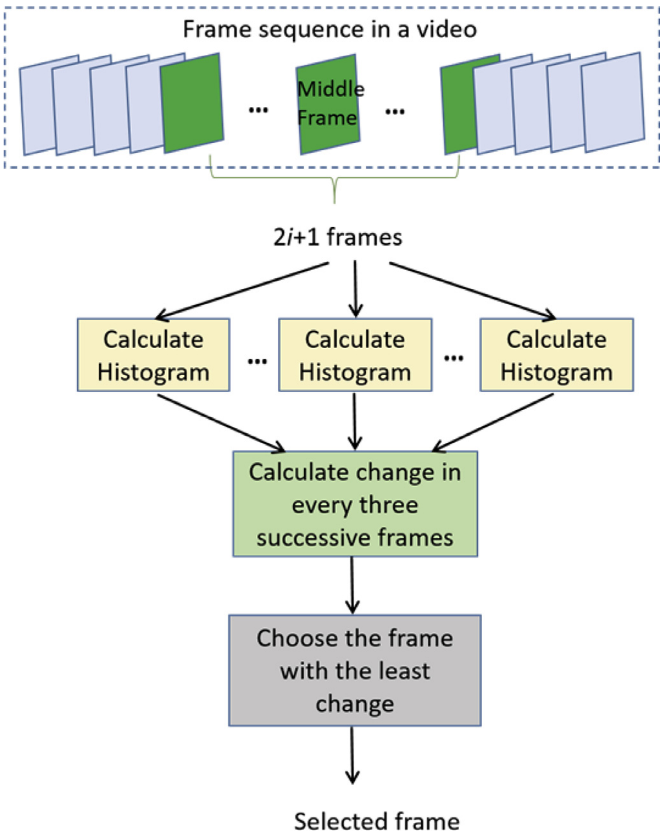

Fig. 3. Process flow-chart of selecting frames from a video in the proposed system.

gion in the video, and a relational autoencoder. An accuracy of 47.67% was achieved by their model. The authors reported recalls of 57.9% and 59.1% using the two databases, respectively. An audio-visual cloud gaming framework was proposed in [43], where the gaming experience of the users was improved by a feedback based on the recognized emotion of the users. MPEG-7 features from audio and video signals were used to classify emotions.

An emotion recognition system based on multidirectional regression and the SVM was proposed in [44]. An accuracy of 84% was obtained in the eNTERFACE database. The authors found that different directional filters were effective to recognize emotions. A convolutional DBN (CDBN) was introduced to recognize emotions in [45]. An accuracy of 58.5% was achieved by the authors using the MAHNOB-HCI multimodal database. An emotion recognition system using audio-visual pre-trained models were used in [46]. A Mel-spectrogram was used as the input to the CNN for the audio signal, and the face frames were the inputs to a 3D CNN for the video signal. Using the eNTERFACE database, the system showed around 86% accuracy.

An audio-visual emotion recognition system was proposed in [47], where a multidirectional regression (MDR) and a ridgelet transform based features were utilized. The ELM was used as the classifier. The obtained accuracy was 83.06%. A multimodal system for emotion recognition using prosody and format features for audio and quantized image matrix features for images was introduced in [48]. Using the eNTERFACE database, the system achieved the accuracy more than 77%.

In [49], the authors suggested a system using audio features and facial features to recognize emotion. A triple-state stream DBN model was used as the classifier. A correlation rate of 66.54% was obtained in the eNTERFACE database. In a recent study, audio features from speech signals, dense features from image frames, and CNN-based features from image frames were fused at the score level to recognize emotion [50]. The accuracies were 54.55% and 98.47% using the EmotiW 2015 database and the CK+database, respectively. Table 3 summarizes the previous works on emotion recognition from audio-visual modality using the deep learning techniques.

3. Proposed audio–visual emotion recognition system

From the above literature review, we find that the existing systems were not evaluated in Big Data. Moreover, the obtained accuracies are still below expectation. Therefore, we propose, in this paper, a system that will work well using Big data.

Fig. 1 shows an overall block diagram of the proposed emotion recognition system. There are two modalities of input to the system: speech and video. Speech signals and video signals are processed separately and fused at the later stage before classification. There are two main steps for each of these modalities before fusion. The steps are preprocessing and deep networks using the CNN. We tested different fusion strategies, and finally, proposed an ELM based fusion, which will be described later.

3.1. Speech signal preprocessing

In the proposed system, a Mel-spectrogram is obtained from the speech signal. The steps to get the Mel-spectrogram are given below.

Step 1 – Divide the signal into 40 ms frames, where the successive frames are overlapped by 50%.
Step 2 – Multiply the frames by a Hamming window.
Step 3 – Apply fast Fourier transform to the windowed frame to convert the time-domain segment into the frequency-domain one.
Step 4 – Apply 25 band-pass filters (BPFs) to the frequency-domain signal. The center frequencies of the filters are distributed on a Mel scale, and the bandwidths of the filters follow the critical bandwidth of human auditory perception.
Step 5 – Perform the logarithm function on the filter outputs to suppress the dynamic range.
Step 6 – Arrange the outputs of the previous steps frame by frame to form the Mel-spectrogram of the signal.

Fig. 2(a) shows the preprocessing steps of the speech signal in the proposed system. The Mel-spectrogram is the input to the CNN. We process the signal for every 2.02 s. Therefore, the size of the Mel-spectrogram is 25×100 (5 filters and 100 frames).

Hand-crafted or conventional speech features can achieve good recognition performance with clean or slightly noisy speech data; how-

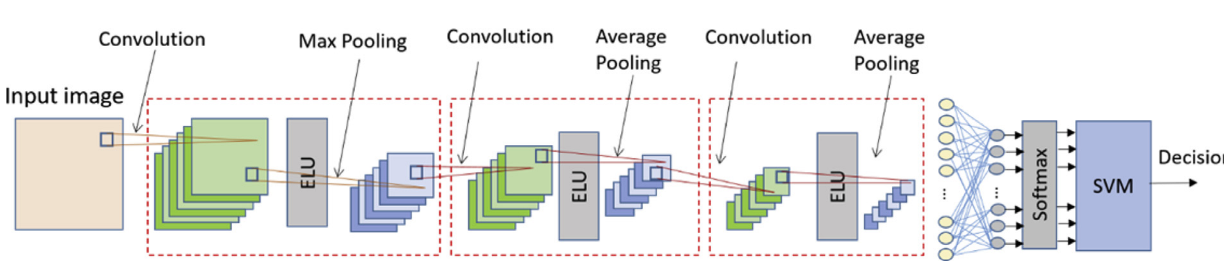

Fig. 4. The structure of the 2D CNN followed by the SVM in the proposed system for speech signals.

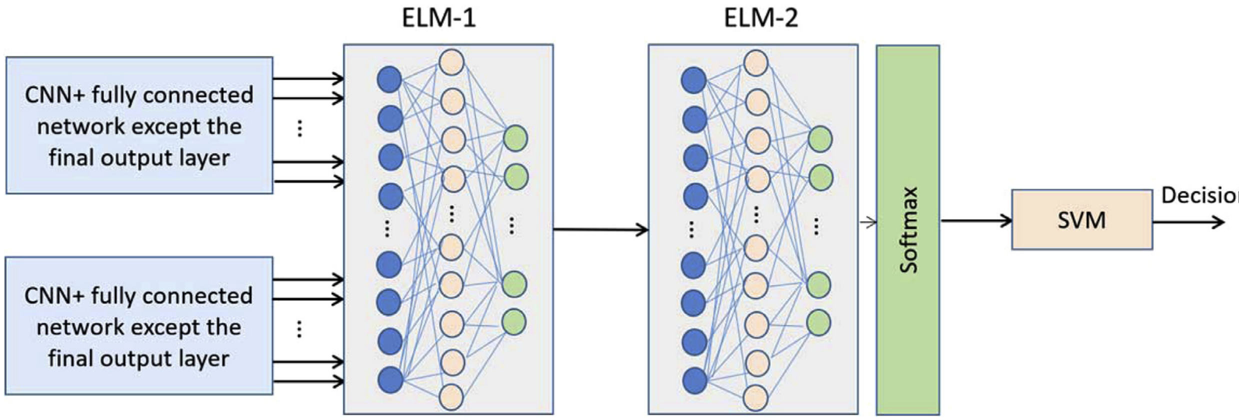

Fig. 5. The proposed ELM-based fusion.

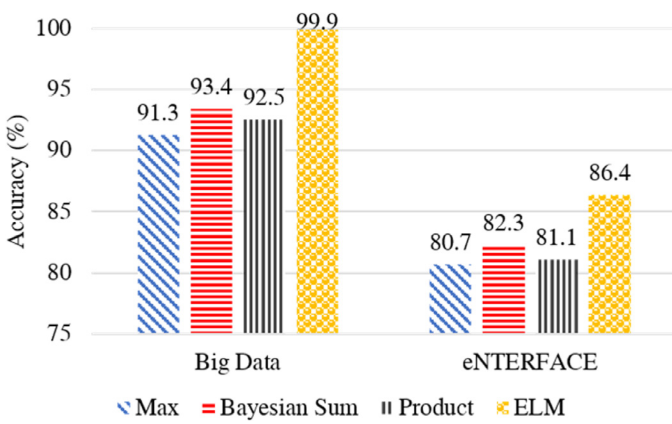

Fig. 6. Accuracy of the proposed system using different fusion strategies.

ever, they fail to a significant amount in noisy data. In contrast, the deep models extract features using a high degree of non-linearity and encode variations of signals. Therefore, we use the CNN models in our system. The CNN models require images as the input. Normally, the images are having three channels (red, green, and blue). To be consistent with this representation, we obtain velocity (delta) and acceleration (double delta) coefficients using a window size of three from the Mel-spectrogram. Therefore, we have the Mel-spectrogram image (converted to gray), its delta image and the double delta image to be analogous with the three channels. The delta and double delta coefficients encode relative temporal information of a speech signal.

3.2. Video signal preprocessing

Fig. 2(b) shows the preprocessing steps of the video signal in the proposed system. The first step is to select some key frames from a 2.02 s video segment. The process of selecting key frames is shown in Fig. 3. In a window of $2i+1$ frames, where i is set to three (empirically), we calculate the histograms of the frames. A chi-square distance is applied to find the difference of successive-frames' histograms. The frame with the least difference is selected as the key frame in that sequence. Before calculating the histograms, we apply a face detection algorithm (in our case, we used the Viola-Jones algorithm [51]) to crop the face area. The histograms are obtained from the cropped face images. If there was no face detected in a frame, we ignored that frame for subsequent pro-

(a) with augmenting

Actual \ Predicted	Anger	Disgust	Fear	Happiness	Sad	Surprise
Anger	93.2	1.3	0.2	1.6	2.7	1
Disgust	1.8	88.4	3.2	1.5	1.4	3.7
Fear	1.3	3.2	81.4	4.3	1.1	8.7
Happiness	1.1	8.7	3.8	81.5	1.3	3.6
Sad	0.2	0.3	0.5	0.3	98.5	0.2
Surprise	1.2	8.3	3.9	5.1	1.1	80.4

(b) without augmenting

Actual \ Predicted	Anger	Disgust	Fear	Happiness	Sad	Surprise
Anger	84.2	3.2	4.3	2.3	3.9	2.1
Disgust	3.5	82.2	5.4	2.1	2.2	4.6
Fear	3.1	4.7	73	5.9	3.1	10.2
Happiness	2.6	10.6	5.3	73.6	2.7	5.2
Sad	1.5	2.4	2.1	1.4	90.5	2.1
Surprise	3.7	11.2	4.7	6.8	3.1	70.5

Fig. 7. Confusion matrix of the system using the eNTERFACE'05 database. The numbers represent accuracies (%). The diagonal dark-shadowed numbers are the correct recognition accuracies of individual emotions, while the light-shadowed numbers are the confused accuracies in the range between 5% and 50%.

Table 4
2D CNN architecture details.

Layer	Dimension
1. First convolution layer	7 × 7 (64 filters)
1. Max pooling	3 × 3
2. Second convolution layer	7 × 7 (128 filters)
2. Average pooling	3 × 3
3. Third convolution layer	3 × 3 (256 filters)
3. Average pooling	3 × 3
4. Fourth convolution layer	3 × 3 (512 filters)
5. Fully connected (FC) layer	1 × 1 × 4096 (two hidden layers)

cessing. Once the key frame is selected, the frame is converted into a gray-scale image. The mean normalization is applied to the image. We also calculate the LBP image and the IDP image from the gray-scale image. Therefore, we obtain three images (mean-normalized gray-scale, LBP, and IDP) per the key frame.

After detecting the key frame, the window is shifted by 4 frames, and another key frame is selected. The process is repeated until the end of the video segment. In every 2.02 s of a video segment, 16 key frames are selected for the CNN. The images from the key frames are sampled to 227 × 227.

3.3. CNN framework

The deep CNN is a very good learning technique of signals because it learns local and spatial textures of the signals by applying convolution and nonlinearity operations [52]. The deep CNN represents higher-level features as a blend of lower-level features. There are many models of the deep CNN in the literature, each of them is good in some sense.

In our proposed system, the CNNs for the speech signal and the video signal are different, for the speech signal, we use a 2D CNN, while for the video signal, we use a 3D CNN.

3.3.1. 2D CNN for speech signal

In the proposed emotion recognition system, we have developed a 2D CNN architecture shown in Fig. 4 for speech signals. There are four convolution layers and three pooling layers. The last layer is a fully-connected neural network with two hidden layers. Table 4 shows this CNN architecture details. A softmax function is applied to the output of the fully-connected layer. The output of the softmax is then fed into a classifier (or the ELM-based fusion).

In the 2D CNN, there are 64 filters of size 7 × 7 in the first convolution layer, 128 filters of size 7 × 7 in the second convolution layer, and 256 filters of size 3 × 3 in the third convolution layer. The fourth convolution layer has 512 filters of size 3 × 3. The size of filters is chosen to maintain a good balance between phone co-articulatory effect and long vowel phone. The stride in all the cases is 2.

The convolved images are normalized by using an exponential linear unit (ELU) as follows (Eq. (1)):

$$y_{i,j,k} = \begin{cases} x_{i,j,k}, & x_{i,j,k} > 0 \\ e^{x_{i,j,k}} - 1, & x_{i,j,k} \le 0 \end{cases} \tag{1}$$

In the proposed architecture, a max pooling is used in the first pooling layer, while an average pooling is used in the next two pooling layers. The pooling is obtained in every 2 × 2, with a stride of 2.

In the fully-connected network, there are 4096 neurons in each hidden layer. The final output layer is followed by a softmax function to provide a probability distribution of the output values. All the weights in the architecture were initialized by using a random function. A dropout with 50% probability is used at the beginning.

3.3.2. 3D CNN for video signal

For the 3D CNN we have adopted a pre-trained model as described in [53]. This 3D CNN model was originally developed for sports action

Table 5
3D CNN architecture details.

Layer	Dimension
1.C. First convolution layer (Conv1a)	3 × 3 × 3 (64 filters)
1.P. Max pooling	1 × 2 × 2
2.C. Second convolution layer (Conv2a)	3 × 3 × 3 (128 filters)
2.P. Max pooling	2 × 2 × 2
3.C. Third convolution layer (Conv3a)	3 × 3 × 3 (256 filters)
4.C. Fourth convolution layer (Conv3b)	3 × 3 × 3 (256 filters)
3.P. Max pooling	2 × 2 × 2
5.C. Fifth convolution layer (Conv4a)	3 × 3 × 3 (512 filters)
6.C. Sixth convolution layer (Conv4b)	3 × 3 × 3 (512 filters)
4.P. Max pooling	2 × 2 × 2
7.C. Seventh convolution layer (Conv5a)	3 × 3 × 3 (512 filters)
8.C. Eighth convolution layer (Conv5b)	3 × 3 × 3 (512 filters)
5.P. Max pooling	2 × 2 × 2
Fully connected layer (fc6)	1 × 1 × 4096
Fully connected layer (fc7)	1 × 1 × 4096

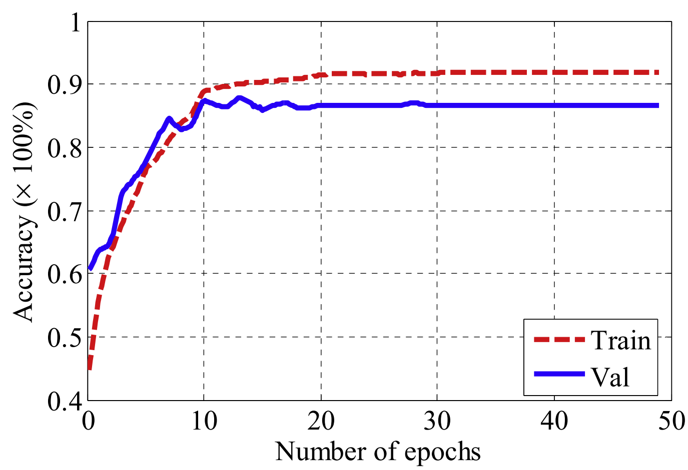

Fig. 8. Training and validation accuracy using the eNTERFACE database with augmentation.

recognition purpose. Later, the model was utilized in many video processing applications including emotion recognition from the video [46]. The structure of the 3D CNN model is shown in Table 5. There are eight convolution layers and five max-pooling layers. At the end, there are two fully-connected layers, each having 4096 neurons. A softmax layer follows the fully-connected layers. The stride of the filters is one. The input to the model is 16 key frames (RGB) resized to 227 × 227.

The output of the 3D convolution can be formulated as follows (Eq. (2)):

$$o_{i',j',k'} = \sum_{i,j,k} \omega_{i,j,k,k'} x_{i+i',j+j',k}, x : \text{input}, \omega : \text{weight}, \quad k : \#\text{offrames}, k' : \#\text{offilters} \tag{2}$$

To use the 3D CNN pre-trained model, first, we use all the weights of the convolution layers and the pooling layers from the model in [54]. Then, we replace the softmax layer to the number of emotion classes that we have in our system. After that, we fine-tune the model using this new softmax layer and update all the weights using a backpropagation algorithm.

3.4. ELM-based fusion

The ELM is based on a single hidden layer feed-forward network (SHLFN), which was introduced in [55]. There are some advantages of the ELM over the conventional CNN, such as fast learning, no need for weight adjustment during training, and no overfitting. In the proposed emotion recognition system, we used two ELMs successively for fusion

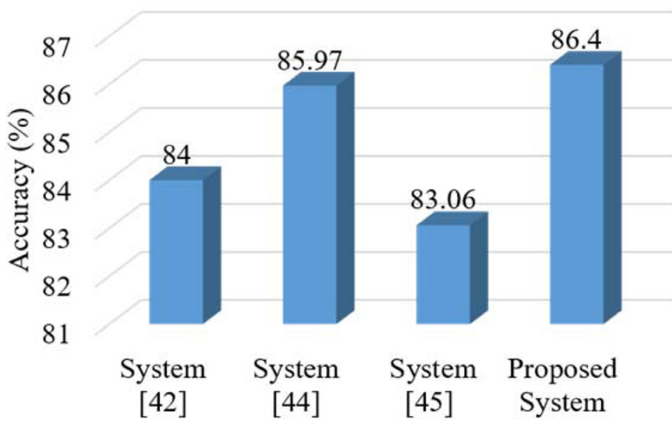

Fig. 9. Accuracy comparison between various systems.

of scores from the two modalities (see Fig. 5). In the proposed approach, the outputs of the fully-connected networks except for the final output layer (softmax) are the inputs to the first ELM. The number of nodes at the hidden layer of the ELM corresponds to 50 times the number of classes to provide the sparsity of the network. The first ELM (ELM-1) is trained according to the gender (two classes), while the second ELM (ELM-2) is trained to the emotions based on gender. As there are two output classes in the ELM-1, the number of hidden layer neuron is 100. Once the ELM-1 is trained, we remove the output layer of this ELM and make the trained hidden layer of the EML-1 as the input to the ELM-2. If there are five emotion classes, there are 250 hidden layer neurons in the ELM-2. The output scores are converted into probabilities using the softmax function. These output probabilities are fed into the SVM-based classifier.

If there are L number of hidden nodes in the ELM, and $\boldsymbol{\varphi}_{\mathrm{q}}$ (.) is the activation function, w_{q} is the input weight, σ_{q} is the bias of qth hidden node, α_{q} is the output weight, we get the output function as follows (Eq. (3)).

$$y_L(\mathbf{x}) = \sum_q \alpha_q \boldsymbol{\varphi}_q\left(w_q \circ \mathbf{x} + \sigma_q\right); q \in \{1, L\} \tag{3}$$

The optimum output weights are calculated using the following equation (Eq. (4)), where, P is the number of training samples.

$$\hat{\alpha} = \begin{cases} \left(\mathbf{M}^T\mathbf{M} + \frac{\mathbf{I}}{\varepsilon}\right)^{-1}\mathbf{M}^T\mathbf{N}, & P > L \\ \mathbf{M}^T\left(\mathbf{M}\mathbf{M}^T + \frac{\mathbf{I}}{\varepsilon}\right)^{-1}\mathbf{N}, & P \le L \end{cases} \tag{4}$$

In Eq. (4), $\mathbf{M}$ represents the output matrix $[\varphi(x_1), \varphi(x_2), ..., \varphi(x_P)]^T$, $\mathbf{I}$ is the identity matrix, and ε is the regularization coefficient and $\varepsilon > 0$. The value of ε was empirically set to 1 during our experiments. A Gaussian kernel is used as an activation function. The kernel parameter was set to 8, which gave the best result among {1,…,10}.

The two layers of the ELM bring a nonlinearity to the fusion in a way, which is fast in calculation but deep in nature. It can be noted that fusion based on the deep network already exists in the literature [46]; however, this type of fusion is computationally expensive, while our proposed one is computationally less demanding. The two-stage ELM inherently does the emotion recognition based on gender, and thereby improves the accuracy. It has been shown in the literature that the gender-based emotion recognition performs better than the gender-independent emotion recognition [56].

Other types of fusion that we considered:

We investigated other types of fusion in the experiments. These fusions include two decision-level fusions: 'max' and 'product' [57], and one score-level fusion: Bayesian sum rule [43]. In the decision-level fusion and the score-level fusion, two separate SVM classifiers, one for the speech modality and the other for the video modality, are used after the softmax layers of the CNNs.

3.5. SVM-based classifier

The probability distribution of the outputs of the ELM fusion is the input to the SVM. The SVM projects the input dimension to a higher dimension so that the samples of two classes are separable by a linear plane. The projection if often done using a kernel; we evaluate a polynomial kernel and a radial basis function (RBF) kernel separately, and the RBF kernel performed better in the experiments. The optimization parameter of the SVM was set to 1 and the kernel parameter was 1.5. We adopt a one-vs.-the rest approach to the SVM classifier.

It can be noted that the SVM is used as the classifier of the system, while the CNN models are used to extract features from the speech signal and the video signal, and the ELMs are used to fuse the features. The SVM is a powerful binary classifier, where the input data are projected to a high dimensional space by a kernel function so that the data of two classes are separated by a hyperplane. The objective is to find an optimal hyperplane that has maximum separation from the support vectors. We use the SVM in our system to exploit its powerful capability to classify different classes of data.

4. Experiments

This section presents a description of databases used in the experiments, some experimental setups, results, and discussion.

4.1. Data and setup

The proposed method of emotion recognition is evaluated using a Big data of emotion. The database was created using bimodal inputs: speech and video. 50 university-level male and female students were recruited for the database. They were trained to mimic different emotional expressions, namely, happy, sad, and normal. The emotions were both facial and spoken. The training for each emotion lasted for five minutes. During actual recording, we used a smartphone iPhone 6s. The recording was taken place in a single office environment. There were eight sessions for each emotion recording. Each session lasted for 15 min per participant. We selected a fixed sentence and some expressive sounds like /ah/, /uh/, and /ih/ to speak by the participants to express an emotion. The speech data amounted approximately 110 GB and the video data approximately 220 GB. The data are partitioned into three subsets: training, validation, and testing. The training, validation, and testing subsets accounted for 70%, 5%, and 25% of the total data.

To evaluate the proposed system on a publicly available database, we used the eNTERFACE' 05 audio-visual emotion database [21]. There are six emotions in the audio-visual signals; the emotions are anger, disgust, fear, happiness, sad and surprise. The speech signals are from read sentences posing different emotions, and the video signals are face videos posing the emotions. The faces are frontal. The average length of the video per subject per emotion per sentence is around three seconds. There are 42 subjects and six different sentences.

The amount of data in the eNTERFACE database is much less compared to the Big Data. Hence, we used five-fold cross-validation approach in the experiments. We also investigated the performance of the proposed system with and without augmenting the eNTERFACE database. In case of augmenting, the face images are rotated at various angles (5°, 15°, 25°, and 35°), and white Gaussian noise was added to the speech signal at the signal to noise ratio (SNR) = 30 dB, 20 dB, 15 dB, and 10 dB.

The training parameters of the CNN models were as follows: learned with a stochastic gradient descent with a group size of 100 samples, a learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.00005. A Gaussian distribution with zero mean and 0.01 standard deviation was utilized to initialize the weights in the final layer. It is already mentioned before that the other layers' weights were taken from the pre-trained model. There were 10,000 iterations during the training.

(a) Using RBF kernel in SVM

Actual \ Predicted	Happy	Pain	Normal
Happy	99.82	0.03	0.15
Pain	0.03	99.91	0.06
Normal	0.01	0.02	99.97

(b) Using Polynomial kernel in SVM

Actual \ Predicted	Happy	Pain	Normal
Happy	98.14	0.61	1.25
Pain	0.25	98.43	1.32
Normal	0.42	0.63	98.95

Fig. 10. Confusion matrix of the system using Big Data.

50% dropout was used in the last two fully-connected layers to lessen overfitting.

4.2. Experimental results and discussion

Fig. 6 shows the accuracies of the proposed system using different fusion strategies. There are four types of fusions: 'max', 'product', Bayesian sum rule, and the ELM. Using the Big Data, the highest accuracy of 99.9% was obtained using the ELM fusion. The least accuracy (91.3%) was with the 'max' fusion. Using the eNTERFACE database, the maximum accuracy (86.4%) was again with the ELM fusion; this accuracy was achieved with augmentation. From these results, we easily see that the ELM has a great potential to fuse information from various modalities. The gap of accuracies between using Big Data and the eNTERFACE database can be attributed to the fact that in Big Data we have only a fixed sentence and some short phrases, while in the eNTERFACE database we have six different sentences. Therefore, the accuracies using the eNTERFACE database were sentence-independent. Also, the number of emotions in Big Data is less and somewhat clearly distinguishable. In addition to this, the system is trained well using the Big Data rather than the limited data in the eNTERFACE database.

Fig. 7 shows the confusion matrix of the proposed system using the eNTERFACE database with and without augmenting. The results are with the ELM fusion. Clearly, we see that the augmentation improved the accuracy of the system by a significant amount. Fig. 8 shows the training and the validation accuracies with the number of epochs using augmentation of the eNTERFACE database. As we can see from the figure, the proposed system has higher accuracy using the validation dataset than that using the training dataset at the initial epochs. This overfitting phenomenon occurs because the number of samples in the eNTERFACE database is limited. Fig. 9 shows a comparison of accuracies obtained by various systems using this database. The performance of the proposed system is slightly better than that of the system in [46].

Fig. 10 shows the confusion matrices of the proposed system using Big Data. As mentioned earlier, we investigated two types of kernels in the SVM. From the confusion matrices, we find that the RBF kernel performed better in the system. 'Normal' emotion had as high as 99.97% accuracy. All these accuracies were with the ELM based fusion. Fig. 11 shows the training and the validation accuracies of the system (with the RBF kernel in the SVM).

We compared the performance of the proposed system with that of another system described in [58] using the same Big Data. In [58], the LBP features for speech and IDP features for face images were used together with the SVM based classifier. The score-level fusion was utilized.

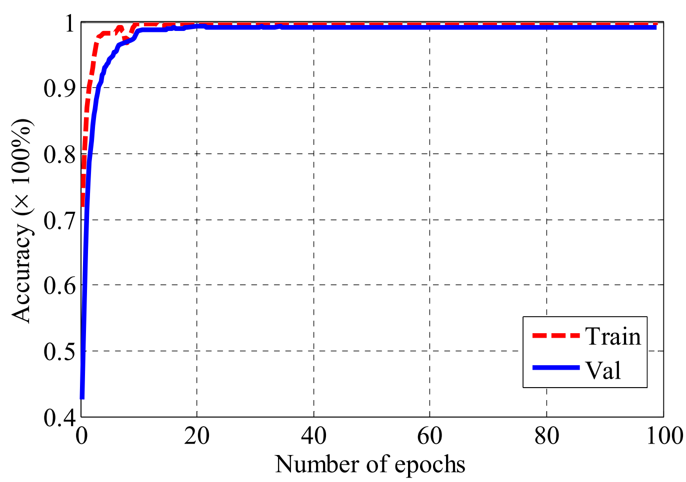

Fig. 11. Training and validation accuracy of the proposed system using Big Data.

The system in [58] achieved 99.8% accuracy, while our proposed system had 99.9% accuracy.

5. Conclusion

An audio-visual emotion recognition system was proposed. The 2D CNN for the speech signal and the 3D CNN for the video signal were used. Different fusion strategies including the proposed ELM-based fusion were investigated. The proposed system was evaluated using Big Data of emotion and the eNTERFACE database. In both the databases, the proposed system outperformed other similar systems. The ELM-based fusion performed better than the classifiers' combination. One of the reasons for this good performance is that the ELMs add a high degree of non-linearity in the features' fusion. The proposed system can be extended to be a noise-robust system by using a sophisticated processing of speech signals instead of using the conventional MFCC features, and by using some noise-removal techniques in key frames of video signals. In case of the failure to capture either speech or face, an intelligent weighting scheme in the fusion can be adopted to the proposed system for a seamless execution.

The proposed system can be integrated in to any emotion-aware intelligent systems for a better service to the users or customers

[59,60,54]. Using edge technology, the weights of the deep network parameters can easily be stored for a fast processing [61].

In a future study, we will evaluate the proposed system in an edge-and-cloud computing framework. We also want to investigate other deep architectures to improve the performance of the system using the eNTERFACE database and emotion in the wild challenge databases.

Acknowledgement

The authors are thankful to the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia for funding this research through the Research Group Project no. RGP-228.

References

[1] M. Chen, Y. Zhang, M. Qiu, N. Guizani, Y. Hao, SPHA: smart personal health advisor based on deep analytics, IEEE Commun. Mag. 56 (3) (2018) 164–169.
[2] F. Doctor, C. Karyotis, R. Iqbal, A. James, An intelligent framework for emotion aware e-healthcare support systems, in: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Athens, 2016, pp. 1–8. 2016.
[3] K. Lin, F. Xia, W. Wang, D. Tian, J. Song, System design for big data application in emotion-aware healthcare, IEEE Access 4 (2016) 6901–6909.
[4] J.M. Harley, S.P. Lajoie, C. Frasson, N.C. Hall, An integrated emotion-aware framework for intelligent tutoring systems, in: C. Conati, N. Heffernan, A. Mitrovic, M. Verdejo (Eds.), Artificial Intelligence in Education. AIED 2015. Lecture Notes in Computer Science, 9112, Springer, Cham, 2015.
[5] S.K. D'Mello, A.C. Graesser, Feeling, thinking, and computing with affect-aware learning technologies, in: R.A. Calvo, S.K. D'Mello, J. Gratch, A. Kappas (Eds.), Handbook of Affective Computing, Oxford University Press, 2015, pp. 419–434.
[6] K. Meehan, T. Lunney, K. Curran, A. McCaughey, Context-aware intelligent recommendation system for tourism, in: Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), „ San DiegoCA, San DiegoCA, 2013, pp. 328–331. 2013.
[7] Y. Zhang, M. Chen, D. Huang, D. Wu, Y. Li, iDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization, Future Gen. Comput. Sys. 66 (2017) 30–35.
[8] B. Guthier, R. Alharthi, R. Abaalkhail, A. El Saddik, Detection and visualization of emotions in an affect-aware city, in: Proceedings of the First International Workshop on Emerging Multimedia Applications and Services for Smart Cities (EMASC '14). ACM, New York, NY, USA, 2014, pp. 23–28.
[9] M. Chen, J. Yang, X. Zhu, X. Wang, M. Liu, J. Song, Smart Home 2.0: innovative smart home system powered by botanical IoT and emotion detection, Mob. Netw. Appl. (2017), doi:10.1007/s1103.
[10] M.S. Hossain, G. Muhammad, M. AL Qurishi, Verifying the Images Authenticity in Cognitive Internet of Things (CIoT)-Oriented Cyber Physical System, Mobile Netw. Appl. 23 (2) (2018) 239–250.
[11] M.L.R. Menezes, A. Samara, L. Galway, et al., Towards emotion recognition for virtual environments: an evaluation of EEG features on benchmark dataset, Pers. Ubiquitous Comput. (2017), doi:10.1007/s00779-017-1072-7.
[12] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, M. Pietikäinen, Multi-modal emotion analysis from facial expressions and electroencephalogram, Comput. Vis. Image Underst. 147 (2016) 114–124.
[13] M. Valstar, J. Gratch, B. Schuller, et al., AVEC 2016: depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the Sixth International Workshop on Audio/Visual Emotion Challenge (AVEC '16). ACM, New York, NY, USA, 2016, pp. 3–10.
[14] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44.
[15] M. Chen, Y. Hao, K. Hwang, L. Wang, Disease prediction by machine learning over big healthcare data, IEEE Access 5 (1) (2017) 8869–8879.
[16] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," Proc. INTERSPEECH 2014, pp. 223–227, Singapore, 14–18 September 2014.
[17] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.
[18] C.K. Yogesh, M. Hariharan, R. Ngadiran, A.H. Adom, S. Yaacob, C. Berkai, K. Polat, A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal, Expert Syst. Appl. 69 (2017) 149–158.
[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of german emotional speech, in: Proceedings of the INTERSPEECH, Lisbon, Portugal, 2005.
[20] J. Deng, Z. Zhang, E. Marchi, B. Schuller, Sparse autoencoder-based feature transfer learning for speech emotion recognition, in: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, 2013, pp. 511–516. 2013.
[21] O. Martin, I. Kotsia, B. Macq, I. Pitas, The enterface'05 audiovisual emotion database, IEEE Workshop Multimed. Database Manag. (2006).
[22] J.B. Alonso, J. Cabrera, M. Medina, C.M. Travieso, New approach in quantification of emotional intensity from the speech signal: Emotional temperature, Exp. Syst. Appl. 42 (2015) 9554–9564.
[23] M.S. Hossain, G. Muhammad, Cloud-based collaborative media service framework for health-care, Int. J. Distrib. Sensor Netw. (2014) 11 Article ID 858712, February 2014.
[24] E.M. Schmidt, Y.E. Kim, Learning emotion-based acoustic features with deep belief networks, in: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, 2011, pp. 65–68. 2011.
[25] W. Zhang, D. Zhao, Z. Chai, L.T. Yang, X. Liu, F. Gong, S. Yang, Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services, Softw. Pract. Exper. 47 (2017) 1127–1138, doi:10.1002/spe.2487.
[26] Haytham M. Fayek, Margaret Lech, Lawrence Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Netw. 92 (2017) 60–68.
[27] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.
[28] J. Tao, F. Liu, M. Zhang, H.B. Jia, Design of speech corpus for mandarin text to speech, in: Proceedings of the Blizzard Challenge 2008 Workshop, 2008.
[29] E. Trentin, S. Scherer, F. Schwenker, Emotion recognition from speech signals via a probabilistic echo-state network, Pattern Recognit. Lett. 66 (2015) 4–12.
[30] Niu, Yafeng; Zou, Dongsheng; Niu, Yadong; He, Zhongshi; Tan, Hua, "A breakthrough in speech emotion recognition using deep retinal convolution neural networks," eprint arXiv:1707.09917, 2017.
[31] H.-W. Ng, V.D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15), New York, NY, USA, 2015, pp. 443–449.
[32] G. Muhammad, M. Alsulaiman, S.U. Amin, A. Ghoneim, M.F. Alhamid, A facial-expression monitoring system for improved healthcare in smart cities, IEEE Access 5 (1) (2017) 10871–10881.
[33] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings of the IEEE International Conference on Automation Face Gesture Recognition, 2000, pp. 46–53.
[34] G. Muhammad, M.F. Alhamid, User emotion recognition from a larger pool of social network data using active learning, Multimedia Tools Appl. 76 (8) (2017) 10881–10892.
[35] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A.M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649.
[36] M.S. Hossain, G. Muhammad, An emotion recognition system for mobile applications, IEEE Access 5 (2017) 2281–2287.
[37] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1–10. 2016.
[38] H. Ding, S.K. Zhou, R. Chellappa, FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition, in: Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 118–126. 2017.
[39] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, D. Tao, Deep neural networks with relativity learning for facial expression recognition, in: Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Seattle, WA, 2016, pp. 1–6. 2016.
[40] N.B. Kar, K.S. Babu, S.K. Jena, Face expression recognition using histograms of oriented gradients with reduced features, in: Proceedings of the International Conference Computer Vision and Image Processing (CVIP), 2, 2016, pp. 209–219.
[41] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 3687–3691. 2013.
[42] S.E. Kahou, X. Bouthillier, P. Lamblin, et al., EmoNets: multimodal deep learning approaches for emotion recognition in video, J. Multimodal User Interf. 10 (2) (2016) 99–111 June.
[43] M.S. Hossain, G. Muhammad, B. Song, M. Hassan, A. Alelaiwi, A. Alamri, Audio-visual emotion-aware cloud gaming framework, IEEE Trans. Circuits Syst. Video Technol. 25 (12) (2015) 2105–2118 December.
[44] M.S. Hossain, G. Muhammad, M.F. Alhamid, B. Song, K. Al-Mutib, Audio-visual emotion recognition using big data towards 5G, Mobile Netw. Appl. 221 (5) (2016) 753–763 October.
[45] H. Ranganathan, S. Chakraborty, S. Panchanathan, Multimodal emotion recognition using deep learning architectures, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1–9. 2016.
[46] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio-visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 99 (2017) 1.
[47] M.S. Hossain, G. Muhammad, Audio-visual emotion recognition using multi-directional regression and ridgelet transform, J. Multimodal User Interf. 10 (4) (2016) 325–333.
[48] M. Bejani, D. Gharavian, N. Charkari, Audiovisual emotion recognition using ANOVA feature selection method and multiclassifier, Neural Computing Appl. 24 (2) (2014) 399–412.
[49] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, H. Sahli, Audio visual emotion recognition based on triple-stream dynamic bayesian network models, in: D'Mello (Ed.), ACII,Part I, LNCS 6974, 2011, pp. 609–618.
[50] H. Kaya, F. Gürpınar, A.A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image Vision Comput. 65 (2017) 66–75.
[51] Paul Viola, J.Jones Michael, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 2001, pp. 511–518.
[52] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4489–4497.

[54] M. Chen, P. Zhou, G. Fortino, Emotion communication system, IEEE Access 5 (2017) 326–337.
[55] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501.
[56] I.M.A. Shahin, Gender-dependent emotion recognition based on HMMs and SPHMMs, Int. J. Speech Technol. 16 (2) (2013) 133–141.
[57] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.
[58] M.S. Hossain, G. Muhammad, Emotion-aware connected healthcare big data towards 5G, IEEE Internet Things J. 5 (4) (2018) 2399–2406 August, doi:10.1109/JIOT.2017.2772959.
[59] M. Chen, Y. Tian, G. Fortino, J. Zhang, I. Humar, Cognitive internet of vehicles, Comput. Commun. 120 (2018) 58–70 May 2018, doi:10.1016/j.comcom.2018.02.006.
[60] M. Chen, F. Herrera, K. Hwang, Human-centered computing with cognitive intelligence on clouds, IEEE Access 6 (2018) 19774–19783, doi:10.1109/ACCESS.2018.2791469.
[61] M. Chen, Y. Qian, Y. Hao, Y. Li, J. Song, Data-driven computing and caching in 5G networks: architecture and delay analysis, IEEE Wireless Commun. 25 (1) (2018) 70–75 Feb.