



Deep neural network-based fusion model for emotion recognition using visual data

Luu-Ngoc Do¹ · Hyung-Jeong Yang² · Hai-Duong Nguyen² · Soo-Hyung Kim² · Guee-Sang Lee² · In-Seop Na²

Accepted: 13 February 2021 / Published online: 10 March 2021
© The Author(s) 2021

Abstract

In this study, we present a fusion model for emotion recognition based on visual data. The proposed model uses video information as its input and generates emotion labels for each video sample. Based on the video data, we first choose the most significant face regions with the use of a face detection and selection step. Subsequently, we employ three CNN-based architectures to extract the high-level features of the face image sequence. Furthermore, we adjusted one additional module for each CNN-based architecture to capture the sequential information of the entire video dataset. The combination of the three CNN-based models in a late-fusion-based approach yields a competitive result when compared to the baseline approach while using two public datasets: AFEW 2016 and SAVEE.

Keywords Emotion recognition in the wild · Convolutional neural network · Long-short term memory network · Fusion model

✉ Hyung-Jeong Yang
hjyang@jnu.ac.kr

Luu-Ngoc Do
doluungoc@gmail.com

Hai-Duong Nguyen
duong.talkin@gmail.com

Soo-Hyung Kim
shkim@jnu.ac.kr

Guee-Sang Lee
gslee@jnu.ac.kr

In-Seop Na
ypencil@daum.net

¹ Department of Radiology, Chonnam National University, Gwangju, Korea

² Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea

1 Introduction

Emotions have an essential influence not only on the interactions among human beings but also on the human–computer interactions. Because that the emotional state of a person may affect his/her concentration, task solving, and decision-making skills, the vision of affective computing is to enable systems to recognize human emotions and influence them to enhance productivity and effectiveness when working with computers [1]. Thus, an automatic emotion recognition model contains several attentions and a variety of applications. The computer vision and psychological research have been combined in several applications such as monitoring the conditions of the driver (e.g., state of fatigue) and monitoring signs of attention to enhance driver's safety [2], detection of depression in individuals, and diagnosis of developmental disorders of children by monitoring their facial expressions and gaze during social interactions [3]. Emotion recognition from video data has also been revolutionizing marketing strategies pertaining to the quantification of advertisement preferences by automatically gathering the human expressions information exposed from specific contexts [4]. Moreover, capturing the emotions automatically can help blind people to understand facial expressions, help robots to interact smartly with people for better service. Emotion recognition in conversation also can be used to extract a huge amount of opinions between participants from massive conversational data in social networks. The task of emotion recognition is particularly difficult for two reasons: (1) a large database containing training images does not currently exist; and (2) the definition of emotional expressions varies in different people, and thus a unified definition cannot be provided. However, numerous studies were conducted in the last two decades to gradually overcome the challenges in automatic emotion recognition [5–7].

Emotions can be recognized based on a variety of approaches, such as facial expressions, speech, and body gestures. Generally, facial information is considered as the most common characteristic for recognizing emotion as it does not demonstrate significant challenges such as recording sound in a noisy environment or analyzing body gestures while dealing with problems of occlusion. However, facial expression recognition is still a challenging concern owing to the variety of head poses and background settings. The challenge in facial expression recognition is to effectively locate and understand facial regions-of-interest. In the past few years, these tasks were performed by traditional computer vision methods, such as landmark detection and object modeling. However, it was assumed that emotion recognition was performed in a controlled environment such as indoor conditions with frontal view of the face. This implies that the face and the background were not complicated for identification. Analyzing facial expression in the wild requires the system to handle various aspects of an unconstrained scenario such as dynamic illumination, occlusion, and head poses. Thus, facial expression recognition in the wild remains as a challenging problem and has been attracting increasing attention.

In this study, we present an ensemble of deep neural network-based models for emotion recognition in the wild using visual data. The proposed fusion

model focuses on the analyses of facial information to estimate seven emotional categories, which are anger, disgust, fear, happiness, sadness, surprise, and neutral state. Three face representations were developed based on the convolutional neural network (CNN) [8]. The first model uses a multilevel CNN (MLCNN) to extract facial features from each video frame and 3DCNN [9] for analyzing temporal information. The second model is a combination of the VGG-FACE [10] and the long short-term memory (LSTM) modules [11]. The third model fine-tuned the Xception network [12] and encoded the features statistically. The fusion of three models was investigated using four fusion methods such as features fusion, average scoring, max voting, and weight fusion to identify the best combination. The major contribution of this study is the construction and training of deep neural networks that perform well in emotion recognition tasks using visual data in both unconstrained scenario, where the data are captured in the wild, and constrained scenario, for the data captured in a controlled environment such as indoor conditions.

The paper is organized as follows. In Sect. 2, the recent state-of-the-art methods are reviewed. The proposed method is presented in Sect. 3. In Sect. 4, the performance of the proposed method on public datasets is evaluated. Section 5 outlines the conclusion and future improvement.

2 Related works

Most applications for emotion recognition examine sequences of images of facial expressions. There were several approaches to resolve this problem, including the use of pyramidal histograms of gradients (PHOG) [13], action-unit-aware facial features [5], and boosted LBP descriptors [14]. These handcrafted features were then used to train various classifiers. Examples include spatial classifications using support vector machines (SVMs) and temporal classifications using dynamic Bayesian networks [15]. Kaya et al. [16] combined several traditional features and employed least-squares-based learners. Traditional computer vision approaches with handcrafted features yielded satisfactory results while analyzing data acquired in controlled environments, where the eyes, nose, and mouth were identified in an image using simple measurements. However, emotions expressed in the wild are sensitive to various aspects of an unconstrained scenario, such as head poses and dynamic illumination, which are extremely complicated to analyse using handcrafted features. Recently, deep neural networks were considered to solve this problem in an unconstrained sampling scenario.

The initial framework of deep learning, CNN, was developed by LeCun in the 1990s for handwritten digit recognition [8]. CNN utilizes a sequence of convolution layers to automatically extract features from a single image. The classification result is produced by fully connected (FC) layers and Softmax classifier. The whole network is trained for updating the parameters by the back-propagation algorithm. In 2012, Krizhevsky et al. proposed Alexnet, which extends CNN to a deeper structure for the task of 1000-class image classification problem using ImageNet data [17]. It significantly outperformed the runner-up and led to a breakthrough in the ImageNet

Challenge. Adding more layers with many stacks of convolution potentially improve the performance of classification problem. Simonyan et al. proposed VGGNet and carried out an analysis of how CNN's depth improves classification accuracy [18]. Szegedy et al. introduced the Inception module, which includes 1×1 , 3×3 , and 5×5 filters in the same layer [19]. Inception's architecture can analyse images with multiple convolution filters parallelly and outperformed the other models in ImageNet Challenge 2014. Francois proposed Xception with the use of depth-wise separable convolution instead of normal convolution [12]. The depth-wise separable convolution not only deals with the spatial dimensions but also with the depth dimension as well. Even though the invention of the Inception module led to significant improvements, it was noticed that small filter sizes such as 3×3 performed well for most image classification problems. Therefore, recent work focuses on improving performance by adding layers to the network. However, by increasing the number of layers, information is more difficult to flow, which is known as degradation problem or gradient vanishing. ResNet, the winner model of ImageNet Challenge 2015, outperformed previous ensemble models by using the Identity mapping [20]. This connection copies the learned shallower layer directly to the later.

Tang introduced a CNN that jointly learned with a linear SVM for facial expression recognition [21]. With the use of a simple CNN and SVM, instead of a Softmax classifier, this model outperformed the others and won the first place in the FER 2013 challenge [22]. Inspired by the success of GoogLeNet, Mollahosseini et al. proposed an architecture that contained four inception modules [23]. However, this did not result in a better performance in the FER 2013 challenge. In 2016, Zhou et al. proposed a multiscale CNN [24]. It consisted of three other networks with different input sizes. In addition, they used late-fusion to obtain the final classification results. By combining multiple CNNs and modifying the loss function, Yu et al. obtained a higher accuracy compared to previous approaches [25].

Emotion recognition in the wild (EmotiW) is the leading competition for facial expression recognition from videos. This grand challenge was initiated in 2013 and provided a benchmark for evaluating methods using "in the wild" data [26]. The participants are encouraged to surpass the baseline result, which is computed by LBP-TOP descriptor [9] and the SVR classification [26]. A major issue in this competition was the lack of training data. To address this issue, most competitors used transfer learning to take advantage of the available deep neural networks and proposed multimodal learning models considering the data types [7, 27]. In 2016, Fan et al. proposed a hybrid network that combines LSTM, 3DCNN, and an audio module based on SVM, and they won the challenge [27].

In this study, we use only visual data for the task of emotion recognition. We propose an ensemble of three deep neural networks for analyzing the sequential facial expressions. The first network, called MLCNN, is inspired by ResNet [20] with the use of a mapping from shallow layers to deeper layers. The second network utilizes VGG-FACE [10], which is trained by VGGNet for face data. The third network is fine-tuned from Xception [12]. Each of the three CNN-based models can capture the emotional features of the human face from a single video frame. A sequential model is adjusted to each CNN-based model separately to generate the feature of the whole video. The final emotion label of the video is decided by combining three models in

the fusion step. The proposed system is evaluated with both unconstrained scenario, where the data are captured in the wild, and constrained scenario, where the data are captured with indoor conditions.

3 Proposed method

Our system is the fusion of three different CNN-based face representations. The first face model utilized a MLCNN for extracting high-level features of face images. The temporal information was captured by a 3DCNN structure. The second face model consisted of VGG-FACE [10] and LSTM modules [11]. The third model fine-tuned the exit flow of Xception [12] and utilized a statistic encoding module for representing the sequential information of the videos. We also conducted our own dataset, called the Chonnam Emotion Videos (CEV) data, for training. The pipeline of the proposed system is shown in Fig. 1.

3.1 Face detection and selection

Although the entire set of video frames, including faces and background, can be fed into a deep CNN for feature extraction, face detection is considered to be an essential preprocessing step to achieve facial expression recognition. This step excludes the unnecessary background information and forces the CNNs to focus on only the human face region, which is known to contain the most crucial emotion expressions of the video frames. A face selection also is required in case two or more of the human faces can appear in one video frame, make it challenging to analyse the primary expressions. Our system is built based on the assumption that each video contains only one emotion expression from one human character. Hence, our face detection system consists of two

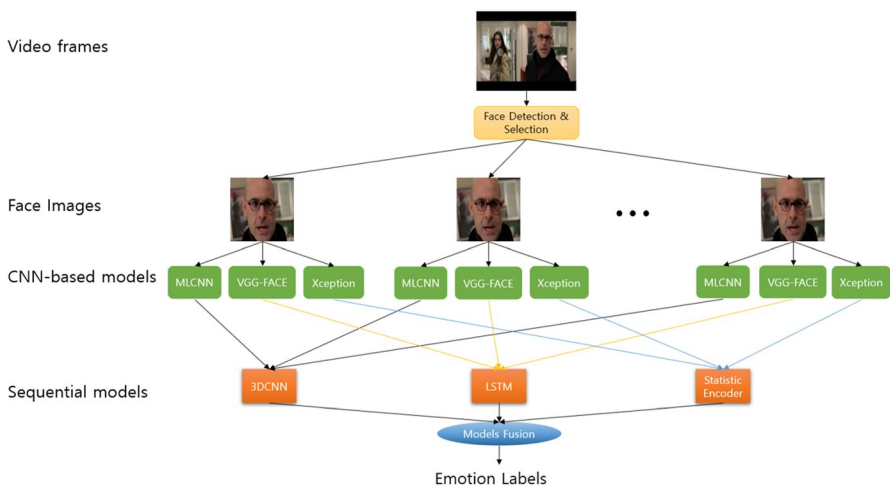


Fig. 1 Proposed system

components: (1) a face detector for locating all the faces in each video frame using the tiny face detector [28], and (2) our proposed clustering method for selecting the major set of faces as a video may include other face regions without any significant expressions. We assumed that $F_i = \{(x_{i,g}, y_{i,g}, w_{i,g}, h_{i,g}) \in \mathbb{N}^4\}$ was a set containing face information of the i_{th} frame in a video, where $g, x_{i,g}, y_{i,g}, w_{i,g}$, and $h_{i,g}$, define the face clusters, face center coordinates, and face sizes, respectively. Initially, the set F_i must be empty, and g was equal to zero as no face groups were detected. If $f_j = (x_j, y_j, w_j, h_j)$ is a detected face region in the j_{th} frame, then from (1), we can measure the difference between this region and the latest detected object from each face cluster. This metric is defined by the changes in the location of the face and size,

$$d_j(g) = \frac{1}{2} \|(x_j - x_{j-1,g}, y_j - y_{j-1,g})\|_2 + \frac{1}{2} (|w_j - w_{j-1,g}| + |h_j - h_{j-1,g}|) \tag{1}$$

To reduce the sensitivity during clustering, we compared this measurement with a threshold t ; subsequently, we assigned the group index g_j for f_j based on the following formula,

$$g_j = \begin{cases} g' & \text{if } d_j(g') < t \\ \max\{g\} + 1 & \text{otherwise} \end{cases}, \tag{2}$$

where $g' = \underset{g}{\operatorname{argmin}}\{d_j(g)\}$. In our experiment, t was equal to 50. To select the most significant faces, we assumed G to be a set of face group indices. After calculating g_j for each detected face region, g_j will be added to G , that is, $G = G \cup \{g_j\}$. Given a set of faces in a video dataset $(g, x_k, y_k, w_k, h_k), g \in G$, we determined the face cluster g that was detected more often than the others in this video for further processing,

$$\bar{g} = \operatorname{mode}\{G\} \tag{3}$$

3.2 Multilevel CNN

The backbone of the MLCNN is a deep plain CNN with 18 weighted layers organized into five blocks (Fig. 2). Each block contains 2, 3, or 4 convolutional layers, followed

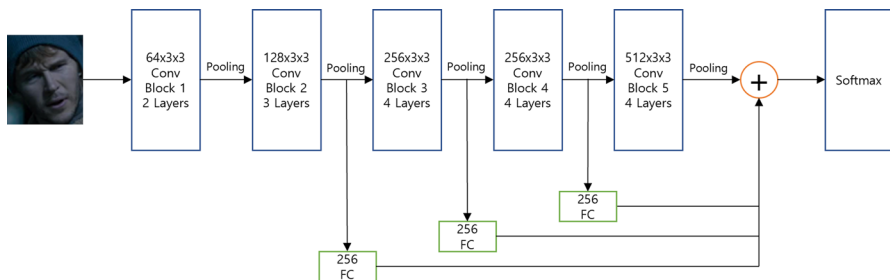


Fig. 2 Multilevel CNN

by max-pooling. The input of the network is a 48×48 face image, and the output of the bottleneck layer contains 512×1 filters, i.e., the fully connected layer contains only a seven-way Softmax classifier. It is to be noted that the bottleneck layer is the layer just before the final output layer of the network. Because we did not use any fully connected layers before the Softmax classifier, it is reasonable to analyse the contribution of the filter to the final classification. The network connection in our architecture is the vector concatenation operator defined as,

$$(x_1^1, x_2^1, \dots)^T \oplus (x_1^2, x_2^2, \dots)^T = (x_1^1, x_2^1, \dots, x_1^2, x_2^2, \dots)^T \tag{4}$$

where $(x_1^1, x_2^1, \dots)^T$ and $(x_1^2, x_2^2, \dots)^T$ are feature vectors from different network levels. When the input is an output of the convolutional layers, it needs to be vectorized before this operator is applied. After extracting features from each face image using MLCNN, we used a fully connected layer with 256 units to reduce their dimensions before feeding them into the 3DCNN. Subsequently, all these vectors were reshaped into a 16×16 array. A temporal model for video-based facial expression recognition containing four convolutional layers, two pooling layers, and a Softmax classifier was used to predict the emotion based on 32 randomly selected frames. The MLCNN was trained using the FER 2013 dataset [22], while the AFEW dataset [26] was used to train the 3DCNN.

3.3 VGG-FACE and LSTM

We used the pre-trained VGG-FACE [10] model for feature extraction. As shown in Fig. 3, the output of the final pooling layer of VGG-FACE was converted into sequences by turning each video into a 60-frame sequence. The sequential information can be captured by the traditional recurrent neural network (RNNs) following:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{5}$$

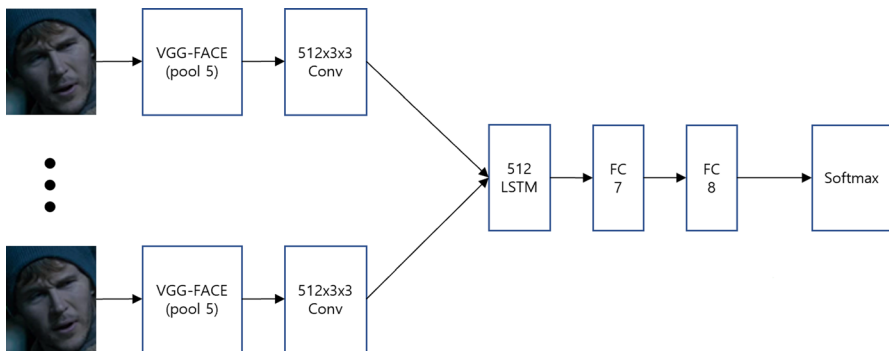


Fig. 3 Combined VGG-FACE and LSTM model

$$y_t = W_{hy}h_t$$

where x_t is the input tensor at time step or frame t , h_t is the hidden state, and y_t is the output at time t . However, for a long-term dynamic data like a 60-frames emotion video, the traditional RNNs is not efficient by vanishing and exploding the gradients through many layers. LSTM [11] proposed a solution by learning the confidence of hidden states: when to ignore the previous hidden states and when to update the hidden states given the new information, as shown in Eq. (6).

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$C_t = f \odot C_{t-1} + i \odot g \quad (6)$$

$$h_t = o \odot \tan h(c_t)$$

where c_t is the memory cell, i is the input gate, o is output gate, f is forget gate, g is input modulation gate, σ is the sigmoid function of x . This cell structure enables LSTM to decode a complex temporal dynamic sequence data.

We used a convolution layer with a kernel size of 3×3 , and a single, 512-wide LSTM layer, followed by two 512 dense layers (FC layers), with some dropout in between for training the VGG-FACE features with AFEW and CEV data. Videos that contained more than 60 frames were downsampled based on the random selection of a continuous sequence, which was composed of 60 frames. Videos which were shorter than 60 frames were repeatedly padded with the first frame into the beginning part of the sequence to form a new sequence with 60 frames. To avoid overfitting and to balance the training data among classes, we augmented the original samples by horizontal flipping, rotating at small angles, and Gaussian noise adjustment. The network is trained by Adam optimizer with the initial learning rate α is $1e-7$. Adam is robust and well-suited to sparse gradients by step size annealing as follow [29]:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \cdot g_t^2$$

$$mb_t = \frac{m_t}{1 - \beta_1^t} \quad (7)$$

$$vb_t = \frac{v_t}{1 - \beta_2^t}$$

$$\theta_t = \theta_{t-1} - \alpha mb_t / (\sqrt{vb_t} + \epsilon)$$

where θ is the update parameter, m_t is the first momentum, v_t is the second momentum, mb_t is the bias-corrected of the first momentum, vb_t is the bias-corrected of the second momentum, β_1 is 0.9, β_2 is 0.999, and ϵ is $1e-7$.

3.4 Xception and statistic encoder

The Xception's architecture is a replacement of the Inception module based on the use of depth-wise separable convolution layers [12]. We fine-tuned the exit flow of Xception with the use of a subset of AFEW and CEV data. We adjusted the fully connected layers to Xception, and then computed the 512-D features for each video frame. Using the set of feature vectors from each video, we converted it into a single feature vector that represented the entire video sequence with the use of a statistical encoder. The set of the 512-D features was encoded based on the computation and concatenation of the mean, variance, minimum, and maximum of the feature dimensions, thus resulting in a 2048-D feature vector. We normalized this feature and used it for classification. The other settings for this model are similar to the VGG-FACE model in Sect. 3.3.

3.5 Fusion models

We used four fusion methods to investigate the best combination of the three models. The first method was fusion at the feature level. The feature vectors from the three models were extracted and normalized using L2 normalization. They were concatenated to one feature vector for classification with the use of the MLP. The second method estimated the average of the scores from the three models. The third method considered voting of the maximum scores from the three models. The fourth method assigned a proper weight to each model according to their performance. The final score vector can then be formulated as follows,

$$S = w_1 S^1 + w_2 S^2 + w_3 S^3 \quad (8)$$

where S^1 , S^2 , and S^3 are the scores of Multilevel CNN, VGG-FACE+LSTM, and Xception+Statistic Encoder, respectively. Based on the performance of individual model shown in Table 2, the weight values are selected so that the ratios between w_1 , w_2 , w_3 are corresponding to the ratio of accuracy between the three models. The comparison between this strategy and the equal-weight strategy in Table 3 shows that assigning the weight value to each model according to their performance is an appropriate way to fuse the three CNN-based models.

4 Experimental results

The proposed model was evaluated using two visual emotion databases of the constrained and unconstrained scenario. The first database contains the videos of the AFEW 2016 data [26] from the Emotion Recognition in the Wild (EmotiW)



Fig. 4 AFEW data samples



Fig. 5 CEV data samples

challenge [16]. These videos, extracted from Hollywood movies, can be considered as data in the unconstrained scenario, which expresses emotion in the wild. The second database is the visual part of the Surrey audio-visual expressed emotion (SAVEE) database [30]. This data are captured with indoor conditions, where the camera is set up to focus on the human frontal face. SAVEE data can be considered as data in the constrained scenario. The baseline results of the AFEW 2016 and SAVEE data were surpassed by the proposed model.

4.1 Data description

In the AFEW 2016 dataset, the training and validation data consisted of short video clips (1–3 s) that were extracted from classic movies, while the testing data also included certain reality TV shows in addition to the movie clips. This data are more realistic and contain more challenges when compared to videos of facial actions that are deliberately produced and captured indoors. There were 773 training samples, 383 validation samples, and 653 testing samples in total. Each of the training and validation video clips was assigned to one label from a list of seven emotional states (anger, disgust, fear, happiness, sadness, surprise, and neutral). In contrast, the test video clip labels were not provided. Therefore, in this study, we used the validation set of AFEW 2016 for evaluating the proposed models.

We also utilized our CEV data for training. The CEV data contained more than 1200 short clips with seven emotional conditions similar to the AFEW data. Figures 4 and 5 illustrate samples of the AFEW and CEV datasets, respectively. The distribution of the AFEW training and validation data is shown in Table 1.

The SAVEE database contained entries recorded from four male subjects (identified as KL, JE, JK, DC) at the University of Surrey with ages in the range from 27 to 31 years. These data contained seven emotional categories similar to the AFEW data, namely anger, disgust, fear, happiness, sadness, surprise, and neutral. A total of 480 three-second videos were recorded. Each subject was associated with 120

Table 1 AFEW data distribution

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Training	133	74	81	150	117	74	144	773
Validation	64	40	46	63	61	46	63	383

videos: 30 videos demonstrated neutral emotional states, and the remaining videos presented each of the remaining six emotion categories (15 videos for each state). Most of the data captured the frontal view of the human face with indoor conditions, as shown in Fig. 6. We evaluated the proposed model with SAVEE data using both subject-dependent and subject-independent strategies.

4.2 Evaluation of AFEW 2016 data

We used the AFEW data allocated for training and CEV data for training our models. Four types of data augmentation techniques, namely horizontal flips, rotation left, rotation right, and Gaussian noise, were applied to prevent overfitting during training. More than ten thousand videos were used in the training session. The validation part of the AFEW data with 383 samples was used for the evaluation of the proposed models. Table 2 lists the recognition accuracies of every single model versus the baseline method. The baseline performance provided for the EmotiW challenge was based on the computation of the LBP-TOP descriptor [9] and the SVR classification [26]. The MLCNN combined with 3DCNN was the best single model that improved the accuracy by more than 10% when compared to the baseline results. Both the Xception-statistic encoder (STAT) model and the combined VGG-FACE and LSTM model yielded marginally higher accuracies when compared to the baseline results.

Table 3 lists the performance of the four fusion methods that was applied to the three proposed single models. The weight fusion achieved the best performance with an accuracy of 54.83%. The equal-weight fusion, where the same weight value is assigned to all three models, has lower accuracy comparing to the weight fusion shown in Sect. 3.5, where a proper weight is assigned to each model according to their individual performances. As the fact that MLCNN,

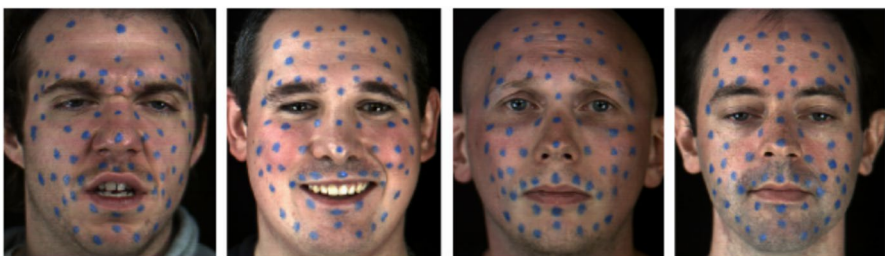
**Fig. 6** SAVEE data samples

Table 2 Recognition accuracy of separated models

Model	Accuracy
Baseline [26]	38.81%
MLCNN and 3DCNN	49%
VGG-FACE and LSTM	40.46%
Xception and STAT	39.68%

VGG-FACE, and Xception have different structures which are inspired by different types of network, they can capture different facial representations which can be used for the analysis of emotion expressions. An improvement of 5% accuracy shown in Table 2 and 3 indicated that the combination of different facial representations can boost the performance of emotion recognition in the wild.

Table 4 shows that the proposed deep learning framework with weight fusion is comparable with the 1st and 2nd winner of the Emotiw2016 Challenge [7, 27]. Their performances for validation data are about 52%, while our proposed model is 54.83%. The 1st and 2nd winner of Emotiw2016 also used audio data to combine with deep learning models from video data, while our proposed framework uses only video data. However, the utilization of CEV data makes our training data larger, therefore, the diversity of the emotional information that CNN models can learn is increased.

Table 5 lists the recognition accuracy of each emotion class using the weight fusion model. The proposed system performed well while determining the emotional states of anger, happiness, sadness, and neutral emotion. However, the performances while determining surprise and disgust were significantly inferior. This result is highly affected by the unbalancing data distribution shown in Table 1. The video samples for anger, happiness, and neutral appear the most in the training data, while a small number of videos for disgust, fear, and surprise is presented. Moreover, as shown in Fig. 7, emotions such as anger and happiness are strongly expressed by human faces. In contrast, emotions such as fear

Table 3 Performance of fusion models

Fusion model	Accuracy
Feature fusion	51.69%
Average	53%
Maximum voting	53.26%
Weight fusion	54.83%
Equal-weight fusion	51.69%

Table 4 Proposed fusion model performance

Model	Accuracy
Emotiw 2016 winner [27]	51.96%
Emotiw 2016 2nd winner [7]	51.96%
Proposed fusion model	54.83%

Table 5 Model performance based on the emotional category

Emotional class	Accuracy
Angriness	75.51%
Disgust	17.5%
Fear	30%
Happiness	69.44%
Neutral	53.88%
Sadness	51.25%
Surprise	14.28%

and surprise are ambiguous, make it hard to be distinguished even by human's cognitive.

4.3 Evaluation of SAVEE data

4.3.1 Subject-dependent evaluation

In this evaluation, the data from each subject were divided into training (80%) and testing (20%). We only applied horizontal flips for augmenting the training data. The testing data were evaluated with the Xception-STAT and the combined VGG-FACE and LSTM models. Table 6 lists the comparison of the performances of every single model with the baseline result [30]. The proposed models outperformed the baseline result by improving the accuracy by approximately 7%. All the subjects have more than 90% accuracy. Table 7 presents the performance of the four fusion methods.



Fig. 7 Examples of emotions expressed "in the wild": **a** Angry, **b** Happy, **c** Fear, **d** Surprise

The maximum voting method achieved the best performance in this situation. However, weight fusion still has one of the best performances with an improvement of more than 6% accuracy for all subjects.

4.3.2 Subject-independent evaluation

In this evaluation, we performed the leave-one-subject-out procedure, and the results were averaged based on the four tests. We applied the horizontal flips, rotation left, and rotation right, for the augmentation of the training data. Table 8 lists a comparison of the performances of every single model with the baseline result [30]. The accuracy was significantly improved by approximately 20%. Table 9 lists the performance of the four fusion methods. The maximum voting and weight fusion still achieve the best accuracy for this study. The accuracy of each emotional category is shown in Table 10. The emotions that are strongly expressed by human faces, such as anger and happiness, also achieved the highest accuracy, which is similar to the performance of AFEW data. However, emotions such as disgust and surprise, with the constrained conditions of SAVEE data, perform well compared to the corresponding category in AFEW data.

4.4 Discussions

CNN with a deep network structure containing a stack of convolutional layers, achieved outstanding results in many popular visual classification problems

Table 6 Subject-dependent for single models

Model	KL	JE	JK	DC
Baseline [30]	89%	89.8%	88.6%	84.7%
VGG-FACE and LSTM	93.3%	93.3%	94.1%	91.67%
Xception and STAT	95%	93.3%	95.83%	93.3%

Table 7 Subject-dependent for fusion models

Model	KL	JE	JK	DC
Baseline [30]	89%	89.8%	88.6%	84.7%
Feature fusion	95%	93.3%	94.1%	93.3%
Average	95%	95%	95.83%	96.67%
Maximum voting	100%	97.5%	97.5%	98.3%
Weight fusion	97.5%	95%	97.5%	95%

Table 8 Subject-independent for single models

Model	Accuracy
Baseline [30]	65%
VGG-FACE+LSTM	84.37%
Xception+STAT	82.29%

Table 9 Subject-independent for fusion models

Model	Accuracy
Feature fusion	85.83%
Average	85.83%
Maximum voting	86.45%
Weight fusion	86.45%

Table 10 Subject-independent performance of each emotional category

Emotional class	Accuracy
Anger	91.67%
Disgust	83.33%
Fear	83.33%
Happiness	91.67%
Neutral	87.5%
Sadness	83.33%
Surprise	83.33%

[17–20]. By extending the depth of the network with more blocks of convolution, the high-level features can be captured with better information for category discrimination. However, CNN learns via a stochastic training process, which leads to a high variance of outcomes. A solution to reduce the variance of CNN is the ensemble approach, which is motivated by training multiple models instead of a single model and combining the predictions from these models. Ensemble learning was proved to not only reduce the variance of predictions but also can achieve better performance than any single model [31].

In this paper, we proposed three different CNN-based networks to extract the high-level features of facial expressions. The predictions of emotional categories were decided by the ensemble of the three networks. The competitive evaluation results of the proposed system on unconstrained scenario data such as AFEW 2016 suggest that the ensemble of different CNN-based architectures, with the support of a sequential model, has the potential to improve the performance of facial expressions recognition in the wild. The utilization of CEV data also makes our training data larger; therefore, the diversity of the emotional information that CNN models can learn is expanded. However, the middle level of accuracies from AFEW 2016 data shows that it still has many rooms for improvement, especially the mission of extending and balancing training data to cover various expressions of “ambiguous” emotions such as surprise, disgust, and fear. Even though the proposed system mainly focuses on the unconstrained scenario data, it still achieved good performance in a controlled environment such as SAVEE data. Hence, the proposed system has the potential to be enhanced for real scenarios applications with the purpose of recognizing facial expressions such as human–robot interactions, drivers monitoring, etc.

5 Conclusions

In this study, we presented an ensemble of deep neural network models for recognizing emotions from visual data. The proposed models focused on the analysis of the face regions extracted from the video sequence. Three face representations were developed based on CNN. The first model used a multilevel CNN to extract facial features from each video frame and the 3DCNN to analyse the sequential information. The second model was a combination of the VGG-FACE and LSTM modules. The third model fine-tuned the Xception network and encoded the features statistically. We investigated four fusion strategies to combine these three models, which were the features fusion, average scoring, maximum voting, and weight fusion strategies. The weight fusion strategy exhibited the best performance in most cases. The proposed ensemble system achieved a competitive result by outperforming the baseline result obtained using two public datasets for emotion recognition. In future, we will conduct a multimodal test with different types of input information, including speech or text, to overcome the limitation of unimodal information.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Kołakowska A, Landowska A, Szwoch M, Szwoch W, Wróbel MR (2014) Emotion recognition and its applications. In: Hippe ZS, Kulikowski JL, Mroczek T, Wtorek J (eds) Human-computer systems interaction: backgrounds and applications, vol 3. Springer, Cham, pp 51–62
2. Qiang J, Zhiwei Z, Lan P (2004) Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Trans Veh Technol* 53(4):1052–1068. <https://doi.org/10.1109/TVT.2004.830974>
3. Rehg JM, Abowd GD, Rozga A, Romero M, Clements MA, Sclaroff S, Essa I, Ousley OY, Li Y, Kim C, Rao H, Kim JC, Presti LL, Zhang J, Lantsman D, Bidwell J, Ye Z (2013) Decoding children's social behavior. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp 3414–3421. doi:<https://doi.org/10.1109/CVPR.2013.438>
4. McDuff D, El Kaliouby R, Senechal T, Demirdjian D, Picard R (2014) Automatic measurement of ad preferences from facial responses gathered over the Internet. *Image Vis Comput* 32(10):630–640. <https://doi.org/10.1016/j.imavis.2014.01.004>
5. Yao A, Shao J, Ma N, Chen Y (2015) Capturing AU-aware facial features and their latent relations for emotion recognition in the wild. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA
6. Kahou SE, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent neural networks for emotion recognition in video. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA

7. Yao A, Cai D, Hu P, Wang S, Sha L, Chen Y (2016) HoloNet: towards robust emotion recognition in the wild. Paper presented at the Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan
8. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324. <https://doi.org/10.1109/5.726791>
9. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp 4489–4497
10. Omkar M, Parkhi AVaAZ (2015) Deep face recognition. Paper presented at the Proceedings of the British Machine Vision Conference (BMVC)
11. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
12. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21–26. pp 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
13. Sarangi P, Mishra B, Dehuri S (2017) Pyramid histogram of oriented gradients based human ear identification. *Int J Control Theory Appl* 10:125–133
14. Shan C, Gong S, McOwan P (2009) Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis Comput* 27:803–816. <https://doi.org/10.1016/j.imavis.2008.08.005>
15. Walecki R, Rudovic O, Pavlovic V, Pantic M (2015) Variable-state latent conditional random fields for facial expression recognition and action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 4–8. pp 1–8. <https://doi.org/10.1109/FG.2015.7163137>
16. Kaya H, Gürpınar F, Afshar S, Salah A (2015) Contrasting and combining least squares based learners for emotion recognition in the wild. In: 2015 ACM International Conference on Multimodal Interaction, pp 459–466. <https://doi.org/10.1145/2818346.2830588>
17. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. *Neural Inf Process Syst*. <https://doi.org/10.1145/3065386>
18. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv* 14091556
19. Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7–12. pp 1–9. <https://doi.org/10.1109/CVPR.2015.7298594>
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27–30. pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
21. Tang Y (2013) Deep learning using linear support vector machines. Paper presented at the Workshop on Challenges in Representation Learning, International Conference on Machine Learning, 06/02
22. Courville PCaA (2013) Challenges in representation learning: Facial expression recognition challenge. <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge>
23. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 7–10. pp 1–10. <https://doi.org/10.1109/WACV.2016.7477450>
24. Zhou S, Liang Y, Wan J, Li S (2016) Facial expression recognition based on multi-scale CNNs. In: 2016 11th Chinese Conference on Biometric Recognition, pp 503–510. https://doi.org/10.1007/978-3-319-46654-5_55
25. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA
26. Dhall A, Goecke R, Joshi J, Hoey J, Gedeon T (2016) EmotiW 2016: video and group-level emotion recognition challenges. <https://doi.org/10.1145/2993148.2997638>
27. Fan Y, Lu X, Li D, Liu Y (2016) Video-based emotion recognition using CNN-RNN and C3D hybrid networks. Paper presented at the Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan
28. Hu P, Ramanan D (2017) Finding tiny faces. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition

29. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. Paper presented at the International Conference on Learning Representations, 12/22
30. Haq S, Jackson PJB Speaker-dependent audio-visual emotion recognition. In: AVSP, 2009.
31. Amin-Naji M, Aghagolzadeh A, Ezoji M (2019) Ensemble of CNN for multi-focus image fusion. *Inf Fusion* 51:201–214. <https://doi.org/10.1016/j.inffus.2019.02.003>

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