A Dynamic Emotion Recognition System Based on Convolutional Feature Extraction and Recurrent Neural Network

Yida Yin[1], Misbah Ayoub[1], Andrew Abel[2], and Haiyang Zhang[1]

[1] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China,
yida.yin20@student.xjtlu.edu.cn, m.ayoub@student.xjtlu.edu.cn, haiyang.zhang@xjtlu.edu.cn

[2] Computer and Information Sciences, University of Strathclyde, Glasgow, Scotland, andrew.abel@strath.ac.uk

Abstract. Over the past three decades, there has been sustained research activity in emotion recognition from faces, powered by the popularity of smart devices and the development of improved machine learning, resulting in the creation of recognition systems with high accuracy. While research has commonly focused on single images, recent research has also made use of dynamic video data. This paper presents CNN-RNN (Convolutional Neural Network - Recurrent Neural Network) based emotion recognition using videos from the ADFES database, and we present the results in the arousal-valence space, rather than assigning a discrete emotion. As well as traditional performance metrics, we also design a new performance metric, PN accuracy, to distinguish between positive and negative emotions. We demonstrate improved performance with a smaller RNN than the initial pre-trained model, and report a peak accuracy of 0.58, with peak PN accuracy of 0.76, which shows our approach is very capable distinguishing between positive and negative emotions. We also present a detailed analysis of system performance, using new valence-arousal domain temporal visualisations to show transitions in recognition over time, demonstrating the importance of context based information in emotion recognition.

Keywords: emotion recognition, deep learning, convolutional neural network, recurrent neural network, visualisation

1 Introduction

Over the past three decades, there has been sustained research activity in emotion recognition from faces, powered by the popularity of smart devices and improved machine learning, resulting in the creation of facial emotion recognition systems with high accuracy. Facial expression has been found to play an major role in emotion recognition. According to a recent study of spoken communication, 7% of the emotion information is transmitted through words, 38% through voice and 55% through facial expressions [1].

This is a peer-reviewed, accepted author manuscript of the following conference paper:
Yin, Y., Ayoub, M., Abel, A., & Zhang, H. (2022). A dynamic emotion recognition system based on convolutional feature extraction and recurrent neural network. In K. Arai (Ed.), Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference IntelliSys Volume 2 (pp. 134-154). https://doi.org/10.1007/978-3-031-16078-3_8

There are two main stages in a emotion recognition system: feature extraction and classification. In earlier research, most systems extracted handcrafted features [2–4], and then separately used traditional machine learning algorithms for classification, such as Naive–Bayes classifiers, Gaussian Tree-Augmented Naive Bayes (TAN) classifiers and hidden hidden Markov models (HMM) [5]. However, there have been a lot of improvements in machine learning in recent years. Using increased computational power and deep learning, researchers can apply more complex models to recognise emotions and obtain good results. Convolutional neural networks (CNNs) are widely used for static emotion recognition and feature extraction. These can extract self-learned features as a feature vector, which can be input into a fully connected neural network as the input value for emotion classification [6] [7]. Similarly, 3D convolutional neural networks (C3D), which are suitable for temporal data [8] and recurrent neural networks (RNNs) are widely used for dynamic emotion recognition [9].

The theory behind facial emotions has also developed in recent decades. Ekman and Friesen [10] identified six basic human expressions, happiness, sadness, surprise, fear, anger, and disgust, and designed the Facial Action Coding System (FACs). This described the facial changes corresponding to each expression, including how eyebrows, eyes, eyelids and lips change. Many emotion recognition systems use similar discrete labels. However, this is a very limited approach, as discrete labels make it hard for computer to fully understand the slight difference between different emotions and to take account of dynamic information. To overcome these limitations, based on the theory of constructed emotions [11], it is possible to use continuous emotion labels to describe emotion. The two-dimensional continuous model [12] is a commonly used approach for defining expressions. It uses 2 values, valence and arousal, to describe emotion, where valence is how pleasant or unpleasant a face is, and arousal is the degree of alertness it would cause in an observer. There is also a three-dimensional continuous model [13] which uses valence, arousal and dominance.

This paper presents a dynamic emotion recognition system and evaluates its performance on posed emotions. There are several key contributions. Firstly, we use videos from the ADFES database [14], rather than the much more commonly used static single images [15–17]. We use windowing to input a sequence of frames, and use transfer learning with a CNN-RNN [18] to demonstrate good dynamic emotion results, showing the highest discrete emotion label accuracy is 58%, and that good results can be achieved with a small and lightweight model. We also thoroughly investigate misclassifications based on continuous values, and show that the system is generally correctly distinguishing between positive and negative emotions, explaining some misclassifications. Many other papers simply investigate discrete classifications, and focus on overall accuracy, without considering in detail exactly what is contributing to the misclassifications. Finally, a key contribution, we visualise the transition in emotion states in the arousal-valence domain, showing that with posed emotions, it is not simply a case of the labelling being incorrect, instead we can see that during the frames of a single video, a person's facial movements may transition and be recognised

as several different emotions. These transitions are rarely considered in current work in the literature. We demonstrate how these transitions can be simply and clearly visualised, showing the benefits of a logical and understandable approach to evaluation of emotion recognition performance.

The remainder of this paper is organized as follows. Section 2 summarises relevant machine learning emotion recognition background research. Section 3 introduces the continuous emotion labelling and the dataset we use. In section 4, we introduce our proposed emotion recognition system, including our model structure, loss function and performance metric. We present results in section 5, and finally, in section 6, we conclude the paper and provide future research directions.

2 Related Work

Emotion recognition includes static emotion recognition, which is based on recognising a single image, and dynamic emotion recognition, which takes into account temporal features in video sequences. Many methods have been proposed, including Zhang et al. [16], who applied discrete wavelet transform and biorthogonal wavelet transform to images, aiming to recognise 7 basic emotions using SVM classifiers, they reported peak accuracy of 0.967 on their dataset with a real time response time, using the JFFEd dataset.

Mistry et al. [19] used the CK+ and MMI image datsets. They proposed an mGA (micro Genetic Algorithm) embedded technique and Ensemble classifier and reported an accuracy of 0.9466. They compared their work positively with other heuristic algorithms. Oh et al. [20] used videos for micro emotion recognition. They used higher order riesz transform techniques with an SVM classifier, reporting 0.9 accuracy with the CASME II and SMIC datasets. Zen et al. [21] proposed Transudative parameter transfer: a framework for building personalized classification models, using visual data and gestures with the PAINFUL, CK+, and SWGR datasets. They achieved 0.9 accuracy with low computational cost but their accuracy was based on parameters of a personalised classifier. They used adaptive techniques in the proposed model with the help of transfer learning and trained their regression framework to learn the relation between the unlabeled data distribution and the classifier's parameters.

Boubenna and Lee [22] proposed a feature extraction algorithm named GA-LDA. GA-LDA is a genetic algorithm (GA) combined with a linear discriminant classifier (LDA), to improve accuracy and calculation speed. They use the Radboud Faces Database (RaFD) [23], and their results show that GA-LDA has the same accuracy (98.67%) as using the VGG-Face CNN model. However, the feature vector size is decreased and training time is shorter. The advantage of GA-LDA is it can extract features from images into a smaller feature vector with high accuracy. It can reduce training time and save storage space. The limitation of this research is that they only test this algorithm on a small dataset.

2.1 Convolutional Neural Network based methods

The CNN approach has become widely used in video emotion recognition. VGG-19 [24] is a widely used CNN architecture, containing 16 convolution layers and 3 fully connected layers. VGG-19 contains many small kernels of the same size (such as 3x3). With this architecture, VGG-19 can extract a large number of features from images with high accuracy. In [25], Cheng and Zhou improved VGG-19 by using 2 fully connected layers with 4096 and 7 neurons to replace the previous 3 fully connected layers with 4096, 4096 and 1000 neurons. They report 96% accuracy on the Extended Cohn-Kanade Dataset (CK+) [26], and demonstrate that a smaller model can be better. Riaz et al. [27] also applied VGG-19 to recognize emotion. They reported 59.02%. accuracy with the FER-2013 dataset [28].

The 3D convolutional neural network (C3D) has a similar architecture. The main difference between CNN and 3DCNN is the C3D input is a three-dimensional matrix rather than 2D. This contains a number of images stacked on top of each other. In [8], Fan et al. designed a emotion recognition system containing C3D and CNN-RNN. C3D is used to extract features from image sequences, and CNN-RNN can extract features from each image by CNN and then input the sequences into an RNN to extract time features. Audio information could also be used. They reported 59.02%. accuracy with the AFEW dataset [29].

2.2 CNN-RNN

The CNN-RNN architecture is another widely used method for dynamic emotion recognition. Here, the CNN is the feature extractor and RNN is used to deal with the time sequence. Rangulov and Fahim [18] proposed a CNN-RNN model. The CNN they use is a simple CNN architecture [30] and RNN uses gated recurrent units (GRU). They used databases including Aff-Wild2 [31], CK+ [26] and Japanese Female Facial Expression (JAFFE) [32]. This system's accuracy is 0.95 on CK+ and 0.50 on JAFFE. Because Aff-Wild2 is a database labeled with continuous numeric annotations, there is only recording for loss function. The lowest CCC for valence is 0.23 and the lowest CCC for arousal is 0.39. They also report that the performance is unbalanced between negative and positive emotion. Ebrahimi Kahou et al. [33] also used a CNN-RNN architecture. They compared 4 emotion recognition methods, including aggregated CNN, audio, activity, and CNN-RNN. They use long short-term memory (LSTM) approach. Databases in this research includes AFEW [29], the Toronto Face Database (TFD) [34] and the Facial Expression Recognition dataset (FER2013) [28]. They report a peak accuracy of 0.4. In this paper, we will follow a similar approach.

2.3 Context-Aware Emotion Recognition Networks

In [35], Lee et al. argue that the change of facial expression is not distinct and regular in a natural state, and recognition accuracy can be improved by using context information, like body gestures. They proposed CAER-Net, which contains

two CNN networks, one to extract facial features, and another to extract features from context. They use the AFEW [29] and CAER [35] databases. The accuracy of CAER-Net over AFEW is 43.21%, CAER is 38.65% and AFEW+CAER is 51.68%.

Ronak Kosti et. al [36], presented emotions in a context database in a non-controlled environment (Emotic). people were annotated with 26 emotions according to people's apparent emotional state, according to continuous labels of valence, arousal, and dominance [37]. To show the importance of considering context in image for recognizing people's emotion, they trained CNN that analyses the person and the whole scene to recognize rich information about emotional states. Results showed that the Jaccard index for the categories recognition is higher than 0.4 and the error on the continuous dimension is lower than 0.5.

Overall, we can see that the majority of lab based CNN approaches report very good results, but on single images, whereas using dynamic emotional data from videos is less widely investigated and tends to produce poorer results. In table 1, we provide a summary of relevant research.

Research	**Method**	**Database**	**Type**	**Accuracy**
[24]	VGG-19	CK+	Static	0.96
[22]	GA-LDA	RaFD	Static	0.99
[8]	C3D & CNN-RNN	AFEW	Dynamic	0.59
[18]	CNN-RNN	Aff-Wild2, CK+ and JAFFE	Dynamic	0.95(CK+), 0.50(JAFFE)
[33]	CNN-RNN	FER2013	Dynamic	0.40
[35]	CAER-Net	AFEW+CAER	Dynamic	0.52

Table 1. Emotion recognition methods overview

3 Continuous Emotion Recognition Labelling

3.1 Dataset

This paper uses the Amsterdam Dynamic Facial Expression Set (ADFES) provided by the Amsterdam Interdisciplinary Centre for Emotion (AICE) [14]. ADFES has 648 posed emotion videos recorded at 25 fps, and 10 different posed emotions: anger, contempt, disgust, embarrass, fear, joy, neutral, pride, sadness and surprise. ADFES has not only face-forward videos but also turn-toward and turn-away videos. Each of those 3 kinds of videos has 216 samples, including 22 videos for each emotion, ranging from 5.6s to 6.5s, using 22 models, including 10 female models and 12 male models. There are 10 Mediterranean models and 12 North-European models. Figure 1 displays examples of ADFES. From left to right: anger, contempt, disgust, embarrass, fear, joy, neutral, pride, sadness and surprise.

Previous studies showed that ADFES has a high recognition rate. Wu and Lin [38] developed the Weighted Center Regression Adaptive Feature Mapping (W-CR-AFM) method for static image emotion recognition from images as opposed to dynamic emotion recognition on videos. Wingenbach et al. [39] found that each

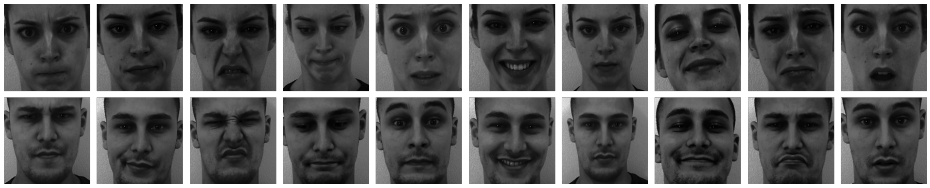


Fig. 1. Examples of ADFES

ADFES emotion is initially neutral and the expression intensity changes from low to high over time. They therefore divided each video into low, middle and high expression intensity videos. This is the ADFES-BIV dataset. Abdulsalam et al. [7] designed a deep convolutional neural network (DCNN) model for ADFES-BIV with images, reporting a highest accuracy of 0.95. However, this dataset does does not contain all the videos from ADFES. In this paper, we use the last 50 frames for training, as they have the highest intensity, and use all ADFES forward facing videos.

We use ADFES rather than alternatives such as Aff-Wild2 [31] and Dynamic Facial Expression in the Wild (DFEW) [40] because the videos are high quality and well organized, and ADFES is also a gender-balanced and region-balanced database, which which can decrease biases. Finally, previous research, including [38, 39], has shown that ADFES can be successfully used for recognition.

3.2 Continuous Emotion Labelling

In this research, we apply the two-dimensional valence-arousal continuous model presented by Scherer [12] to describe emotion. Scherer argued that any emotion can be described using valence and arousal values, where valance measures how positive and pleasant an emotion is, and arousal measures the agitation level of the person, ranging from non-active to active. Figure 2 displays the valence-arousal space taken from [12], showing valence along the x-axis, and arousal along the y-axis. For example, excitement is assigned both high valence and high arousal, whereas joy has a similar valence, but a lower arousal. Haag et al. [41] used the valence and arousal model, calculated separately, and achieved 89.9% and 96.6% accuracy rate for valence and arousal respectively, with neural network classifiers. A third dimension known as dominance is also sometimes incorporated in this model, which measures the control level of the situation by the person ranging from non- control to in-control. Some recent work [36] [37] uses continuous dimensions of VAD (valance, arousal and dominance) to describe emotions using three numeric dimensions, but we focus on a two dimensional approach.

Table 2 gives the equivalent valence-arousal values for the emotions represented in the ADFES dataset. The advantage of the two-dimensional continuous model is that it can distinguish the nuances of different expressions, and can help the computer to better understand human expressions. For our training,

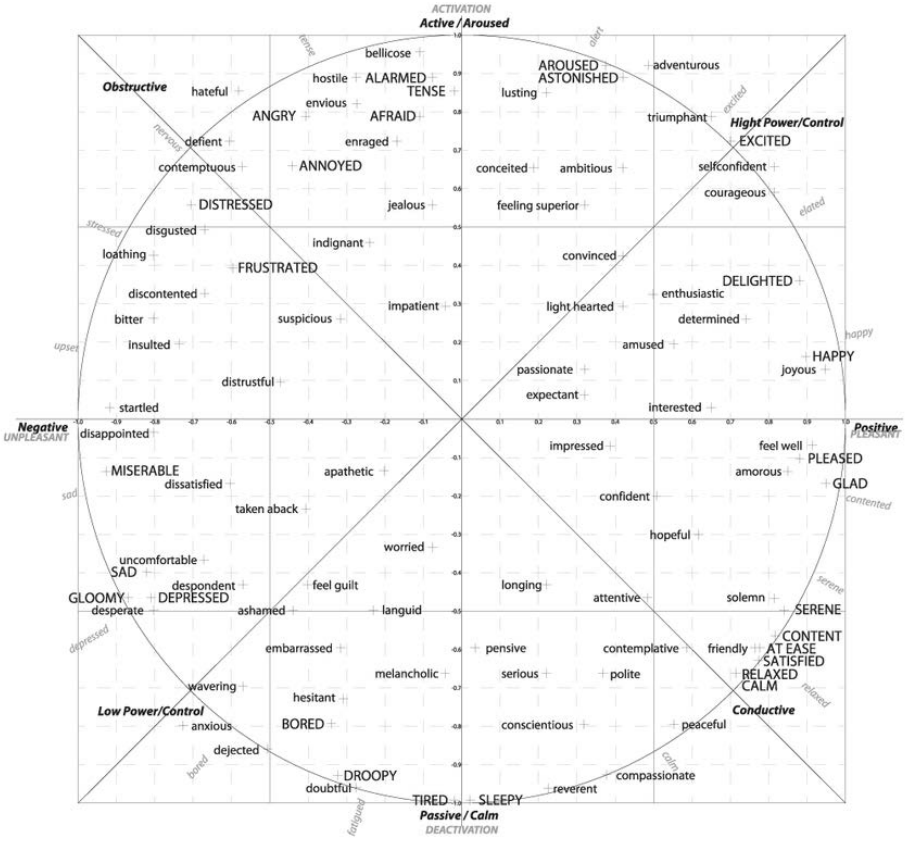


Fig. 2. Valence and arousal space, taken from from Scherer [12]

we label the last 50 frames in each video with the values shown in table 2. For example, if a video is labeled as Anger, we will set valence and arousal values of all frames in this video to (-0.4, 0.8).

However, continuous labels are harder to understand. The output of our proposed approach is a pair of arousal-valence vales, which is then assigned the emotion corresponding to its closest match. For the whole video, we find the most widely assigned basic emotion and use that basic emotion to describe this video.

4 CNN-RNN Emotion Recognition System

We conducted many preliminary experiments, not reported here due to space limitations, and the approach we proposed to use for emotion recognition is the CNN-RNN system.

Emotion	Valence	Arousal		Emotion	Valence	Arousal
Disgust	-0.68	0.50		Fear	-0.11	0.79
Contempt	-0.57	0.66		Neutral	0.00	0.00
Sadness	-0.50	-0.86		Pride	0.31	0.55
Anger	-0.40	0.80		Surprise	0.38	0.92
Embarrass	-0.31	-0.60		Joy	0.95	0.12

Table 2. Examples of continuous emotion labels

4.1 CNN

The CNN architecture used in this paper is based on the model used by Rangulov and Fahim [18] and Khorrami et al. [30], and is shown in figure 3. This architecture has 3 convolution layers and 3 pooling layers. Every convolution layer is followed by Rectified Linear Unit (ReLU) activation functions. The first layer is the batch normalization layer to suppress overfitting [42]. The second layer is a convolution layer containing 64 filters with the size 5x5. The layer after the first convolution layer is a max-pooling layer. The fourth layer is another convolution layer containing 128 filters with the size 5x5. The fifth layer is another max-pooling layer. The third convolution layer has 258 filters with the size 5x5. After the third convolution layer is a quadrant pooling layer [43]. Finally, the system flattens all matrices and creates a feature vector containing 300 elements. We use a pre-trained CNN-RNN model from Rangulov and Fahim [18], which has previously demonstrated good results [18, 33], and is easy to train and use. In addition, compared with methods like C3D, CNN-RNN regards emotion data as a kind of time series data and focuses on the video's time features, which is consistent with our understanding of emotion.

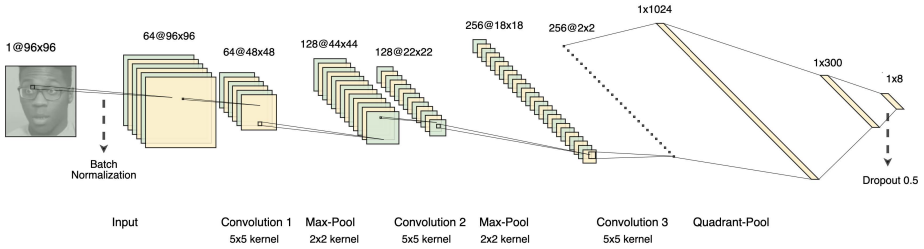


Fig. 3. CNN architecture, taken from Rangulov and Fahim [18]

4.2 RNN

An RNN is a deep learning model that can extract time features from sequential data. Compared with feedforward neural networks, RNNs receive not only an input value x_t but also h_{t-1}, provided by the previous time point $(t-1)$. With those input values, an RNN will provide output value y_t for time point t and hidden state h_{t+1} for the next time point $(t+1)$ [33] [44]:

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{1}$$
$$y_t = g(W_{hz}h_t + b_z) \tag{2}$$

where $g()$ is the activation function and W_{xh}, W_{hh} and W_{hz} are the weight matrices. The Gated recurrent unit (GRU) is a special type of RNN. It is designed to solve problems such as long-term memory and gradients in back propagation [45]. The GRU output can be described as follows [46]:

$$r(t) = \sigma_g(W_r x(t) + U_r h(t-1) + b_r) \tag{3}$$
$$z(t) = \sigma_g(W_z x(t) + U_z h(t-1) + b_z) \tag{4}$$
$$\hat{h}(t) = \sigma_h(W_h x(t) + U_h(r(t) \circ h(t-1)) + b_h) \tag{5}$$
$$h(t) = (1 - z(t)) \circ h(t-1) + z(t) \circ \hat{h}(t) \tag{6}$$

where $x(t)$ is the input at time point t, $h(t-1)$ is the hidden state from time point $(t-1)$, $h(t)$ is the hidden state and $\hat{y}(t)$ is the output value, $\sigma_g()$ is the Sigmoid activation function and $\sigma_h()$ is the TanHyperbolic activation function. The relationship between output $\hat{y}_t$ and hidden state $h(t)$ is the same as standard RNN.

We conducted preliminary experiments with different RNN structures. When we applied the original pre-trained model [18], the final accuracy was 38%. When we decreased the number of parameters by decreasing the number of hidden layers and hidden units, the accuracy increased to 58%, thus identifying that a model with less parameters has higher accuracy than the initial model. We are interested in 2 different RNN architectures and will compare their performance. One is a 1-layer RNN model having 10 hidden units, denoted as CNN+RNN(10). The other is a 3-layer RNN model having 10, 10 and 5 hidden units in hidden layers, denoted as CNN+RNN(10, 10, 5).

4.3 Loss Function

Rather than regression evaluation metrics such as mean squared error (MSE) and root mean squared error (RMSE), we choose to use concordance correlation coefficient (CCC), as proposed by Lawrence and Lin [47], to measure the correlation between actual value and predicted value. The formula of CCC is shown in equation 7:

$$\rho_{ccc} = \frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \tag{7}$$

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} \tag{8}$$

where σ_X is the variance of X, μ_X is the mean value of X, $\rho_{X,Y}$ is the correlation coefficient between X and Y.

CCC contains 2 kinds of information. Firstly, a higher absolute value of CCC means two datasets have a strong linear relationship. Secondly, it shows the information of the slope of the regression line for the two sets of data. If CCC is close to 1, it means two data sets have a strong linear correlation and their regression line's slope is close to 1. If CCC is close to -1, it means two data sets have a strong linear correlation and their regression line's slope is close to -1. To make the CCC of actual and predicted value close to 1, we use the loss function as given in equation 9:

$$L_{ccc} = 1 - \frac{1}{2}[\rho_{valance} + \rho_{arousal}] \tag{9}$$

The process of minimizing L_{ccc} will maximize $\rho_{valance}+\rho_{arousal}$, which makes the predicted value more consistent with actual value. Compared with RMSE and MSE, CCC has its advantages. RMSE and MSE only focus one each sample's accuracy and they do not consider the relationship between data in different time points. However, the data we deal with is time series data and the relationship between output value in different time points is important. In order to keep the relationship between output values, we use CCC in our loss function.

4.4 PN Accuracy

To measure the system's ability to distinguish between positive and negative emotion, we design a performance metric named *PN Accuracy*. Emotions with negative valence are labelled as negative emotion, such as disgust, contempt, sadness, anger, embarrass and fear. Emotions with positive valence are regard as positive emotion, such as pride, surprise and joy. PN accuracy is shown in equation 10.

$$PN\ Accuracy = \frac{n_{correctNegative} + n_{correctNeutral} + n_{correctPositive}}{n_{total}} \tag{10}$$

where $n_{correctNegative}$ is the number of negative emotion samples that are correctly recognised as negative emotions, $n_{correctNeutral}$ is the number of neutral emotion samples that are correctly recognised as neutral emotions, $n_{correctPositive}$ is the number of positive emotion samples that are correctly recognised as positive emotion and n_{total} is the total number of samples.

4.5 System Configuration

Dataset preprocessing We pre-process the ADFES dataset with following steps:

1. First, we extract frames from videos at 25fps.
2. We perform grayscale processing and histogram equalization. We use the average value of RGB to set each pixel's gray level, and apply histogram equalization to each gray image.

3. We use the widely used Dlib algorithm to identify and crop the face region.
4. The fourth step is data augmentation. To have a larger dataset for model training, we use image flipping to double our dataset. We regard the speaker in the videos generated by image flipping as different from the original. After data augmentation, our dataset has 44 speakers and 432 videos.
5. Finally, we group our dataset into 3 sets: training, validation and testing, according to person. We keep speakers separate and all videos of one speakers will be in the same group. The ratio of speakers in the three groups is 8:2:1. The speakers are randomly shuffled each run.
6. Because expression intensity changes from the beginning to the end [39], we only label and use the last 50 frames in each video.

Computer configuration

- Processor: Intel ® Xeon® W-2133 @ 3.6 GHz.
- RAM: 32 GB.
- GPU: Quadro P4000, 8G.
- Windows 10 professional.
- PyCharm (IDE), Python (3.8), OpenCV-python (4.5.4.58), tensorflow (2.6).

Deep learning network configuration The basic structure of our deep learning model is discussed in section 4. We apply transfer learning using the pre-trained CNN model from Rangulov and Fahim [18]. The batch size is set to 128, the time step of RNN is 10 (meaning we use 10 image frames) and the max epoch of both CNN and RNN is 100. As discussed in section 4, we used 2 model configurations, CNN+RNN(10) and CNN+RNN(10, 10, 5).

5 Results and Analysis

5.1 Model Evaluation

Using the configurations discussed in section 4, we compare the performance of CNN+RNN(10) and CNN+RNN(10, 10, 5), with the results shown in table 3. The results are an mean of 5 runs.

Firstly, table 3 shows that the results are generally good, with both models reporting an accuracy of 0.58, with a low interquartile range (IQR). However, as might be expected, performance was considerably worse for unseen data, with accuracy of 0.44 and 0.43 for both models. This is in line with other research into dynamic emotion recognition [33]. However, of interest is our PN accuracy measure. This is much higher, with validation means of 0.76 for both models, and test means of 0.62 and 0.64. This indicates that our system is very good at distinguishing between positive and negative emotions, even with unseen data, but considerably worse at identifying the specific emotion. Another interesting finding is in the confusion matrices in figure 4. The precision and recall of positive emotions are higher than negative emotions. Contempt has the lowest

	CNN+RNN(10)		CNN+RNN(10, 10, 5)	
	Training and Validation (IQR)	Testing (IQR)	Training and Validation (IQR)	Testing (IQR)
Average Macro-Precision	0.59	0.39	0.59	0.41
Average Macro-Recall	0.58	0.43	0.58	0.43
Average Accuracy	0.58 (0.08)	0.44 (0.03)	0.58 (0.06)	0.43 (0.10)
Average PN Accuracy	0.76 (0.06)	0.62 (0.08)	0.76 (0.09)	0.64 (0.08)

Table 3. Summary of model evaluation. The numbers in brackets are the IQR value.

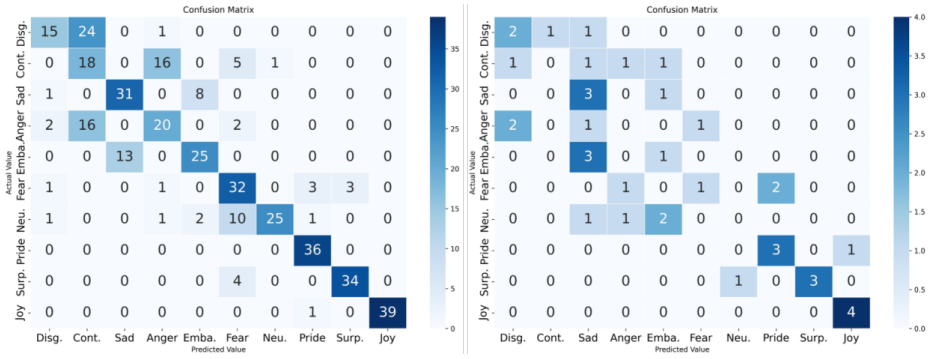


Fig. 4. CNN+RNN(10)'s confusion matrix in round 2 using training and valuation (left) and testing (right) dataset

precision and recall in all 5 rounds. All positive emotions' accuracy are close to 100%. However, with negative emotions, the system produces much worse results. The confusion matrices show that for example, disgust is recognised as contempt, contempt as anger, and anger as contempt. Both CNN+RNN (10) and CNN+RNN (10, 10, 5) have similar performance, but here, we focus only on CNN+RNN(10). Both models have poor performance when they are used to recognise unseen speakers' emotions, however, both successfully distinguish between positive and negative emotions. Finally, there is considerably better performance for positive emotions than negative emotions.

5.2 Comparison with Related Works

We compare this research with 4 related researches, including 2 papers [18, 33] that use the same recognition method and 2 that use the same database [38, 7]. These are summarised in table 4.

Compared with Rangulov and Fahim [18] and Ebrahimi et al. [33], our accuracy is only lower than Rangulov and Fahim's [18] accuracy on CK+. We experimented with varying numbers of CNN-RNN architecture's parameters, and identified that a smaller model is more suitable for this smaller ADFES database. The key difference between this research and others [38, 7] is that we

focus on the emotion of the whole video and they focus on the emotion of single images.

Research	Method	Database	Type	Annotation	Accuracy
This research	CNN-RNN	ADFES	Dynamic	Continuous	0.58
[18]	CNN-RNN	Aff-Wild2, CK+ and JAFFE	Dynamic	Continuous	0.95(CK+), 0.50(JAFFE)
[33]	CNN-RNN	FER2013	Dynamic	Discrete	0.40
[38]	W-CR-AFM	ADFES	Static	Discrete	0.92
[7]	DCNN	ADFES-BIV	Static	Discrete	0.95

Table 4. A comparison with related researches

5.3 Output Visualisation

While our results are good, and we are able to identify that the poor performance is primarily with negative emotions, we are interested in investigating model performance in more depth. To do this, we perform two visualisations. One to display the trajectory of emotion over time, and another to describe the distribution of all samples' output. The output used in this section is the training output.

Output Distribution Figure 5 displays the arousal-valence distribution of all samples' output values. Orange points represent the frames in videos that are correctly identified and blue points represent the frames in the misidentified videos. Firstly, it allows us to visualise the individual emotion accuracy rate. For example, we can see that contempt has a low accuracy, as the blue area is much larger than the orange. Secondly, it describes the characteristics of the distribution of video that are correctly recognised. We can see that the positive emotions such as joy tend to be much better classified, and the negative emotions such as contempt and disgust contain multiple clusters.

We are also interested in misclassification details. For example, with sadness, false points have similar valence values as true points but much higher arousal values. This means that some sadness samples are counted as errors, but are are very close in terms of output. Another example is provided with joy. False points have similar arousal values to true points but lower valence values. We can therefore clarify classification performance, as many of the "misclassifications" are in fact very similar.

Visualisation of Temporal Changes in Dynamic Emotion We also investigated how emotion recognition changed over time. In the videos, the speakers started off with a neutral expression, and then posed the designated emotion. This means that there is a transition in facial expressions. We considered this when we used only the last 50 frames of the video for training and testing, but we also wished to visualise the results of testing the complete video sequences.

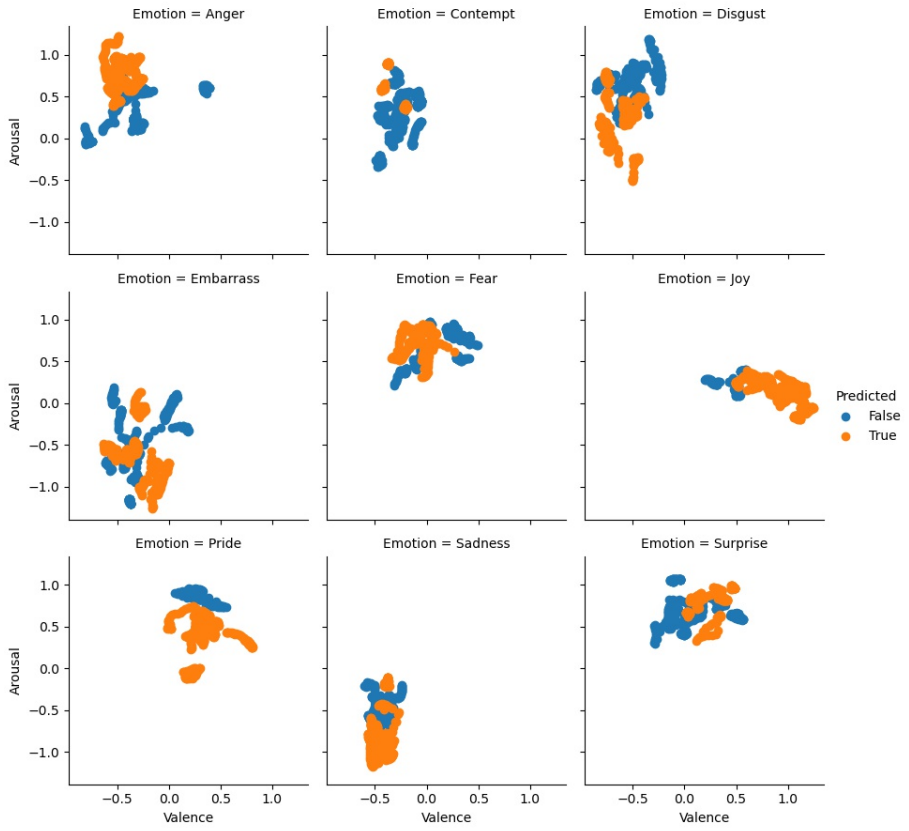


Fig. 5. Distribution of all samples' output values

Figure 6 shows examples of plotting a single video's emotions on the arousal-valence space, with different emotions represented with different colours, and the frame number shown with larger markers over time.

First of all, we can see that in all figures, the emotion tends to be initially plotted as having a neutral location, before then changing over time. This transformation results in some interesting classifications. In 6 (b) the emotion is initially neutral. With the increase in arousal, the emotion becomes closer to fear. When the arousal increases to 0.9, it becomes stable and valence starts to decrease. The final frames are close to the basic point of Anger $(-0.4, 0.8)$ and we can estimate that this video is an Anger video. For a positive emotion like joy or surprise, figure 6 (e), shows a relatively straight forward transform from initially neutral, to then become less neutral and more joyful over time. This confirms the findings of other work, and provides a clear visualisation [14, 39].

In contrast, other emotions are less simple. With contempt, figure 6 (d), the initial frames are first classified as fear, then correctly as contempt, but then

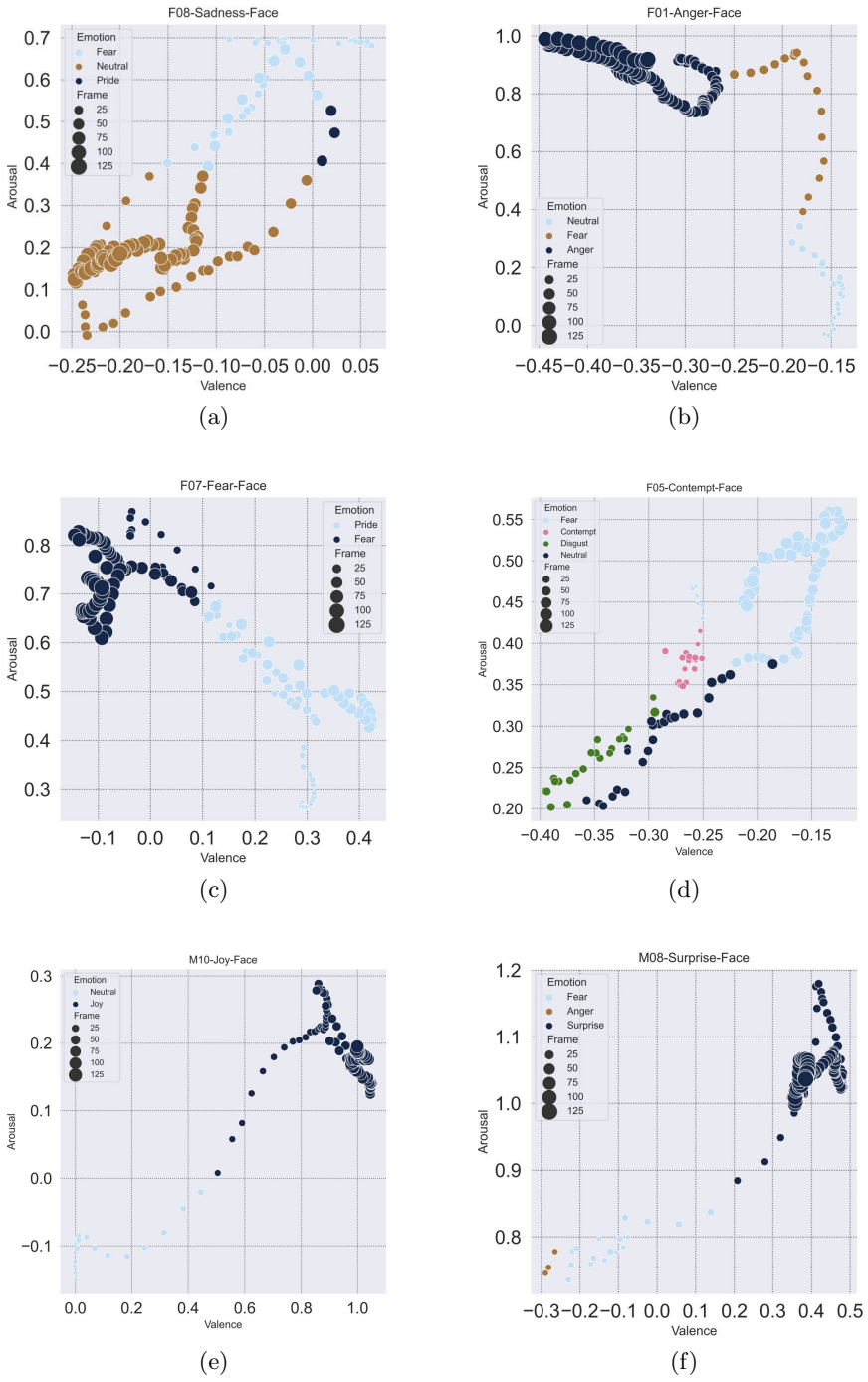


Fig. 6. Plotting of emotion change from different videos

over the course of the video, are considered to be first disgust, then neutral, then fear. The point is close to point $(-0.20, 0.45)$ at the end of the video. As the basic point of Contempt is (-0.57, 0.66), we can clearly see that our recognition system estimates a larger valence value and lower arousal value than it should. This partly explains the mis-classifications in our work (and others), as dynamic emotional videos are not as straightforward to estimate as single videos.

Finally, figure 7 displays all outputs of a single emotion, (a) anger, and (b) joy from all persons. Firstly, looking at joy, we can see that the classification is generally correct. While the emotions may start in different places, over time, they tend to converge to the same cluster, demonstrating that the system is able to easily identify this emotion. In contrast, anger is very different. The results do not all converge to a similar cluster, and there is less consistency in representation. This confirms our findings that negative emotions are harder to identify than positive emotions.

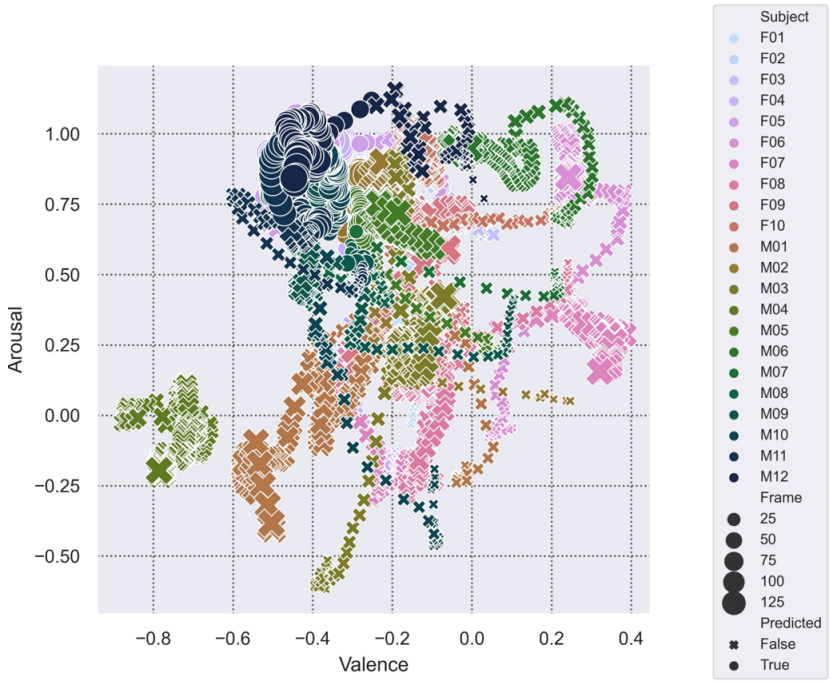
Overall, with these dynamic visualisations, we are able to show how emotion classifications changes over time during a single video, and why we have different classification performance between positive and negative results. for example in the case of anger, it would suggest that either the system is less able to distinguish anger from other negative emotions, or that anger is a less consistently performed emotion than a more positive emotion like joy.

We can therefore see that the CNN-RNN system we use is not only able to identify emotions successfully, but by analysing the output and using continuous values rather than discrete classifications, we can see how classifying videos as a discrete emotion is problematic, but that we can also see transitions between different states.

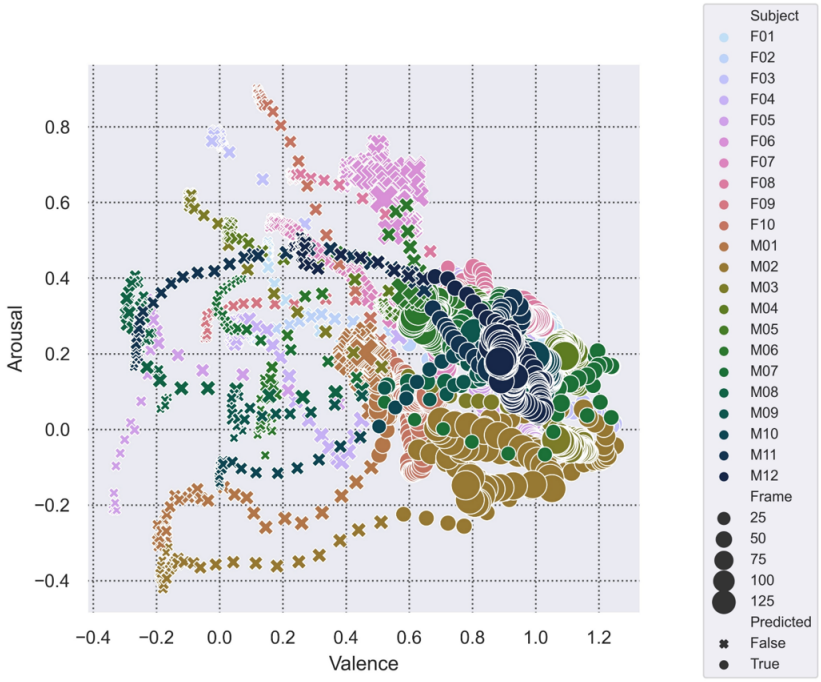
6 Conclusions

This paper proposed a dynamic emotion recognition system based on CNN feature extraction and RNN arousal-valence recognition. The results showed good results, and we also designed a new performance metric, the PN accuracy to measure whether our recognition models can distinguish between positive and negative emotions. We used two different RNN configurations, and our results show that the highest PN accuracy is 0.76 and our recognition system has a high ability to distinguish between positive and negative emotions. It also reflects that a lot of incorrect classifications are within with same set of emotions (i.e. negative emotions tend to be misclassified as other negative emotions). We also find that in comparison to previous research [18, 33], with a smaller dataset, a smaller deep learning model is more suitable, and that even an RNN with a single hidden layer can deliver good performance.

The plotting of emotional transitions on the arousal-valence space shows transitions over time, reasons for misclassifications, and clearly demonstrates which emotions perform better than others. Rather than simply classifying a discrete emotion, by plotting in the arousal-valence space, we can explain how emotion changes during a video and evaluate the performance of the recognition



(a) Anger



(b) Joy

Fig. 7. The outputs of the same emotion from different subjects

model. We can also identify that with regard to classification performance, some of the "misclassifications" can be explained by transforms over time. As shown by us and discussed by previous research [39], face expression usually starts from neutral and the expression intensity increases. It is therefore reasonable if some frames are misidentified. The more interesting visualisation is how the emotion changes over time. Finally, it shows that the information provided by single frame or single face is limited, and stresses the importance of temporal context. In future research, we will improve the model we use, and experiment with alternative configurations, including deeper networks. We have also noticed that negative emotion classification is less accurate, and we aim to improve this. We also propose to use more information, including facial, audio and gesture data, to increase emotion recognition accuracy and also improve the explanation and visualisation of model performance still further.

References

1. Mehrabian, A.: Communication without words. In: Communication theory, pp. 193–200. Routledge (2017)
2. Yacoob, Y., Davis, L.S.: Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on pattern analysis and machine intelligence **18**(6), 636–642 (1996)
3. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE transactions on image processing **16**(1), 172–187 (2006)
4. Ma, L., Khorasani, K.: Facial expression recognition using constructive neural networks. In: Signal Processing, Sensor Fusion, and Target Recognition X, vol. 4380, pp. 521–530. International Society for Optics and Photonics (2001)
5. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and image understanding **91**(1-2), 160–187 (2003)
6. Zhang, B., Quan, C., Ren, F.: Study on cnn in the recognition of emotion in audio and images. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–5. IEEE (2016)
7. Abdulsalam, W.H., Alhamdani, R.S., Abdullah, M.N.: Facial emotion recognition from videos using deep convolutional neural networks. International Journal of Machine Learning and Computing **9**(1), 14–19 (2019)
8. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: Proceedings of the 18th ACM international conference on multimodal interaction, pp. 445–450 (2016)
9. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y.: Spatial–temporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics **49**(3), 839–847 (2018)
10. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17**(2), 124 (1971)
11. Barrett, L.F.: How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt (2017)
12. Scherer, K.R.: What are emotions? and how can they be measured? Social science information **44**(4), 695–729 (2005)

13. Schlosberg, H.: Three dimensions of emotion. Psychological review **61**(2), 81 (1954)
14. Van Der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B.: Moving faces, looking places: validation of the amsterdam dynamic facial expression set (adfes). Emotion **11**(4), 907 (2011)
15. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.S., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 47–56 (2014)
16. Zhang, Y.D., Yang, Z.J., Lu, H.M., Zhou, X.X., Phillips, P., Liu, Q.M., Wang, S.H.: Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Access **4**, 8375–8385 (2016)
17. Go, H.J., Kwak, K.C., Lee, D.J., Chun, M.G.: Emotion recognition from the facial image and speech signal. In: SICE 2003 Annual Conference (IEEE Cat. No. 03TH8734), vol. 3, pp. 2890–2895. IEEE (2003)
18. Rangulov, D., Fahim, M.: Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network. In: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), pp. 14–20. IEEE (2020)
19. Mistry, K., Zhang, L., Neoh, S.C., Lim, C.P., Fielding, B.: A micro-ga embedded pso feature selection approach to intelligent facial emotion recognition. IEEE transactions on cybernetics **47**(6), 1496–1509 (2016)
20. Oh, Y.H., Le Ngo, A.C., Phari, R.C.W., See, J., Ling, H.C.: Intrinsic two-dimensional local structures for micro-expression recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1851–1855. IEEE (2016)
21. Zen, G., Porzi, L., Sangineto, E., Ricci, E., Sebe, N.: Learning personalized models for facial expression analysis and gesture recognition. IEEE Transactions on Multimedia **18**(4), 775–788 (2016)
22. Boubenna, H., Lee, D.: Image-based emotion recognition using evolutionary algorithms. Biologically inspired cognitive architectures **24**, 70–76 (2018)
23. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion **24**(8), 1377–1388 (2010)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Cheng, S., Zhou, G.: Facial expression recognition method based on improved vgg convolutional neural network. International Journal of Pattern Recognition and Artificial Intelligence **34**(07), 2056,003 (2020)
26. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, pp. 94–101. IEEE (2010)
27. Riaz, M.N., Shen, Y., Sohail, M., Guo, M.: Exnet: An efficient approach for emotion recognition in the wild. Sensors **20**(4), 1087 (2020)
28. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International conference on neural information processing, pp. 117–124. Springer (2013)
29. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE multimedia **19**(03), 34–41 (2012)
30. Khorrami, P.R.: How deep learning can help emotion recognition. Ph.D. thesis, University of Illinois at Urbana-Champaign (2017)

31. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
32. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE transactions on pattern analysis and machine intelligence **21**(12), 1357–1362 (1999)
33. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 467–474 (2015)
34. Susskind, J.M., Anderson, A.K., Hinton, G.E.: The toronto face database. Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep **3** (2010)
35. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10,143–10,152 (2019)
36. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotic: Emotions in context dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 61–69 (2017)
37. Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genetic, social, and general psychology monographs (1995)
38. Wu, B.F., Lin, C.H.: Adaptive feature mapping for customizing deep learning based facial expression recognition model. IEEE access **6**, 12,451–12,461 (2018)
39. Wingenbach, T.S., Ashwin, C., Brosnan, M.: Validation of the amsterdam dynamic facial expression set–bath intensity variations (adfes-biv): A set of videos expressing low, intermediate, and high intensity emotions. PloS one **11**(1), e0147,112 (2016)
40. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2881–2889 (2020)
41. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First steps towards an automatic system. In: Tutorial and research workshop on affective dialogue systems, pp. 36–48. Springer (2004)
42. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
43. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
44. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634 (2015)
45. Wang, W., Yang, N., Wei, F., Chang, B., Zhou, M.: Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 189–198 (2017)
46. Wang, Y., Liao, W., Chang, Y.: Gated recurrent unit network-based short-term photovoltaic forecasting. Energies **11**(8), 2163 (2018)
47. Lawrence, I., Lin, K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics pp. 255–268 (1989)