IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Emotion Classifier Using Deep Learning

[1] Mr. Dileep Kr. Kushwaha, ,[2]Kartik kumar [2] Ishita Sundriyal ,[2]Himanshu, [2]Ayushi Saxena
[1]Department of IT, Noida Institute of Engineering and Technology, Greater Noida 201308, Uttar Pradesh, India
[2]Department of IT, Noida Institute of Engineering and Technology, Greater Noida 201308, Uttar Pradesh, India

Abstract. Emotions are essential to comprehending human interactions. Efforts are being made to discover techniques that can mimic the human capacity to recognize emotions conveyed through facial expressions, variations in tone while speaking, and images of faces. Human Expression Recognition (HER) is among these disciplines. This paper reviews machine learning classification and deep learning algorithms for human expression recognition systems using multimodal signals. This work would assist individuals in forming relationships and is applicable in various fields, including the HCI (Human-Computer Interaction) and pharmaceutical industries. The speech and video inputs are selected and intend to develop a model that collects data from each respective data set and predicts the emotion class. The primary purpose of this paper is to enable researchers to assess the feasibility of human-computer interfaces that are sensitive to a person's emotions. The reuse of a previously learned model on a new problem is known as transfer learning, which is popular in deep learning now since it can train deep neural networks with a small amount of data. In this paper, we have applied a Deep learning model, i.e. CNN, and compared it with the existing models such as Multilayer perceptron and Decision tree classifier. The study aims to approach and improve continuous human expression recognition via video and audio and report the most recent developments in this field. To improve the accuracy of the existing model for speech, we have used two different combinations of datasets, i.e. RAVDESS and TESS and accomplished 87.08% accuracy using CNN. For the facial expression model, we have used the FER 2013 dataset using a transfer learning algorithm, and the model reached an accuracy approx. 99% over seven classes.
Keywords: Emotion recognition, Speech, Video, Deep Learning.

1 Introduction

Feelings are a fundamental piece of human communication. Most of the communication takes place using emotion. People are fit for delivering many feelings during correspondence that differ in intricacy, power, and importance. To perceive the right emotion, an accurate interpretation of these emotions is essential. This system will help maintain relationships in our society or social contact, and perceived emotion will help us understand human emotion or his/her activities. Expressive speech, facial gestures, body language, and other non-verbal and verbal ways for people to communicate their sentiments are only a few examples [1,2]. As a result, a multi-modal model may accurately anticipate human emotions.Notwithstanding, the single-modular model can only, with significant effort, anticipate the feelings of people. Hence, a multi-modal model is the best way to understand emotion. Affective computing is a branch of machine learning and computer science that looks at identifying and explaining human effects. It is already in use in the fields of smart cards, education, and automobiles, as well as entertainment and healthcare. This holds for human-computer interface design, intelligent robots, and safety control, among other things. Based on the most current system, deep learning [3, 4] has had significant success in many areas over the last several years. This subject is gradually gaining attention since it gives a rich probability in theoretical and practical domains. It would be exciting to use objective computer tools to verify psychological hypotheses [20].Here, we present a survey of the new exploration in emotional recognition frameworks utilizing multi-modular signs and characterization philosophies utilizing profound learning. This survey means approaching and improving continuous human expression recognition through video and audio and providing the most recent advancement in this innovation.

This paper is divided into four sections such that section 2 briefly describes the work related to both facial and speech emotion recognition separately. Section 3 describes the work related to multi-modal expression recognition. Section 4 presents different models and classifier. Section 5 presents our paper methodology.Section 6 shows the experimental results of the model. Section 7 provides the future scope of the project.Section 8 concludes this paper.

2 Expression Recognition and Deep Learning

2.1 Classification of emotions

The communication between individuals, the feeling is the fundamental approach to conveying. The look and voice regularly have the passionate condition of the speaker. For people, it is a very little challenging to distinguish the passion condition of the speaker, however, let the PC get the speaker. The enthusiastic state is a difficult matter. To perceive the enthusiastic condition of the speaker, the PC, as a matter of first importance, instate both the look and discourse sign of the speaker, brings outs the passionate elements of the sign, and afterward makes a specific face and discourse passionate acknowledgment model [21]. At last, the inclination class of the face and discourse is set by some arrangement strategies. The human articulation acknowledgment

framework is primarily involved signal handling, including extraction, and feeling acknowledgment. At long last, the regularly utilized signal acknowledgment model is presented.

2.2 Facial Expression Recognition

Human beings share seven universal emotions. Neutral, disgust, pleasure, anger, fear, sorrow, and surprise are the basic emotions, and human facial expressions can be used to classify them. It proved that humans, regardless of their cultural backgrounds, can discern these feelings. It is a tough job to understand facial expressions as each face varies from person to person. There are a lot of factors affecting the features like physical characteristics as well as sex, genes, and age [5]. 7% of information moves among individuals through composition, 38% through voice, and 55% through look [6]. Feelings could be communicated utilizing two symmetrical aspects: valence and excitement. It expresses that different individual has a different method for conveying their sentiments. People groups' feelings might contrast when requested to communicate occasional feelings. The enthusiasm might be calm to energize, and the valence can be good or negative. [7]. A feeling acknowledgment gadget in light of mouse-type equipment had revealed by analysts at IBM. Picard and partners at the MIT Media Laboratory have been applying their endeavors to execute a viable PC since the last part of the 1990s. Even though they outline the practicability of a physiological sign-based feeling acknowledgment framework, a few highlights of its exhibition are expected to be upgraded before it very well may be utilized as a viable framework. In the first place, their calculation advancement and execution tests were achieved with information that reflects deliberately communicated feelings. The utilization of a Deep Neural Network model to confront appearance examination is an interesting issue in facial acknowledgment right now [9]. Facial expression recognition models based on the Xception model, transfer learning using Keras, and TensorFlow with OpenCV [8].

2.3 Speech Emotion Recognition

In recent years, experts working in the fields of pattern recognition and speech signal processing have been increasingly interested in emotion recognition via speech. Emotion identification is critical for distinguishing a speaker's emotional state from the audio stream. Emotional speech recognition can automatically deduce a person's emotional or physical status from his or her voice. Emotional aspects of speech refer to a speaker's emotional or physical state and are included in the so-called paralinguistic aspects of communication. A message-free technique for discourse feeling order, which applies a brief time frame log recurrenceconversation signals is addressed using low-frequency power coefficients (LFPC), and a Hidden Markov Model (HMM) is used as a classifier. The author proposes a paradigm for describing the passionate state of expressions. An average accuracy of 78% was achieved & LPCC and MFCC were used to analyze the performance. Also follows the fact that system performance can be enhanced by grouping the emotions with the same characteristics [11]. The meaning of brain science and phonetics in communication in language man-machine. interfaces. It likewise needs mental and etymological investigation, going with the procedures in signal handling and examination. Average accuracy of 77% is attained. Fear is tough to classify whereas anger and neutral can be classified easily [10]. Exploratory investigation through boundaries like principal recurrence, formant recurrence, and measurable examination was directed toward a multi-facet brain organization (MLNN). The ordinary accuracy achieved through this procedure is 75.93% for multi-word sentences while that for single-word sentences is 81.67% [12].

3 Literature Survey

Multi-modular Emotion Recognition is an emerging field that focuses on incorporating both audio and video inputs to enhance emotion recognition. Deep Learning (DL) architectures have played a significant role in advancing this area, with approaches like profound conviction nets, profound Convolutional brain networks LSTM, and support vector machines (SVM) being utilized. One approach, known as the multi-modular consideration organization (MMAN), integrates visual and textual cues for discourse emotion recognition. To model linguistic emotions, a pre-trained "BERT-like" architecture for self-regulated learning (SSL) is employed [14]. Another method involves extracting acoustic features using techniques such as the SincNet layer and band-pass filtering, followed by applying these features to a DCNN [15]. Additionally, a recent approach incorporates a convolutional brain network with cross-sequence attention for waveform-based emotion recognition. In terms of voice analysis for consumer evaluations, SVM-based AI models have demonstrated reliability, leading to the proposal of an SVM-based multi-dimensional speech emotion recognition method [19].

Table 1. Combining signals from speech, text and video

Author	Neural Network architecture and deep learning algorithm used	Accuracy	Year	Dataset used
[13]	Long Short Term Memory(LSTM)	73.98%	2020	IEMOCAP
[14]	Self-supervised Learning(SSL model)	----	2020	IEMOCAP, CMUMOSEI, CMUMOSI),
[15]	Deep Convolutional Network (DCCN), Recurrent Neural Network(RNN)	80.51%	2020	IEMOCAP
[18]	1D Convolutional Neural Network (CNN)	71.85%	2020	IEMOCAP
[19]	Support Vector Machine (SVM)	72.52%	2019	Berlin Emotional DB
[16]	Convolutional Neural Network(CNN)	----	2020	Asian Character from the TV

				drama series
[17]	Long Short Term Memory(LSTM)	79.7%	2020	777songs(Music Mood Classification Data Sets)

A significant learning model that incorporates multiple measures, including facial images and text-based cues, is employed to analyze emotions in the context of the Korean TV series "Misaeng: The Incomplete." The model categorizes the facial images of the characters into seven emotions: Anger, Disgust, Joy, Surprise, Fear, Sadness, and Neutral. Two multi-modal models, utilizing both images and textual data, were developed to capture emotions in this context [16].In the domain of music, a novel multi-modal framework for music emotion analysis was proposed, considering both the audio quality of music and its accompanying lyrics. The framework utilizes LSTM networks for audio feature extraction and demonstrates superior performance compared to other AI methods in terms of emotion recognition [17].

4 Model and Classifier

Our goal is to develop the grouping model utilizing AI calculations to perceive the enthusiastic state. Among all approaches we have, the now and again utilized arrangement calculations are the Maximum Likelihood Model (MLB), Artificial Neural Network(ANN), Support Vector Machine (SVM), and Xception model.A few different classifiers that need a reference here are Fuzzy Classifier, Decision Tree, LDC (Linear Discriminant Classifier), and some more. How about we momentarily examine a couple of classifiers utilized in feeling acknowledgment.

4.1 Multinomial Naive Bayes

The multinomial naive Bayes algorithm applies the Bayes theorem: it is based on the preferably strong assumption that in the context of classification, every characteristic is independent of the others. This classifier will always result in the category with the highest priority probability using the Bayes theorem. This algorithm has a simple and intuitive design and is a good benchmark for classification purposes.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B)Marginal Probability: Probability of Evidence.

Author	Year	Based on	Dataset	Accuracy
[22]	2019	Text	Bangla Text Corpus	78.6%
[23]	2018	Speech	The Arabic Natural Audio Dataset (ANAD)	85.27%
[24]	2019	Text	Facelecture posts from official diabetes support groups	82%

Table 2. Combining signals based on multinomial naive Bayes

4.2 Support Vector Machine

This technique doesn't zero in on potential outcomes however centers around making a discriminant work f: X −→ y. The instinct of SVM in the directly detachable case is to place a line in two classes so the distance to the closest certain or negative occurrence gets expanded. It is vital to take note that this overlooks the class circulation P(X|y). The SVM discriminant work has the structure:

$$f(X) = w\,T\,x + b\ 16(1)$$

The order rule is sign(f(X)), and the straight choice limit is determined by f(x) = 0. On the off chance that f isolates the information, the numerical distance between a point x and as far as possible is (yf(X))/(||w||).
Given preparing information, the point is to observe a choice limit w, b that expands the mathematical distance of the nearest point. The advancement objective is accordingly:

$$max\ (w,b)\ min\ (i=1)n = yi(wT\ xi + b)/\ ||w|| \qquad (2)$$

This improvement goal can be re-composed with an extra requirement, considering the way that the intention is no different for k w ˆ, k ˆb for any non-zero scaling factor k:

$$min\ (w,b) \quad \tfrac{1}{2}||w||2 \qquad (3)$$

Table 3. Combining signals based on support vector machine

Year	Author	Based on	Dataset	Accuracy
2019	[25]	Speech	Chinese speech emotion corpus	83.75%
2020	[26]	Video	JAFFE database, CK+ database, and FG net database	____
2019	[19]	Speech	Berlin Emotional DB	72.52%

4.3 Convolution Neural Network

In the field of computer vision, convolutional neural networks (CNNs) have become the new standard. CNNs are specialized types of neural networks designed to process data with a grid-like structure, such as images. In traditional SVM approaches, a significant part of the work involved selecting and designing filters (e.g., Gabor filters) to extract as much relevant information from the image as possible. However, with the advancements in deep learning and increased computational capabilities, this process can now be automated. The term "convolutional" in CNNs refers to the process of convolving the input image with a set of filters. The main challenge lies in determining the number and characteristics of these filters. The characteristic of a filter is referred to as the stride length, which typically falls within the range of 2 to 5. In essence, we are constructing a convolved output that has a volume, no longer a 2-dimensional image. The filters themselves may not be easily interpretable by humans, especially when a large number of them are utilized. Some filters are designed to recognize curves, while others capture edges or textures. Once this volume is extracted, it can be flattened and passed into a dense neural network for further processing.

Mathematically, the convolution operation is expressed as:

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)\, g(t - \tau)\, \partial\tau \quad (1)$$

The convolution addresses the level of the region of the channel g that covers the information f at time τ in general time t. In any case, since $\tau < 0$ and $\tau > t$ have no importance, the convolution can be decreased to:

$$(f * g)(t) = \int_{0}^{t} f(\tau)\, g(t - \tau)\, \partial\tau \quad (2)$$

At each convolutional step, an activation function is applied to each input, typically using the Rectified Linear Unit (ReLU) activation function. This process introduces additional dimensionality to the original image input. To reduce the dimensionality, a pooling step is performed. Pooling involves down sampling the features, aiming to learn fewer parameters during training. The most common type of pooling is max-pooling. For each element of the input images, a maximum pooling operation is applied over a specified height and width, usually 2x2. This operation selects the maximum value among the four pixels in the pooling region. The intuition behind max-pooling is that the maximum value is more likely to be significant when recognizing an image. With these convolutional and pooling operations, we have described the elements of a convolutional neural network:

- The convolution layer
- The enactment
- The pooling layer
- The completely associated layer, like a thick brain organization

The request for the layers can be exchanged:

ReLU(MaxPool(Conv(X)))=MaxPool(ReLU(Conv(X)))

Table 4. Combining signals based on convolution neural network

Year	Author	Based on	Dataset	Accuracy
2021	[27]	Images	UMD Faces dataset	The precision of about 95%
2021	[18]	Video	IEMOCAP	71.85%
2020	[28]	Speech	IEMOCAP	77.01%

The above arrangement models enjoy a few benefits and impediments as per their application region. For Different boundaries (discourse, video) various classifiers are prevalent.

5 Experimental Analysis and Methodology

5.1 Experiment Setup

There are many open source and licensed tools available to perform face recognition such as image pre-processing, normalization, feature selection, and classification. We used open source deep learning/machine learning along with Python 3.6 scripting, Kerason the top of Tensor flow libraries to perform hybrid feature selection and classification approaches to get optimal results on benchmark datasets FER 2013dataset and RAVDESS and TESS dataset.

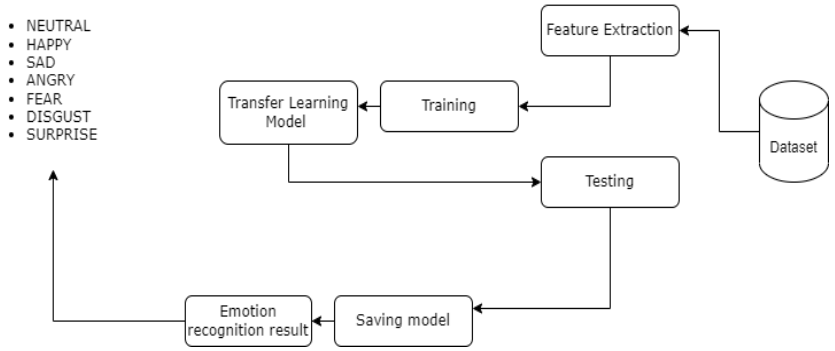

Fig.1 Proposed Methodology for Experimental Analysis.

5.2 Source Dataset

The experimental results of human expression recognition using images or video and using speech are based on FER 2013dataset and RAVDESS and TESS dataset respectively. In FER 2013 dataset, there are 28709 images in the train set, and 7178 images in the set of tests. Each image's data set comprises the grayscale color of 2304 pixels (48x48) as well as the emotion associated with it, such as anger, disgust, fear, happiness, neutrality, sadness, and surprise whereas in RAVDESS and TESS datasetRAVDESS and Tess dataset, we used 5252 samples from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and the Toronto emotional speech set (TESS) dataset to recognize speech elicits emotion.RAVDESS has 1440 speech files and 1012 song files.TESS has 2800 files.

Table 5: Data Collections

Dataset	Description	Used for
FER 2013 Dataset	The train set has 28709 images. The test set has 7178 images. Total 35887 images in a dataset. Gender: male and female 48 x 48 pixels. Emotions include neutral, happy, sad, angry, fear, surprise, and disgust.	Facial expression recognition using images.
RAVDESS and TESS Dataset	Total 5252 samples. Combining both RAVDESS (speech and songs) and TESS dataset Emotions include neutral, calm, happy, sad, angry, fearful, surprise, and disgust.	Speech emotion recognition.

5.3 Feature Extraction

In facial expression recognition, the image of the face is used to extract the relevant features. Scale, pose, level translation, and differences in illumination are all intrinsic issues in picture classification whereas in recognizing emotion using speech, the extraction of elements is a basic advance in assessing and finding connections between different items. We realize that the sound information given by the models can't be deciphered straight by the models, so we want to transform it into an arrangement that the models can comprehend, which is where include extraction comes in. We are using MFCC, Chroma, stft and MelSpectogramto train our model.

5.4 Model Evaluation

Evaluation of the Model can be done using the following methods:

a) Precision

Precision evaluates the algorithm's predictive capacity by estimating the predictive value of a label, which can be positive or negative depending on the class for which it is derived. The percentage of correctly assigned expressions with the total number of aspects is known as precision.

$$\text{precision} = \frac{tp}{tp+fp} \qquad (1)$$

b)Recall

Recall is a function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Recall is the percentage of correctly assigned expressions in relation to the total number of expressions.

$$\text{recall} = \frac{tp}{tp+fn} \qquad (2)$$

c) F-score

F-score is a composite measure which benefits algorithms with higher sensitivity and challenges algorithms with higher specificity. The F-score is evenly balanced when $\beta = 1$. It favors precision when $\beta > 1$, and reviews in any case.

$$F - measure = \frac{(\beta^2+1)*precision*recall}{\beta^2*precision*recall} \quad (3)$$

6 Experimental Analysis

In the experiments, we used dataset for human expression recognition using images or video and using speech are based on FER 2013 dataset and RAVDESS and TESS dataset respectively. For facial expression recognition, transfer learning classification is performed, we set thekeras application to MobileNetV2 and parameter values loss= sparse_categorical_crossentropy, optimizer= adam, metrics= accuracy, activation= Relu, epochs=5. Similarly, for recognizing emotions through speech we used three different classifications i.e. multilayer perceptron, decision tree classifier, convolutional neural network and choose the best model for system to predict emotions. In multilayer perceptron classification, we set the parameter values alpha=0.01, batch_size=256, epsilon=1e-08, hidden_layer_sizes= 300, learning_rate=adaptive, max_iter=500. In decision tree classifier, we mportDecisionTreeClassifier from sklearn. tree. In convolutional neural network, we set the parameter values optimizer= rmsprop, loss=sparse_categorical_crossentropy, metrics= accuracy,activation= Relu, batch_size=16, epochs=200, epsilon=1e-07.

6.1 Experimental results of facial expression recognition model using images

Deep learning based transfer learning model achieved a model accuracy of approx. 0.99.

```
In [60]:  1 historymodel=new_model.fit(X,Y,epochs=5) ##training

Epoch 1/5
100/100 [==============================] - 1163s 10s/step - loss: 0.0412 - accuracy: 0.9896
Epoch 2/5
100/100 [==============================] - 847s 8s/step - loss: 2.8126e-06 - accuracy: 1.0000
Epoch 3/5
100/100 [==============================] - 812s 8s/step - loss: 5.5272e-07 - accuracy: 1.0000
Epoch 4/5
100/100 [==============================] - 812s 8s/step - loss: 2.8947e-07 - accuracy: 1.0000
Epoch 5/5
100/100 [==============================] - 812s 8s/step - loss: 2.0406e-07 - accuracy: 1.0000
```

Fig 2. Transfer learning model reached the accuracy approx. 99% over 5 epochs

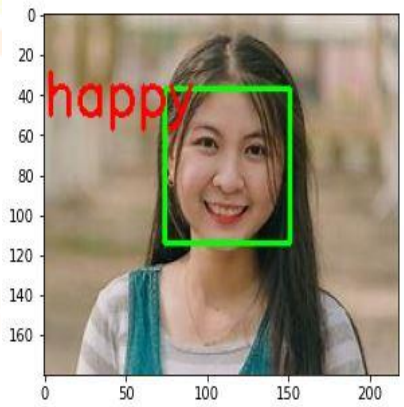

Fig 3. The Deep face model predicted the emotion Happy

6.2 Experimental results of speech emotion recognition model

After implementing all three different classifications, we found that MLP classifier achieved an F1 score of 0.82 over the 8 classes, Decision tree classifier achieved an F1 score of 0.65 and CNN model that obtained an F1 score of 0.87. So our final choice model that we choose for predicting emotions through speech would be CNN model.

Table 6. Result on the CNN model on the test set per each class.

	Precision	**Recall**	**F1 Score**
0	0.97	0.89	0.92
1	0.70	0.91	0.79
2	0.89	0.83	0.86
3	0.91	0.82	0.86
4	0.92	0.90	0.91
5	0.79	0.91	0.85
6	0.89	0.87	0.88
7	0.88	0.87	0.87
Accuracy			0.87
Macro avg	0.87	0.87	0.87
Weighted avg	0.88	0.87	0.87

Table 7.F1 score for each class compared to the baselines (MLP, Decision tree classifier, CNN)

Class	MLP	Decision tree Classifier	CNN
Angry	0.90	0.80	0.92
Calm	0.75	0.60	0.79
Disgust	0.81	0.56	0.86
Fearful	0.83	0.65	0.86
Happy	0.82	0.69	0.91
Neutral	0.88	0.63	0.85
Sad	0.73	0.62	0.88
Surprised	0.80	0.66	0.87

7 Future Scope

Human expression recognition methods have improved a lot over the past decade. The focus has shifted from posed expression recognition to spontaneous expression recognition. Promising results can be obtained under face registration errors, fast processing time, and high correct recognition rate (CRR) and significant performance improvements can be obtained in our system. The model can work with the images feed. It can recognize spontaneous expressions. Our model can be used in Digital Cameras wherein the image can be captured only when the person smiles. Security systems can identify a person in any form of expression he presents himself. Rooms in homes can set the lights and television to a person's taste when they enter the room. Doctors can use the system to understand the intensity of pain or illness of a deaf patient. Our system can be used to detect and track a user's state of mind, and in mini-marts and shopping centers to view the customers' feedback to enhance the business.

8 Conclusion

The accurate assessment and depiction of emotions present several challenges. Researchers in the field are exploring various combinations of features that influence emotion recognition. In this study, the authors utilized seven distinct emotions from the FER2013 dataset, as well as a combination of RAVDESS and TESS datasets. The process involved preprocessing facial images captured from individuals, extracting features, and classifying emotions based on the FER2013 dataset using transfer learning. The study also involved preprocessing audio data, extracting relevant features, and classifying speech emotions based on a combination of the RAVDESS and TESS datasets using a Convolutional Neural Network (CNN). The dataset used for training and testing purposes was divided into separate sets for each task. The transfer learning model achieved an accuracy of approximately 99% for facial emotion recognition across the seven classes. For speech emotion recognition, the multi-layer perceptron, decision tree classifier, and convolutional neural network achieved accuracies of 81.72%, 65%, and 87.08% respectively. Overall, this study demonstrates the effectiveness of utilizing transfer learning for facial emotion recognition and employing different classifiers for speech emotion recognition, providing promising results in both domains.

9 References

[1] A. Clark, S. Abdullah, and S. Ameen, "A comparison of decision feedback equalizers for a 9600 bit/s modem," Journal of the Institution of Electronic and Radio Engineers, vol. 58, pp. 74-83, 1988.
[2] S. Ammen, M. Alfarras, and W. Hadi, "OFDM System Performance Enhancement Using Discrete Wavelet Transform and DS-SS System Over Mobile Channel," ed: ACTA Press Advances in Computer and Engineering, 2010. [3] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition," in 2019 8th International Conference on Affective Computing and Intelligent Interaction (ASCII), 2019, pp. 552-558.
[4] K. B. Obaid, S. Zeebaree, and O. M. Ahmed, "Deep Learning Models Based on Image Classification: A Review," International Journal of Science and Business, vol. 4, pp. 75-81, 2020.
[5] P. Ekman and W. V. Friesen, Unmasking the face: A guide to recognizing emotions from facial clues: Ishk, 2003.
[6] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," Procedia Computer Science, vol. 175, pp. 689-694, 2020.
[7] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience," Journal of personality and social psychology, vol. 69, p. 153, 1995.
[8] T. D. Nguyen, "Multimodal emotion recognition using deep learning techniques," Queensland University of Technology, 2020.
[9] S. Dou, Z. Feng, X. Yang, and J. Tian, "Real-time multimodal emotion recognition system based on the elderly accompanying robot," in Journal of Physics: Conference Series, 2020, p. 012093
[10] Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.
[11] Tin Lay New, Say Wei Foo and Liyanage C. De Silva. Speech emotion recognition using Hidden Markov Models. Speech Communications 41, 603-623, 2003.
[12] Jana Tuckova and Martin Sramka. Emotional speech analysis using Artificial Neural Networks. Proceedings of the International Multiconference on Computer Science and Information Technology, 141-147, 2010.
[13] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal Attention for Speech Emotion Recognition," arXiv preprint arXiv:2009.04107, 2020.
[14] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," arXiv preprint arXiv:2008.06682, 2020.

[15] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention Driven Fusion for Multi-Modal Emotion Recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020, pp. 3227-3231.
[16] J.-H. Lee, H.-J. Kim, and Y.-G. Cheong, "A Multi-modal Approach for Emotion Recognition of TV Drama Characters Using Image and Text," in 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), 2020, pp. 420-424.
[17] G. Liu and Z. Tan, "Research on Multi-modal Music Emotion Classification Based on Audio and Lyirc," in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 2331-2335.
[18] D. Krishna and A. Patil, "Multimodal Emotion Recognition using Cross-Modal Attention and 1D Convolutional Neural Networks," Proc. Interspeech 2020, pp. 4243-4247, 2020.
[19] C. Caihua, "Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM," 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2019, pp. 173-176, doi: 10.1109/ICPICS47731.2019.8942545.
[20] E. C. -C. Kao, C. -C. Liu, T. -H. Yang, C. -T. Hsieh and V. -W. Soo, "Towards Text-based Emotion Detection A Survey and Possible Improvements," 2009 International Conference on Information Management and Engineering, 2009, pp. 70-74, doi: 10.1109/ICIME.2009.113.
[21] Kotti M, Stylianou Y. Effective emotion recognition in movie audio tracks[C]//2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2017: 5120-5124.
[22] S. Azmin and K. Dhar, "Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), 2019, pp. 1-5, doi: 10.1109/EICT48899.2019.9068797.
[23] Zantout, Rached (2018). [Advances in Biochemical Engineering/Biotechnology] || Ensemble Models for Enhancement of an Arabic Speech Emotion Recognition System..(Chapter 15), 174–187. doi:10.1007/978-3-030-12385-7_15
[24] Balakrishnan, V., & Kaur, W. (2019). String-based Multinomial Naïve Bayes for Emotion Detection among Facelecture Diabetes Community. Procedia Computer Science, 159, 30–37. doi:10.1016/j.procs.2019.09.15
[25]Sun, Linhui, Sheng Fu, and Fu Wang. "Decision tree SVM model with Fisher feature selection for speech emotion recognition." *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1 (2019): 1-14.
[26] Kumar, R., M. Sundaram, and N. Arumugam. "Facial emotion recognition using subband selective multilevel stationary wavelet gradient transform and fuzzy support vector machine." *The Visual Computer* 37.8 (2021): 2315-2329.
[27]Said, Yahia, and Mohammad Barr. "Human emotion recognition based on facial expressions via deep learning on high-resolution images." *Multimedia Tools and Applications* 80.16 (2021): 25241-25253.
[28]Anvarjon, Tursunov, and Soonil Kwon. "Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features." *Sensors* 20.18 (2020): 5212.