A Brief Review of Facial Emotion Recognition using Deep Learning Techniques

Priti Singh, Hari Om Sharan, C.S. Raghuvanshi
Faculty of Engineering and Technology, Rama University, Uttar Pradesh, Mandhana, Kanpur
preetirama05@gmail.com

Abstract:

Humans have traditionally had an easy time detecting emotions from facial expression, but performing the same feat with a computer programme is rather difficult. It is now possible to discern emotions from photos because to recent advances in computer vision and machine learning. This study provides an overview of the current phases, methodologies, and datasets for Facial Emotion Recognition (FER). For decades, FER has been identified, and it is a critical topic in the disciplines of computer vision and machine learning. Automatic FER is beneficial in a wide range of applications, including healthcare, safety, education, criminal investigation, and the Human Robot Interface, among others. This study provides a review of the most effective strategies offered in recent years, as well as a brief introduction to the system and database that were employed. For the identical problem mentioned in this document, several writers have utilised alternative algorithms. Various approaches of emotion recognition are reviewed and contrasted in this review paper. The goal of this paper is to review recent research on automatic facial emotion recognition (FER) using a convolutional neural network. We highlight the contributions that have been addressed, as well as the architecture and databases that have been employed, and we illustrate the progress made by contrasting the recommended approaches with the results acquired. The purpose of this paper is to assist and guide scholars by reviewing recent work and offering suggestions for how to enhance this field.

Keywords: facial emotion recognition, convolutional neural network, Deep learning, automatic recognition, database

1. Introduction:

In today's world, facial expression recognition has become a critical issue in a variety of applications. In recent years, there has been a lot of research on face emotion recognition. The goal of facial emotion detection is to use certain face images to help identify the state of human emotion (e.g., neutral, happy, sad, surprise, fear, wrath, disgust, contempt). The goal of facial emotion recognition is to recognize facial emotion states accurately and automatically. As a result, determining the similarity of the same emotion state amongst different people can be difficult because they may display the same emotion state in different ways. For instance, depending on the individual's mood, skin colour, age, and surrounding environment, the expression may change.

Face detection, feature extraction, and emotion classification are the three key stages of FER. An image of a face is detected in the first stage, which is a preprocessing stage, and facial components of the face are detected from the region. Eyes, brows, nose, and mouth are examples of face components. The second stage will extract informative features from various regions of the face. A classifier must be trained before being used to generate labels for the Emotions using the training data in the final stage.

Mehrabian [1] found that visual information accounted for 55% of emotional information, audio information for 38%, and verbal information for 7%. Human emotional states can be determined using verbal and nonverbal data collected by various sensors, such as face changes [2], voice tone [3], and physiological signs [4]. Face alterations during communication are the first evidence of emotional state transmission, which is why this modality has piqued the interest of most studies.

When compared to other modalities to statistics, the automatic FER is the most investigated by researchers, according to Philipp et al.[5], but it is a difficult process because each person expresses his or her emotion in their own unique way. There are several obstacles and challenges in this area that should not be overlooked, such as variations in head pose, luminosity, age, gender, and background, as well as the problem of occlusion caused by Sunglasses, scarfs, skin illnesses, and so on.

To improve classification, extracting traits from one face to another is a complex and delicate operation. In 1978, Ekman and Freisen [6] were among the first scientists to study facial expression and develop FACS (Facial Action Coding System), a system in which facial movements are described by Action Units AUs. The human face is divided into 46 AUs, each of which is coded with one or more facial muscles.

Deep learning has been a very successful and efficient approach in recent years, thanks to the results obtained with its architectures, such as the convolutional neural network CNN and the recurrent neural network RNN, which allow for the automatic extraction of features and classification; this is what prompted researchers to start using this technique to recognize human emotions. Researchers have made several efforts to construct deep neural network topologies, which have yielded highly promising findings in this area.

Provide a summary of current improvements in perceiving emotions through facial expression recognition using various convolutional neural network designs in this paper. Present new findings from 2016 through 2020, together with an analysis of the issues and contributions. The following is how it's laid out: We describe some publicly available databases in section two, give a recent state of the art on the FER utilizing deep learning in section three, and conclude in sections four and five with a discussion and comparisons, followed by a general conclusion with future efforts.

2. Facial Emotion Recognition databases

The training of the neuron network with examples is one of the success factors of deep learning, and several FER databases are now available to researchers to help them do so. Each one differs from the others in terms of the number and size of images and videos, variations in illumination, population, and face pose.

Some of databases are given below:

(1) Multi-PIE Facial Expression Database: The Multi-PIE [7] (Multi Pose, Illumination, Expressions) face database contains more than 750,000 images captured of 337 subjects taken under different pose, illumination and expressions. The pose of imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions. The database contains more than 305 GB of face data. Multi-PIE Facial emotions contained in the database are Anger, Disgust, Neutral, Happy, Squint, Scream, and Surprise**.**

(2) MMI Facial Expression Database: Over 2900 videos and high-resolution still images of 75 subjects cover the MMI Facial Expression Database [8]. It's fully coded for the presence of AUs in videos (event coding) and partially coded on a frame-by-frame basis, indicating whether an AU is in the neutral, onset, apex, or offset phase for each frame. Audio-visual laughters were annotated in a short section. Happiness, sadness, anger, disgust, fear, and surprise are six universal emotions represented in the MMI facial expressions.

(3) GEMEP FERA Facial Expression Database: There are 289 images sequences in the GEMEP FERA [9] face database. GEMEP FERA Facial emotions contained in the database are Anger, Fear, Sadness, Relief, and Happy.

(4) SFEW (Static Facial Expression in the Wild) Facial Expression Database: In a person-independent approach, the database was separated into equal-sized sets. SFEW [10] is a collection of 700 images that have been labeled for six basic emotions: anger, disgust, fear, happy, sad, surprise, and neutral.

(5) CK+ (Extended Cohn-Kanade dataset) Facial Expression Database: The Extended Cohn-Kanade (CK+) [11] dataset contains 593 video sequences from 123 distinct people ranging in age from 18 to 50 years old, as well as gender and ethnicity. Anger, contempt, disgust, fear, happiness, sadness, and surprise were among the seven expression classes in the extended Cohn-Kanade database.

(6) FER-2013 Facial Expression Database: FER-2013 [12] contains 35,887 facial RGB images of various emotions with a size restriction of 48X48 pixels, and the main labels can be grouped into seven categories: 0 means angry, 1 means disgust, 2 means fear, 3 means happy, 4 means sad, 5 means surprise, and 6 means neutral. The Disgust expression has the fewest images – 600 – compared to the other categories, which each have around 5,000.

(7) JAFFE (Japanese Female Facial Expression) Facial Expression Database: The JAFFE [13] dataset contains 213 images of 10 different Japanese female participants with various face expressions. Each subject was requested to make 7 facial expressions (6 basic and one neutral), and the photographs were annotated by 60 annotators with average semantic evaluations for each facial emotion.

(8) BU-3DFE Facial Expression Database: Although 3D facial models have been widely employed for 3D face identification and 3D face animation, their utility for 3D facial emotion recognition remains uncertain. To help researchers in this field, we built the BU-3DFE database, a 3D facial expression library comprising 100 people and 2500 facial expression models. Six universal emotions are represented in the BU-3DFE facial expressions: happiness, sadness, anger, disgust, fear, and surprise [14].

(9) CASME II Facial Expression Database: There are 195 micro-expressions captured at 60 frames per second in the Chinese Academy of Sciences Micro-expression (CASME) database [15]. They were chosen from a pool of over 3000 evoked facial motions, with 247 micro-expressions included in the database. The onset, apex, and offset frames, as well as action units (AUs) and emotions, were all coded in these samples. Happy, Disgust, Surprise, Regression, and other CASME face emotions are included in the database.

(10) Oulu-CASIA Facial Expression Database: The Oulu-CASIA [16] collection contains 2880 images from 80 people representing six emotions: happy, surprise, disgust, fear, sadness, and anger; the majority of the subjects are males aged 23 to 58. This database was created specifically to address the issue of lighting due to environmental changes. It comprises of two imaging systems: one that uses near infrared (NIR) and another that uses visible light (VIS).

(11) AffectNet Facial Expression Database: AffectNet [17] is a huge facial expression dataset with roughly 0.4 million images manually tagged for the presence of eight facial expressions (neutral, happy, angry, sad, fear, surprise, disgust, contempt) as well as valence and arousal intensity.

(12) RAF-DB Facial Expression Database: The Real-world Affective Faces [18] Collection (RAF-DB) is a large-scale face expression database that contains over 30K great-diverse facial images that were downloaded from the Internet. Each image has been tagged separately by roughly 40 annotators using crowd sourcing annotation. The participants' age, gender, and ethnicity, as well as head postures, lighting conditions, occlusions (e.g. spectacles, facial hair, or self-occlusion), and post-processing processes (e.g. various filters and special effects), are all represented in this database. Happiness, sadness, anger, disgust, fear, and surprise are six universal emotions represented in the RAF-DB facial expressions.

(13) RaFD Facial Expression Database: The Radboud Faces Database (RaFD) is a collection of 67 photographs of models with eight different emotional expressions. The RaFD is a high-resolution face database that includes images of eight different emotional expressions. Each model was trained to portray the following expressions: anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral, according to the Facial Action Coding System [19].

3. Facial emotion recognition using convolutional neural network

Despite the widespread success of classical facial recognition systems based on the extraction of handmade features, researchers have turned to deep learning in the last decade because of its great automatic recognition capability. In this regard, we will describe some new FER researches that demonstrate proposed deep learning strategies for enhanced detection. Train and test on a variety of databases, both static and sequential.

Lopes et al. [20] investigated the impact of data pre-processing prior to network training in order to improve emotion classification. Before CNN, which consists of two convolution-pooling layers finishing with two fully linked with 256 and 7 neurons, data augmentation, rotation correction, cropping, down sampling with 32x32 pixels, and intensity normalization were performed. At the test stage, the best weights achieved throughout the training stage are employed. Three databases were used to evaluate this experience: CK+, JAFFE, and BU-3DFE. Researchers have discovered that combining all of these pre-processing processes is more successful than doing it individually.

Cai et al. [21] proposed a novel architecture CNN with Sparse Batch normalisation SBP for the disappearance or explosion gradient problem in 2018. This network has the property of using two convolution layers in succession at the beginning, followed by max pooling and SBP, and the dropout applied in the midst of three fully connected layers to reduce the over-fitting problem. For the face occlusion problem, Li et al. [22] propose a new CNN method in which the data is first introduced into a VGGNet network, and then the CNN methodology with the attention mechanism ACNN is applied. The FED-RO, RAF-DB, and AffectNet databases were used to train and test this architecture.

Using the FER2013 database, Agrawal et Mittal [23] investigate the impact of changing CNN parameters on recognition rate. To begin, all of the photographs are 64x64 pixels in size, with varying numbers of filters and sizes. On a simple CNN, which contains two successive convolution layers, the second layer plays the job of max pooling, and then a softmax function for classification, the type of optimizer chosen (adam, SGD, adadelta) is also important. According to these investigations, researchers developed two innovative CNN models that reach an average accuracy of 65.23 percent and 65.77 percent. The unique feature of these models is that they do not comprise fully connected layers dropout, and the network retains the same filter size.

Kim et al. [24] investigates changes in facial expression as a function of emotional state and proposes a spatio-temporal architect that combines CNN and LSTM. The spatial features of the facial expression are first learned by CNN in all frames of the emotional state, and then an LSTM is used to store the entire sequence of these spatial features.

Yu et al. [25] also provide a unique architecture known as Spatio-Temporal Convolutional with Nested LSTM (STC-NLSTM), which is built on three deep learning subnetworks: 3DCNN was used to extract spatiotemporal features, then temporal T-LSTM was used to maintain the temporal dynamic, and finally convolutional C-LSTM was used to model the multi-level features.

The basic emotions were previously classified by all of the researchers as follows: happy, disgust, surprise, anger, fear, sadness, and neutral.

4. Comparison of Current Research Papers on Facial Emotion Recognition via Convolutional Neural Network

Researchers have shown a strong interest in FER using convolutional neural networks in recent years, according to this publication. The automatic FER task includes several processes, including data processing, model design proposal, and emotion recognition.

Preprocessing is an important step that was present in all of the papers cited in this review and includes techniques like resized and cropped images to reduce training time, normalisation spatial and intensity pixels, and data augmentation to increase image diversity and eliminate the over-fitting problem. Lops et al. [20] describe all of these strategies in a clear and concise manner.

Researchers presented alternative convolutional neural network structures for extracting spatio-temporal information, such as a mix of CNN-LSTM, 3DCNN, and a Deep CNN. The methods presented by Yu et al. [25] produce better precision than the one employed by Kim et al. [24] based on the findings obtained. With a success record of more than 99 percent.

The correctness of several methodologies and contributions provided in this review was high. Li et al. [22] are interested in studying the problem of occlusion photos, as well as getting a better understanding of the network. For the benefit of Agrawal and Mittal. [23] After a thorough examination of the impact of CNN parameters on recognition rate, two new CNN architectures are presented. The majority of these approaches produced competitive results in excess of 90% of the time.

Researchers attain great precision in FER by using CNN networks with spatial data, and for sequential data, they employed a mix of CNN-RNN networks, particularly the LSTM network, indicating that CNN is the deep learning network of choice for FER. The Softmax function and the Adam optimization technique are the most commonly employed CNN parameters by researchers.

Table 1: Comparison of current research Work on FER via CNN

Paper	Expressions	Authors	Databases	Architecture used	Rate of Recognition
20	happiness, sadness, anger, disgust, fear, and surprise	Lopes et al.[22]	CK+[11], JAFFE[13], BU-3DFE[14]	CNN	96.76% for CK+
21	Anger, contempt, disgust, fear, happiness, sadness, and surprise	Cai et al.[21]	JAFFE[13], CK+[11]	SBN-CNN	95.24%, 96.87%
22	neutral, happy, angry, sad, fear, surprise, disgust, contempt	Li et al.[22]	RAF-DB[18], AffectNet[17]	ACNN	80.54%, 54.84%
23	Anger, contempt, disgust, fear, happiness, sadness, and surprise	Agrawal et Mittal.[23]	FER2013[12]	CNN	65%
24	Happiness, sadness, anger, disgust, fear, and surprise	Kim et al.[24]	MMI[8], CASME II [15]	CNN-LSTM	78.61%, 60.98%
25	Anger, contempt, disgust, fear, happiness, sadness, and surprise	Yu et al.[25]	CK+[11], Oulu-CASIA[16], MMI[8], BP4D[29]	STC-NLSTM	99.8%, 93.45%, 84.53%

5. Conclusion

Emotion expression is crucial in communication, which improves the quality of human contact. Furthermore, in the near future, research into facial expression detection may help to better feedback to society as well as interaction between Human Robot Interface (HRI). The geometric section of the face is used to discern emotion the most (eg; eyes, eyebrow, and mouth).The review considers experiments that were carried out in a controlled environment, as well as real-time and wild images.

This paper presented recent FER research, allowing us to stay up to date on the most recent breakthroughs in the field. In order to have and achieve an accurate detection of human emotions, this paper described different architectures of CNN and CNN-LSTM recently proposed by different researchers and presented some different databases containing spontaneous images collected from the real world and others formed in laboratories. This publication also includes a discussion that demonstrates the high rate achieved by researchers, highlighting the fact that machines will be more capable of interpreting emotions in the future, implying that human-machine contact will grow more natural.

FERs are one of the most essential means of delivering information about an individual's emotional state, but they are always constrained by just knowing the six basic emotions plus neutral. It clashes with what is present in everyday life, which has more complicated feelings. This will encourage academics to generate larger databases and powerful deep learning architectures to recognise all basic and secondary emotions in the future. Furthermore, today's emotion recognition has progressed from unimodal to complex system multimodal analysis. Multimodality is one of the conditions for having a perfect detection of human emotion, according to Pantic et Rothkrantz [26]. For example, Zhang et al. [27] and Ringeval et al. [28] researched the integration of audio and visual for audio-visual and physiological modalities, and are now pushing their research to construct and offer effective multimodal deep learning architectures and databases.

References

[1] C. Marechal et al., “Survey on AI-Based Multimodal Methods for Emotion Detection”, in High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet, J. Kołodziej et H. González-Vélez, Éd. Cham: Springer International Publishing, 2019, p. 307-324.

[2] E. Sariyanidi, H. Gunes, et A. Cavallaro, “Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition “, IEEE Trans. Pattern Anal. Mach. Intell., oct. 2014, doi: 10.1109/TPAMI.2014.2366127.[1] E. Sariyanidi, H. Gunes, et A. Cavallaro, “ Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition “, IEEE Trans. Pattern Anal. Mach. Intell., oct. 2014, doi: 10.1109/TPAMI.2014.2366127.

[3] C.-N. Anagnostopoulos, T. Iliou, et I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 ", Artif. Intell. Rev., vol. 43, no 2, p. 155-177, févr. 2015, doi: 10.1007/s10462-012-9368-5.

[4] L. Shu et al., “A Review of Emotion Recognition Using Physiological Signals”, Sensors, vol. 18, no 7, p. 2074, juill. 2018, doi: 10.3390/s18072074.

[5] P. V. Rouast, M. Adam, et R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments", IEEE Trans. Affect. Comput., p. 1-1, 2018, doi: 10.1109/TAFFC.2018.2890471.

[6] M. H. Alkawaz, D. Mohamad, A. H. Basori, et T. Saba, "Blend Shape Interpolation and FACS for Realistic Avatar", 3D Res., vol. 6, no 1, p. 6, janv. 2015, doi: 10.1007/s13319-015-0038-7.

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, et S. Baker, “Multi-PIE”, Proc. Int. Conf. Autom. Face Gesture Recognit. Int. Conf. Autom. Face Gesture Recognit., vol. 28, no 5, p. 807-813, mai 2010, doi: 10.1016/j.imavis.2009.08.002.

[8] M. Pantic, M. Valstar, R. Rademaker, et L. Maat, “Web-based database for facial expression analysis”, in 2005 IEEE International Conference on Multimedia and Expo, juill. 2005, p. 5 pp.-, doi: 10.1109/ICME.2005.1521424.

[9] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, et K. Scherer, “The first facial expression recognition and analysis challenge”, in Face and Gesture 2011, mars 2011, p. 921-926, doi: 10.1109/FG.2011.5771374.

[10] A. Dhall, R. Goecke, S. Lucey, et T. Gedeon, “Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark”, in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), nov. 2011, p. 2106-2112, doi: 10.1109/ICCVW.2011.6130508.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, et I. Matthews, “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, juin 2010, p. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[12] I. J. Goodfellow et al., “Challenges in Representation Learning: A Report on Three Machine Learning Contests”, in Neural Information Processing, Berlin, Heidelberg, 2013, p. 117-124, doi: 10.1007/978-3-642-42051-1_16.

[13] M. Lyons, M. Kamachi, et J. Gyoba, “The Japanese Female Facial Expression (JAFFE) Database”. Zenodo, avr. 14, 1998, doi: 10.5281/zenodo.3451524.

[14] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, et M. J. Rosato, “A 3D facial expression database for facial behavior research”, in 7th International Conference on Automatic Face and Gesture Recognition (FGR06), avr. 2006, p. 211-216, doi: 10.1109/FGR.2006.6.

[15] W.-J. Yan et al., “CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation”, PLoS ONE, vol. 9, no 1, janv. 2014, doi: 10.1371/journal.pone.0086041.

[16] G. Zhao, X. Huang, M. Taini, S. Z. Li, et M. Pietikäinen, “Facial expression recognition from near-infrared videos”, Image Vis. Comput., vol. 29, no 9, p. 607-619, août 2011, doi: 10.1016/j.imavis.2011.07.002.

[17] A. Mollahosseini, B. Hasani, et M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild”, IEEE Trans. Affect. Comput., vol. 10, no 1, p. 18-31, janv. 2019, doi: 10.1109/TAFFC.2017.2740923.
[18] S. Li, W. Deng, et J. Du, “Reliable Crowd sourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild”, 2017, p. 2852-2861.
[19] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, et A. van Knippenberg, “Presentation and validation of the Radboud Faces Database”, Cogn. Emot., vol. 24, no 8, p. 1377-1388, déc. 2010, doi: 10.1080/02699930903485076
[20] A. T. Lopes, E. de Aguiar, A. F. De Souza, et T. Oliveira-Santos, “Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order”, Pattern Recognit., vol. 61, p. 610-628, janv. 2017, doi: 10.1016/j.patcog.2016.07.026.
[21] J. Cai, O. Chang, X. Tang, C. Xue, et C. Wei, “Facial Expression Recognition Method Based on Sparse Batch Normalization CNN”, in 2018 37th Chinese Control Conference (CCC), juill. 2018, p. 9608-9613, doi: 10.23919/ChiCC.2018.8483567.
[22] Y. Li, J. Zeng, S. Shan, et X. Chen, “Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism”, IEEE Trans. Image Process., vol. 28, no 5, p. 2439-2450, mai 2019, doi: 10.1109/TIP.2018.2886767.
[23] A. Agrawal et N. Mittal, “Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy”, Vis. Comput., janv. 2019, doi: 10.1007/s00371-019-01630-9.
[24] D. H. Kim, W. J. Baddar, J. Jang, et Y. M. Ro, “Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition”, IEEE Trans. Affect. Comput., vol. 10, no 2, p. 223-236, avr. 2019, doi: 10.1109/TAFFC.2017.2695999.
[25] Z. Yu, G. Liu, Q. Liu, et J. Deng, “Spatio-temporal convolutional features with nested LSTM for facial expression recognition”, Neurocomputing, vol. 317, p. 50-57, nov. 2018, doi: 10.1016/j.neucom.2018.07.028.
[26] M. Pantic et L. J. M. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction”, Proc. IEEE, vol. 91, no 9, p. 1370-1390, sept. 2003, doi: 10.1109/JPROC.2003.817122.
[27] S. Zhang, S. Zhang, T. Huang, et W. Gao, “Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition”, in Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 2016, p. 281–284, doi: 10.1145/2911996.2912051.
[28] F. Ringeval et al., “Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data”, Pattern Recognit. Lett., vol. 66, p. 22-30, nov. 2015, doi: 10.1016/j.patrec.2014.11.007.
[29] M. F. Valstar et al., “FERA 2015 - second Facial Expression Recognition and Analysis challenge”, in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), mai 2015, vol. 06, p. 1-8, doi: 10.1109/FG.2015.7284874