OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

FACIAL EMOTION RECOGNITION USING DEEP LEARNING: A SURVEY

Ms. Manju Lata Joshi, Department of CS & IT, International School of Informatics & Management, Jaipur
Ms. Suhani Agarwal, Department of CS & IT, IIS (Deemed to be University), Jaipur

Abstract

Recognition of facial expression has been an important and exciting area of research in the past decade. It is still challenging due to high intraclass variations like facial expression, body posture, speech recognition, etc. There is an immense stipulation for facial expression acknowledgement in different areas like medical services, educational institutes, business, entertainment, e-commerce, health, and security. Face expressions play an essential role in identifying a person's emotions. To detect or identify someone's emotions, traditional approaches such as Local Binary Pattern (LBP), Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradient (HOG) rely on handcrafted features, followed by a trained classifier on a database of photos or videos. The majority of the research studies perform reasonably well on datasets of images collected under controlled conditions but cannot perform well on more complex datasets with more image variation and partial faces. In recent years several research studies have suggested an end to end system for Facial Emotion Recognition (FER) using deep learning models. These studies have exhibited outstanding performance and show that deep learning-based FER performs better than a traditional approach based FER. This study provides an in-depth analysis of the existing research for facial emotion recognition through deep learning. Firstly, existing studies based on traditional approaches are analyzed, and then deep learning-based approaches used for FER are discussed.

Keywords: Emotion Recognition, Facial Expression, Facial Expression Recognition, Deep Learning.

Introduction

Emotion forms an indispensable part in any interpersonal; hence emotions play a vital role in human communication, which helps us to understand other people's intentions and moods. This area of research has attracted researchers for several reasons as human-computer interaction, which then helps in better advertising, animations, medicines and, security with improved human communication with the help of emotional quotient or emotional intelligence of a human being. One of the best parts about emotions is that they can be seen with naked eyes. Hence, any indication, preceding or subsequent signals, can be subject to identification and recognition using correct means. There are several ways to examine human expression recognition, ranging from facial expressions, body posture, voice tone, speech etc. Among these characteristics, for various reasons, facial expressions are one of the most common. They can be seen recognized and contain several helpful characteristics for emotional detection. It is comparatively easy to get an extensive dataset of faces. So, essentially FER is a method of identifying human feelings through facial expressions.

As per the feature representation, the system of FER can be categorized into two major categories, i.e. Dynamic Sequence FER and Static Image FER. Spatial information from a single image based on

static methods feature representation is encoded. In contrast, a temporal relationship in the input facial expression series among contiguous frames is considered in dynamic methods. Dynamic FER has a higher recognition rate than static FER as dynamic FER provides additional temporal knowledge. Psychologist Paul Ekman's work(Ekman & Friesen, 1976) has become imperative to the growth of this area.

In the early 20th century (Ekman & Friesen, 1976), research work identified six basic emotions, showing that humans interpret these emotions in a similar path irrespective of culture. The facial features are disgust, anger, happiness, fear, surprise and sadness. Later on, contempt and neutral was added as the fundamental emotions. Based on this concept, a Facial Action Coding System (FACS) was developed by Paul Ekman. Most of the facial emotion recognition studies nowadays are based on FACS.

There are mainly two approaches to identifying facial expression recognition by detecting the action unit (AU) and the facial points. FACS is a framework that is used to detect action units. FACS framework quantifies human facial expression by analyzing changes in the facial muscle when an emotion is triggered. FACS describes the facial muscle movement across 44 areas on the face or action units. Hence, facial expressions may be recognized through the presence and intensity of many AUs. When a framework is categorized by a rundown of prepared images from the given dataset, the captured photograph eventually contributes to discovering the individual's enthusiastic condition alongside his / her picture. It is called the Facial Emotion Recognition System.

The OCC (Ortony/Clore/Collins) model(Clore & Ontony, 2013) is another standard model that characterizes twenty-two types of emotions based on emotional responses to circumstances and is intended to model human emotions. On the other hand, the emotion model proposed by(Plutchik & Conte, 1997) is a dimensional model that provides an integrative theory based on the concepts of evolution and describes eight basic bipolar emotions.

The process of facial expression recognition is divided into four stages, i.e. image pre-processing, face detection, feature extraction and facial expression classification. Feature extraction is the most critical step and trickiest in recognizing facial expressions among all the steps. Once the feature extraction is done, the classifier divides the extracted features into the respective categories based on the classes of emotions.

There are several forms of feature extraction techniques at present, which can be split into texture feature-based, geometric based and the assortment of these two. The geometric-based approach uses facial features like nostrils, eye corners, etc., or facial elements like eyebrows, mouth, eye, etc. It shows the location and shape of facial components. Geometric feature-based methods have shown better performance than appearance-based methods. However, these methods typically require detecting facial components that are sometimes complex to accommodate in many cases. In order to track geometry-based recognition, two different types of models, the Active Appearance Model (AAM) and Active Shape Model (ASM), are used. ASM is a statistical model of object shape defined by a collection of points that iteratively deforms to match a new image.

On the other hand, AAM uses statistical model objects to fit a new image by shape and appearance. Appearance or texture-based approaches use the face's texture, which catches the strength changes associated with various expressions, like bulges, furrows, and wrinkles. Image filters work

with appearance-based methods like Gabor wavelets applied to a specific part of the entire face. However, the convolution of face images with Gabor filters consumes more time and space to derive coefficients of multi-scale and multi-orientation. Hence, the best option is to combine these two approaches for facial expression recognition.

Similarly, several conventional methods have been used to extract handcrafted features traditionally or by sparse learning. The LBP is a texture or appearance classification tool used as a visual descriptor in computer vision. It has been further analyzed that LBP combined with HOG gives better results (Alhindi et al., 2018). Another method, HOG, is a descriptor of features. The technique counts gradient orientation occurrences in localized portions of an image.

Similarly, SIFT is also an algorithm to detect local features in images. The other method SIFT relies on the appearance-based method and is invariant to the scale and rotation of the image. The critical limitation of feature-based approaches is that these require significant effort to design and employ different methods. The modern deep learning-based solution suggests overcoming this limitation, wherein a machine automatically extracts the facial features.

Two types of methods prevail in facial emotion recognition: Analytical and Holistic. The holistic method primarily aims to model human facial deformations that encode the entire face globally. On the other hand, to construct informative & expressive models, analytical methods scrutinize and quantify local or distinct human facial warp such as in eyebrows, eyes, nose, mouth, etc. and their geometrical relationships. Various technical advancements in machine learning have made it easier to identify emotions. However, there is a tremendous need for automatic detection of emotion from facial expression, and deep learning-based algorithms are outperforming it.

Deep learning extends an artificial neural network with representation learning (feature learning). and increasingly uses multiple layers to extract higher-level characteristics or features from the raw input. Deep learning, particularly convolutional neural networks, can extract and learn many facial features for a facial expression recognition system automated neural network & deep belief network. Most modern deep learning models, particularly Convolutional Neural Networks, are based on artificial neural networks. The coming section discusses significant studies conducted for facial emotion recognition with an explicit focus on deep learning techniques.

Literature Survey

In 1978 very first effort was made by(Sown et al., 1978) to analyze facial expressions automatically. A thorough investigation of the initial research studies on this topic is found in(Pantic & Rothkrantz, 2000; Fasel & Luettin, 2003). The study Initiated by(Tian et al., 2001) suggested identifying facial expressions using FACS. After that, several studies have discussed detecting AU occurrence and AU intensity, including(Tian et al. 2005). The different approaches detect distinct facial points and interpret the meaning of expression accordingly. Three significant steps are universal in automatic deep FER, i.e., pre-processing, deep feature learning, and deep feature classification. Out of these three, feature extraction and classification are two critical tasks in FER and Analysis (FERA). Most of the current FERA methods used different pattern recognition techniques to identify facial expressions based on facial characteristics, including approaches based on geometric features, appearance features, texture and hybrid features.

Some of the geometric features are Active Shape Model(ASM), extended ASM etc. A face detection

algorithm based on Haar-like features to detect faces is discussed in(Viola & Jones, 2004). A classifier is trained in the technique proposed by(Milborrow & Nicolls 2008), further extending the original ASM. The improved ASM proved more efficient and accurate in practical applications as it detects the eye and mouth centres to offer correct initialization.(Pantic & Rothkrantz, 2000) extracted the outline of the face components like the mouth, eyes and profile contours. Based on these contours and rule-based reasoning, thirty-two individual facial Action Units (AU) occurring in combination or alone are identified. The study claims eighty-six per cent (86%) of the average recognition rate using their proposed method.

In order to recognize facial expressions, appearance-based features are also used.(Liu & Wang, 2006) used a combination of neural network and Gabor features to offer improved accuracy than the original Gabor system.(Bartlett et al., 2006) proposed a system that selects a subset of Gabor features and trains an SVM classifier.(Whitehill & Omlin, 2006) explored the characteristics of Haar and the AdaBoost for identification of facial action unit that yielded the recognition rates with the SVM approach are equivalent to those of the Gabor but works faster.

A method presented by(Ying et al., 2010) was based on sparse representation and using the local binary patterns (LBP) from facial images; they believed that the proposed approach worked better than conventional LDA and PCA algorithms.

To extract the most discriminating LBP functionality, they formulated Boosted-LBP, and the best accuracy was achieved using an SVM classifier. For low-resolution face images, LBP produces stable and robust outcomes.(Shan et al., 2009).

[Lucey et al., 2010) made use of the Active Appearance Model (AAM) and classified seven facial expressions of a human face using the Extended Cohn Kanade (CK+) database along with SVM. Their classification was based on canonical normalized appearance characteristics, geometric shape characteristics, and combined shape characteristics. They exhibited that the combined characteristics have a greater identification rate than geometric shape features or appearance-based features. The AAM was used by(Mahoor et al., 2011) to detect positions of facial points; they extracted Gabor characteristics then presented a sparse representation based method to classify facial expressions with comparable outcomes.

The study mentioned in(Chen, 2012) compares the performance (in terms of recognition rate) between hybrid feature method and Active Appearance Model that used the CK+ database is classified by SVM classifier. The hybrid function approach gives 95% accuracy in recognition, which performed better than the other approach, giving 83.3 per cent accuracy. After extracting the feature, a classifier is used to identify the expression from the extraction. In order to identify facial expressions, various techniques have been suggested, such as rule-based classifiers, Support Vector Machine (SVM), Neural Network, Bayesian Network (BN), K-Nearest Neighbor and Random Forest.

In recent times, deep learning techniques have drawn much attention, and tremendous research interest has arisen in the study of deep learning approaches to detect facial emotions. Deep learning methods have been tested to identify facial emotions and are primarily focused on supervised learning.

In 2002, the CNN was introduced to classify facial expressions to address the commonly come

across problems of face expressions and irregular elucidation.

A CNN-based series of FER systems developed that, by training a multi-task deep neural network to recognize facial expressions, could predict the positioning of facial feature points. In the case of previously unseen variations in facial pose, a convolutional neural network behaves better than multilayer perceptron (MLP)[Fasel, 2002).

[Cohen et al., 2003) compared various classifiers of Bayes, and among all Gaussian Tree-Augmented-Naive (TAN), Bayes showed the most excellent performance.(Bartlett et al., 2005) systematically compared various techniques for facial expression recognition, including AdaBoost, LDA and SVM, and obtained the most excellent results by using AdaBoost and selecting a subset of Gabor filters. Further, the model is trained using the SVM model.

Much progress has been made for deep learning-based FER in the last decade. Deep CNN and Filter banks recognized emotions from face images(Huang et al., 2016). The study mentioned in(Stolar et al., 2017) concludes that image spectrograms with deep convolutionary networks can also be implemented to perform FER.

The study in(Khorrami et al., 2015) demonstrated that CNNs could attain high emotion recognition accuracy.(Aneja et al., 2016) proposed a model based on deep learning for facial expression recognition for stylized animated characters. The model is trained using human facial expressions and animated faces and mapping human images into animated ones.

A deep network proposed by(Mollahosseini et al., 2016) comprises two convolutional layers, each accompanied by using maximum pooling, followed by four Inception layers, is defined in(Bartlett et al., 2006; Cohen et al., 2003). The network is defined as the architecture of a sole part that takes registered facial images as input and categorizes them into six fundamental expressions or neutral expressions. The method records 66.4 per cent accuracy on the FER dataset. To simulate the temporal behaviours of facial expressions from image sequences, Hidden Markov Models (HMMs) have been commonly used(Cohen et al., 2003).

Dynamic Bayesian Networks (DBNs) have been used to recognize sequence-based expressions in (Kaliouby & Robinson, 2005; Zhang & Ji, 2005). Deep Belief Network (DBN) has obtained important results for facial expression recognition combined with other approaches. For FER(Jung et al., 2015), the powerful classifier AdaBoost and DBN approach had a high expression recognition accuracy rate of expression recognition.

A novel method for extracting salient features from deep faces was proposed, i.e. MLDP-GDA (Modified Local Directional Patterns GDA (Modified Local Directional Patterns, Generalized Discriminant Analysis), which was applied to DBN for training various facial expressions(Uddin et al., 2017).

Deep belief networks and hidden Markov models with unweighted average recall (UAR) has also been used to detect emotions(Hairar et al., 2017). Among the studies discussed above, Most of them achieve substantial performance progress compared to the conventional emotion recognition works, but attaining the relevant emotion detection facial regions seems to be lacking. A framework on the attentional convolutional network was proposed to address this issue that focused on salient face regions and dramatically improved on multiple datasets(Singh & Kaur, 2018).

Inception net v3 model that is a CNN architecture from inception family trains a model of emotion detection and is used for automated image classification and image labelling. It interprets and identifies the emotions from the spectrogram of an image. The spectrogram is nothing but an image from which emotion can be recognized. After refining the structure, every level shows the value of the emotion to be estimated by the trained Model. Transfer learning is also used in this Model. Using this Model, an accuracy of 35.91% per image spectrogram was achieved. It was observed that precision was very low(Roopa, 2019). The model proposed by(Singh & Kaur, 2018), is implemented on the Image Net dataset to see the effects of image recognition 41% accuracy was achieved by the Model.

To detect the occurrence of Action Units, automatic extraction of features is done using a deep convolutional neural network. The observation is that the mean square error (MSE) decreases when the training data raises, and MSE is linear with the size of testing data.(Liliana, 2019).

The research work mentioned in(Li & Deng, 2020) is a systematic, thorough analysis of deep learning for FER tasks based on both dynamic and still images and an overview of the network design and performance of the existing deep FER systems.

For 3D face recognition, Microsoft Kinect is primarily used in(Tarnowski et al., 2017) because of its low price and simplicity. The Kinect system calculated six Action Units (AU). For Emotion recognition using deep learning, an exhaustive study is given in)Li & Zhao, 2019; Giannopoulos et al., 2018), representing the study of classifier for face emotion recognition from machine learning to deep learning, its dataset properties & their respective results.

Conclusion

Recognition of facial expression is a very complex area of research and is still in the elementary and exploratory phases. This paper discussed various techniques for facial emotion recognition, emphasizing deep learning-based techniques. It has been observed that deep learning models are widely used due to their excellent feature-learning ability. However, most of the models are of greater time complexity in terms of the training period taken. There is still a necessity to develop models with lesser time complexity.

References

- Alhindi, T. J., Kalra, S., Ng, K. H., Afrin, A., &Tizhoosh, H. R. (2018). Comparing LBP, HOG and in-depth features for classification of histopathology images. In 2018 joint international conference on neural networks (IJCNN) (pp. 1-7). IEEE.
- Aneja, D., Colburn, A., Faigin, G., Shapiro, L., & Mones, B. (2016). Modelling stylized character expressions via deep learning. In Asian conference on computer vision (pp. 136-153).Springer, Cham.
- Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., &Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behaviour. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 2, pp. 568-573).IEEE.
- Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully automatic facial action recognition in spontaneous behaviour. In 7th International Conference on Automatic Face and Gesture Recognition (FGR06) (pp. 223-230). IEEE.

- Chen, J., Chen, D., Gong, Y., Yu, M., Zhang, K., & Wang, L. (2012). Facial expression recognition using geometric and appearance features. In Proceedings of the 4th international conference on internet multimedia computing and service (pp. 29-33).
- Clore, G. L., &Ortony, A. (2013).Psychological construction in the OCC model of emotion. Emotion Review, 5(4), 335-343.
- Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modelling. Computer Vision and image understanding, 91(1-2), 160-187.
- Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. Environmental psychology and nonverbal behaviour, 1(1), 56-75.
- Fasel, B., &Luettin, J. (2003). Automatic facial expression analysis: a survey. Pattern recognition, 36(1), 259-275.
- Fasel, B. (2002, August). Robust face analysis using convolutional neural networks.In Object recognition supported by user interaction for service robots (Vol. 2, pp. 40-43). IEEE
- Giannopoulos, P., Perikos, I., &Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. In Advances in the hybridization of intelligent methods (pp. 1-16).Springer, Cham.
- Huang, K. Y., Wu, C. H., Yang, T. H., Su, M. H., & Chou, J. H. (2016). Speech emotion recognition using autoencoder bottleneck features and LSTM. The 2016 International Conference on Orange Technologies (ICOT) (pp. 1-4).IEEE.
- Jung, H., Lee, S., Park, S., Kim, B., Kim, J., Lee, I., &Ahn, C. (2015).Development of deep learning-based facial expression recognition system.In 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV) (pp. 1-4). IEEE
- Kaliouby, R. E., & Robinson, P. (2005).Real-time inference of complex mental states from facial expressions and head gestures.In Real-time vision for human-computer interaction (pp. 181-200).Springer, Boston, MA.
- Khorrami, P., Paine, T., & Huang, T. (2015). Do deep neural networks learn facial action units when doing expression recognition?. In Proceedings of the IEEE international conference on computer vision workshops (pp. 19-27).
- Liliana, D. Y. (2019).Emotion recognition from facial expression using deep convolutional neural network. In Journal of physics: conference series (Vol. 1193, No. 1, p. 012004). IOP Publishing.
- Li, C. E. J., & Zhao, L. (2019).Emotion recognition using convolutional neural networks.
- Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing.
- Liu, W., & Wang, Z. (2006). Facial expression recognition based on fusion of multiple Gabor features. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 536-539).IEEE.
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). 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.

- Mahoor, M. H., Zhou, M., Veon, K. L., Mavadati, S. M., & Cohn, J. F. (2011). Facial action unit recognition with sparse representation.In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG) (pp. 336-342).IEEE.
- Milborrow, Stephen &Nicolls, Fred. (2008). Locating Facial Features with an Extended Active Shape Model. Proc. of European Conf. on Computer Vision. 5305. 504-513. 10.1007/978-3-540-88693-8_37.
- Minaee, S., Minaei, M., &Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046.
- Mollahosseini, A., Chan, D., &Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks.In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE
- Pantic, M., &Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. IEEE Transactions on pattern analysis and machine intelligence, 22(12), 1424-1445.
- Plutchik, R. E., & Conte, H. R. (1997).Circumplex models of personality and emotions (pp. xi-484). American Psychological Association.
- Roopa, N. (2019). S "Emotion Recognition from Facial Expression using Deep Learning". International Journal of Engineering and Advanced Technology (IJEAT) ISSN, 2249-8958.
- Singh P. &Kaur A. (2018).Deep Learning-based Face Emotion Recognition, JETIR.5(12).
- Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6), 803-816.
- Sown, M. (1978).A preliminary note on pattern recognition of facial emotional expression.In The 4th International Joint Conferences on Pattern Recognition, 1978.
- Stolar, M. N., Lech, M., Bolia, R. S., & Skinner, M. (2017, December). Real-time speech emotion recognition using RGB image classification and transfer learning. In 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS) (pp. 1-8). IEEE.
- Tarnowski, P., Kołodziej, M., Majkowski, A., & Rak, R. J. (2017).Emotion recognition using facial expressions.Procedia Computer Science, 108, 1175-1184.
- Tian, Y. I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence, 23(2), 97-115.
- Tian, Y. L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In Handbook of face recognition (pp. 247-275).Springer, New York, NY.
- Uddin, M. Z., Hassan, M. M., Almogren, A., Zuair, M., Fortino, G., & Torresen, J. (2017). A facial expression recognition system using robust face features from depth videos and deep learning. Computers & Electrical Engineering, 63, 114-125.
- Viola, P., & Jones, M. J. (2004). Robust real-time face detection.International Journal of computer vision, 57(2), 137-154.

- Ying, Z., Wang, Z., and Huang, M. (2010). Advanced intelligent computing theories and applications. With Aspects of Artificial Intelligence, Lecture Notes in Computer Science, 6216, (2010),457-464.
- Zhang, Y., &Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Transactions on pattern analysis and machine intelligence, 27(5), 699-714.