An Analysis of AI-Based Emotional Recognition: Main Methods Based on Four Modalities

Tian Jing

Tianjin No.1 High School, Tianjin, 300051, China

18622065261@163.com

Abstract. Artificial Intelligence (AI) has become a useful tool in human emotion recognition, with a broad application range. To better cater for the applications, numerous researches are conducted, helping developing the related technologies rapidly. This review broadly explores the main methods in emotion recognition based on AI. It begins with facial emotion recognition (FER), analyzing its general working flow (from constructing database to preprocessing to extracting features to machine learning). It is seen in the following that this flow commonly applies to other three modalities. Then, speech emotion recognition (SER) is briefly discussed, mainly on its feature extraction and classification (classification is a part of machine learning). Subsequently, emotion recognition from physiological signals is deeply explored, due to its passive nature and resistance to artificial control. Among a variety of physiological signals, the review concentrates on electroencephalographic (EEG) and electrocardiographic (ECG) signals. Afterwards, textual emotion recognition (TER) is roughly introduced, outlining four basic methods based on it. Finally, the review concludes the challenges which occur to nearly every experiment regarding emotion recognition. Additionally, the strengths and limitations of each modality are presented in the discussion module. The highlight of the review is that it provides a systematic analysis of basic methods of using AI to recognize emotion.

Keywords: Artificial Intelligence; Emotion Recognition; Method, Modality.

1. Introduction

Nowadays, the interaction between humans and computers serves as an important part of daily life [1]. Emotions are also significant for individuals [1]. The utilization of artificial intelligence for emotion recognition is gaining significant traction, with extensive applications. For instance, AI-based emotion recognition is employed in software engineering, website customization, education and gaming [2]. It also intrigues interests in areas involving human–computer interaction, cognitive-behavioral science and the treatment of emotion-related diseases [3]. Particularly, in human-computer interaction, AI's comprehension of humans' emotions can help computers not only respond more naturally, but also understanding individuals' needs more accurately [4]. As a result, it's important to learn and develop the technologies concerning AI-based emotion recognition.

Therefore, this review aims to provide an easy overview over the main methods used to identify emotions with the help of AI. The literature review is separated into five sections, revolving around four primary data modalities: facial expressions (section 1), speech (section 2), physiological signals (section 3) and text (section 4). Finally, in section 5, the general challenges in all four modalities are discussed. For each modality, the author explores its basic impletion processes, including constructing database, preprocessing, extracting features and machine learning.

When discussing FER, which is the first part of the main body, all four steps are introduced in detail. This aims to provide a comprehensive perception of the total process. Later in SER, emotion recognition from physiological signals and TER, only several parts which the author regards as the most important ones are discussed. In detail, in this review, SER focuses on feature collection and classification in machine learning; emotion recognition from physiological signals concentrates on feature extraction and machine leaning from EEG signals, as well as some novel methodologies employed in ECG-signal-based emotion recognition; TER discusses four basic methods, namely keyword-based approach, rule-based approach, machine learning/classical learning–based approach, and deep learning–based approach.

2. Literature Review

There are basically four modalities in emotion recognition using AI, namely facial emotion recognition (FER), speech emotion recognition (SER), emotion recognition from physiological signals and textual emotion recognition (TER). Among them, FER is the most popular modality, hence the author discusses it firstly.

2.1 Facial Emotion Recognition (FER)

FER is the recognition of human emotion states from static images and videos [5]. To sum up, there are three main original steps of FER: firstly, a comprehensive database should be built; secondly, faces are identified from a video or a photo when non-facial components are removed; thirdly, some features are selectively removed or adapted for following steps; fourthly, data which have undergone pre-processing in step two and three are utilized to train a classifier, which produces affective tags during the training. Next, let's see it in a more detailed way.

2.1.1 Database

Those data used for researches usually come from pictures or images captured in videos [6]. Some of the pictures are taken when individuals pose for the photograph, others record spontaneous emotions, with lower quality compared to the former [6]. Nevertheless, it's expensive and time-costly to collect those data [7]. Also, photos taken from the real world may invade people's privacy [8]. To deal with those problems, researchers generate facial expression images using a wide range of tools. For instance, in the research conducted by Darne, Quan and Luo, images of facial expressions are created by firstly generating facial images using StyleGAN2 model and secondly applying EmoStyle model to the edition of facial expressions [7].

2.1.2 Pre-processing

Pre-processing is necessary for FER, since it can enhance the accuracy of recognition results significantly.

Pre-processing involves three main steps [6]. Firstly, human faces are detected in each picture, followed by the removal of non-facial parts. The Viola-Jones face detector is one of the most commonly utilized technology for this step [9]. Secondly, features in the images are reduced and normalized. Dalvi, Rathod, Patil, Gite and Kotecha suggests that illumination and pose should be normalized in FER, in order to enhance the capabilities of FER models [6]. Thirdly, data are enhanced in the face of the lack for data. This could be achieved by generating poses, glasses and so on [10]. Similar efforts are made in other studies, for example, researchers apply two models to create facial expression images [7], as mentioned above, in "1.1database".

2.1.3 Feature Extraction

There are several ways to extract features from a picture. Global and low-dimensional features can be extracted by Principal Component Analysis (PCA), which is one of the most widely used way in FER [6]. But if more detailed features are needed, Independent Component Analysis (ICA) could be a better way [6]. There are also methods which can be useful in extracting both global features and local features. For example, there are methods based on features regarding texture, such as using Gabor filter; we can also extract features based on edges [6].

2.1.4 Machine Learning

In FER, convolutional neutral network (CNN) is the most popular and effective way [6]. However, this kind of unimodal system has its own limitations. For example, an only CNN applied in FER possesses complex structure, too many training parameters and unideal identification results [11]. Hence, Qiao, Hou and Liu introduce an optimization algorithm, in which CNN is combined with support vector machine (SVM). This new method consists better accuracy and robustness, confirmed by examinations [11]. Also used to improve CNN's capabilities on FER, three CNN models with

different filters and layers are combined to distinguish different kinds of emotions [12]. It can also enhance the accuracy, though in a limited extent.

The model shown by Wang also bases on CNN. Different from other methods, this measure divided the FER process into two stages, namely ESER and ECER. Emotion State Expression Recognition (ESER), which uses DenseNet169, VGG16 and ResNet50 as basic models, is responsible to identify or classify the express of emotion states [5]. Among them, DenseNet169 and VGG16 build on CNN. Emotion Cause Extraction and Recognition (ECER), which employs two models, Xception and ViT, recognizes or extracts the reason of the arousal of a specific emotion [5]. This method possesses striking effectiveness, with the accuracy 95% in recognition of basic emotions and 88% in the recognition of mixed emotions.

2.2 Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) includes two major stages: feature extraction and classification [13]. Using convolutional neural network (CNN) is one of the methods (with two phases: initially, a sparse auto-encoder (SAE) with reconstruction segregation is employed to extract the local invariant features (LIFs) from unlabeled data; then, these LIFs are fed into a feature extractor for further analysis) [14]. Researchers also consider using semi-CNN firstly, and then extract features from contractive convolutional network [15].

2.3 Emotion Recognition from Physiological Signals

Compared with emotion recognition based on the analysis of facial expressions and speeches, technology focused on physiological signals are much more passive [1]. That is because, it's nearly impossible for a person to control his or her brain activities [16], suggesting subjects to be less capable of hiding their emotions deliberately or unconsciously. Thus, emotion recognition from physiological signals could be a rather good way for emotion recognition.

2.3.1 Electroencephalographic (EEG) Signals

Some features extracted from EEG signals can reflect several specific emotional states of individuals. For example, the alpha and beta bands correspond to positive emotions regarding evaluating oneself, such as gratefulness, inspiration and pride; the theta and gamma bands are associated with pleasurable emotions such as entertainment, interest and delight [22]. Hence, numerous researches employ EEG signals to detect people's reactions to external stimuli [22].

2.3.1.1. Feature Extraction

There are several methods of feature extraction developed. In the experiment conducted by Chaudary, Khan and Mumtaz, a large language model called GPT-4 is applied in the hybrid CNN-LSTM-LLM framework [16]. In this framework, CNNs aim to detect spatial information, with LSTMs intended for the extraction of temporal features and LLMs applied to reason the signals by transforming them into vector embedding. Using this method, it attains 60.3% accuracy in three-class emotion recognition. The author of [16] also suggests that the challenges in this technology focus on the prediction of positive moods (which are likely to be overestimated) and neutral ones. In the research done by Mehmood and Lee, Hjorth parameters are employed to detect features after the pretreatment of the input of EEG signals with the help of SVM [17]; in the study conducted by Sohaib, Qureshi, Hagelb¨ack, Hilborn, Jerˇci´c and Petar, Higher- Order Crossings analysis is applied [1].

When it comes to extracting EEG features, there is a common challenge: too many noises. The noises range from physiological artifacts (say, muscle activity and eye movements) to environmental artifacts (say, power lines interference, radiation from lights and other radio frequency emissions from medical devices) [16]. This poses a huge challenge to accurate emotion recognition based on EEG signals. To mitigate this problem, Saha, Kunju, Majid, Kashem, Nashbat, Ashraf, Hasan, Khandakar, Hossain, Alqahtani and Chowdhury introduce a one-dimensional deep learning model, MultiResUNet3p, which helps effectively generate clean EEG signals [23].

2.3.1.2 Machine Learning Technology

There are five main related kinds of machine learning technology [1]. They are: K-Nearest Neighbor (KNN), Regression Tree (RT), Bayesian Network (BNT), Support Vector Machine (SVM) and Artificial Neural Network (ANN). In the experiment, SVM is shown the best classifier (with a highest accuracy 77.78%), while KNN lags slightly after it. Although the other three classifiers don't demonstrate rather ideal results, it's suggested that RT is widely used in medical domain [18]; additionally, ANN can effectively deal with noisy data [19].

Apart from using single models, hybrid models can also be used to enhance accuracy and reduce the impact of noise (such as eye movements). For instance, a EEG-CNN-souping model can be employed [16]. In this combined model, several EEG-CNN models are separately trained, and then their weights are averaged. Consequently, the accuracy of emotion recognition is enhanced, but with the identify time not increased.

2.3.2 Electrocardiographic (ECG) Signals

According to a range of experiments, ECG signals differs when it comes to different emotion states [22]. Additionally, it's more comfortable to record ECG signals compared to EEG signals, making subjects more likely to accept continuous monitoring [3]. As a result, ECG could serve as a useful tool to recognize emotions. Generally, cardiac activity is assessed utilizing key electrocardiographic parameters, which includes the P wave, Q wave, T wave, QRS complex and QT/QTc intervals. Among them, QRS complex is most commonly used in emotion-recognition-related researches, with its duration and amplitude mostly focused on [22].

It is concerned in many ECG studies that individual differences exist in the sense of responses. Thus, the study conducted by Fan, Qiu, Wang, Zhao, Jiang, Wang, Xu, Sun and Jiang develops a method of introducing attentional mechanisms to deep learning models. In more details, this is done by combining the deep convolutional neural network and Convolutional Block Attention Module (CBAM) [21]. In the system, the deep convolutional neural network captures ECG features while CBAM adds weight information to those extracted features. Thanks to this disposal, the neural network can better concentrate on those important and common information while avoid the majority of noise.

However, it's argued that using deep neural networks possesses larger complexity and requires more training time. Hence, Fang, Pan, Yu, Yang and He suggest to employ random convolutional kernel method in ECG emotion recognition [3]. By deploying multiple convolutional kernels with parameters including length, weight, bias, dilation and padding, this method is able to extract a variety of emotional states from ECG signals, with both high efficiency and high accuracy.

2.4 Textual Emotion Recognition (TER)

There are four basic methods for Textual Emotion Recognition (TER), namely keyword-based approach, rule-based approach, machine learning/classical learning–based approach, and deep learning–based approach [24].

The keyword-based approach recognizes different emotions by analyzing the keywords' locations in the input text and comparing them with relevant labels from the dataset [24]. It includes five stages: determining emotional keyword lists from normative vocabulary-related databases, preprocessing of the given text, matching keywords found in the text with the prepared keyword lists, assessing the emotional intensity and carrying out negation checking. In the end, the label regarding emotional state is attained [24].

By contrast, the approach built on rules identifies emotions with the use of regulations in the terms of logic and grammar, while machine learning approach employs complex classifiers with features on n-gram, lexical, semantic and grammatical dimensions [24]. Different from the machine learning approach, in which supervised machine learning algorithms are extensively employed, deep learning algorithm allows unsupervised learning from unlabeled data [24]. This trait also differentiates it from the keyword-based approach, which needs labeled data.

2.5 Challenges

As generally researchers define and distinguish different emotions by themselves to offer tags, significant differences could be seen between different researches [6]. This is a challenge rather hard to be overcome. Although in the experiment produced by Wang, each tag is decided by voting [5], which is a useful method to reduce the error caused by subjectivity, this problem is still inevitable. That is because, it's rather hard to define a kind of emotion precisely, and different people may have different views on which category a specific emotion belongs to.

Additionally, different individuals may behave differently to express identical emotions [1], not to mention people may arise different emotions in a same circumstance. Hence, feature election is definitely no easy task, which is also received by Sohaib, Qureshi, Hagelb¨ack, Hilborn, Jerˇci´c and Petar [1]. Some methods can be adopted to mitigate the problem, such as choosing features which possess larger absolute correlation among subjects [20] and introducing attentional mechanism [21]. But undeniably, the problem cannot be thoroughly eliminated.

Besides, it's clear that the chosen subjects cannot on behave of all human beings. This inevitably leads to errors. The rather that, it's not an unusual case that selected features may not be able to conclude all subjects, leading to errors that cannot be ignored.

3. Discussion

55% of human emotions are conveyed through facial expressions, which represents the highest proportion [6]. Although contains plenty of emotional information, there is an obvious disadvantage in FER, which can also be seen in SER and TER: the detected results may be disturbed, as individuals can control their facial expressions, pitch, writing style and so on. By contrast, EEG signals and ECG signals are impossible to be controlled consciously or unconsciously [16], making them hopeful to help develop accurate emotion recognition technologies. However, they also face plenty of challenges. For instance, many noises affect the result of attaining EEG signals [16], necessitating sophisticated noise reduction models like MultiResUNet3p [23]. Additionally, physiological data-based methods pose many more difficulties to collect data, as specialized equipment are required [3].

Besides, several challenges can be seen in every modality. Firstly, it's hard to define an emotion by oneself, and people may have different opinions on the exact definition. Secondly, different people may have different emotions in a same circumstance; even when they are in the same emotion state, they may behave differently. Thirdly, many chosen features cannot be shown in all subjects, and the subjects cannot on behalf of everyone. Basing on the discussions above, the advantages and disadvantages of the four modalities (FER, SER, EEG/ECG and TER) are analyzed and compared in this review (shown in Table 1).

Table 1. Advantages and disadvantages of FER, SER, EEG/ECG and TER

MODALITY	ADVANTAGES	DISADVANTAGES	
FER	1. With highest proportion of emotional information 2. More developed	Easy-to-be-disturbed detected results	1. Hard to define an emotion 2. Different people behave differently in the same emotion state; Different people have different emotions in a same circumstance 3. Chosen features hard to be shown in all subjects; Subjects cannot on behalf of everyone.
SER	Not invasive	Easy-to-be-disturbed detected results	
EEG/ECG	Impossible to be controlled artificially	1. Noises 2. Hard to collect data	
TER	Capable of completing the picture for digital contexts	Easy-to-be-disturbed detected results	

All of these show that no method can be totally perfect, each of them has its advantages and disadvantages. If it's needed to decide which method to use, it's suggested to balance those pros and cons carefully with users' needs. Or maybe in the future, a multimodal AI system intelligently integrating information from faces, speeches, physiological signals and text will be developed, in order to draw on the strengths from every modality and avoid problems which don't occur to all

modalities, such as subjective disturbance. In the meantime, efforts need to be done to mitigate challenges which generally exist in every method. This is the work of not only computer science learners, but also psychology learners.

4. Conclusion

In conclusion, this review provides a programmatic overview on emotion recognition based on AI, which is a rapidly developing technology integrating two different subjects (artificial intelligence and emotion recognition). Moreover, the author extracts a general working flow after referring to vast amounts of academic literatures. This presents a clear and easy-to-follow methodology framework, especially useful for those beginners. Additionally, four main modalities related to AI emotion recognition are compared in the "discussion" session, helping readers better understand and choose the most suitable method in a specific application occasion.

It can be concluded from the review that there are plenty of methods to recognize human emotions with the help of AI. Different forms of data could be delivered to large language models after preprocess. Among them, FER, SER, emotion recognition based on physiological signals and TER are the four most common modalities. This review demonstrates some core methodologies based on each of them.

The analysis indicates that there is no single superior method. While FER is more developed, emotion recognition based on physiological signals is resistant to deception. By contrast, SER is more non-invasive and TER completes the picture for digital contexts.

Looking forward, hybrid and multimodal systems may be effective to draw on the advantages and avoid the limitations of a single method. This employs the complementary strengths of various methodologies. Especially when it comes to combining different modalities, as they differ from each other in numerous aspects, making it easier to find out complementary fields. Apart from improving algorithm, ethical issues should also be treated rigorously. No method should be utilized or developed when it invades the privacy of recorded individuals, for example. Sticking to this notion can help us develop a more trustworthy technology.

References

[1] Tauseef Sohaib, Ahmad Qureshi, Shahnawaz Hagelb¨ack, Johan Hilborn, Olle Jerˇci´c, Petar. Evaluating Classifiers for Emotion Recognition Using EEG. Found. Augment. Cognit. Lecture Notes Comput. Sci.. 8027(2013). 492-501.10.1007/978-3-642-39454-6 53.

[2] A. Landowska, M. Szwoch, W. Szwoch, M.R. Wróbel, A. Kołakowska, Emotion recognition and its applications, Springer International Publishing, 2014, 10.1007/978-3-319-08491-6_5.

[3] Ancheng Fang, Fan Pan, Weichuang Yu, Linkun Yang, Peiyu He, ECG-based emotion recognition using random convolutional kernel method, https://doi.org/10.1016/j.bspc.2023.105907.

[4] Picard, R. W. (1995). Affective computing. MIT Media Lab Technical Report No. 321. https://affect. media.mit.edu/pdfs/95.picard.pdf.

[5] Shuiping Wang, FEEL: Fast and Effective Emotion Labeling, a Dual Ensemble Approach for Effective Facial Emotion Recognition, PeerJ Computer Science, 2025, 11:e3138, https://doi.org/10.7717/peerj-cs.3138.

[6] Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha (2021), A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions.

[7] Clément Gérard Daniel Darne, Changqin Quan, and Zhiwei Luo, Generating Synthetic Facial Expression Images Using EmoStyle, Applied Sciences, 2025, 15(19), https://doi.org/10.3390/app151910636.

[8] Boutros, F., Huber, M., Siebke, P., Rieber, T., Damer, N., SFace: Privacy-friendly and Accurate Face Recognition using Synthetic Data, arXiv 2022, arXiv:2206.10520, [Google Scholar] [CrossRef].

[9] N. C. Ebner, M. Riediger, and U. Lindenberger, FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation, Behav. Res. Methods, vol. 42, no. 1, pp. 351–362, Feb. 2010, doi: 10.3758/BRM.42.1.351.

[10] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, Data aug-mentation for face recognition, Neurocomputing, vol. 230, pp. 184–196, Mar. 2017, doi: 10.1016/J.NEUCOM.2016.12.025.

[11] QIAO Guifang, HOU Shouming, LIU Yanyan, Facial expression recognition algorithm based on combination of improved convolutional neural network and support vector machine, Journal of Computer Applications, 10.11772/j.issn.1001-9081.2021071270.

[12] K. Chengeta and S. Viriri, A survey on facial recognition based on local directional and local binary patterns, Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS), Mar. 2018, pp. 1–6, 10.1109/ICTAS. 2018. 8368757.

[13] Anvita Saxena, Ashish Khanna, and Deepak Gupta (2020), Emotion Recognition and Detection Methods: A Comprehensive Survey.

[14] Q. Mao, M. Dong, Z. Huang and Y. Zhan, Learning Salient Features for Speech Emotion Recognition Using Con utional Neural Networks, in IEEE Transactions on Multimedia. 16(2014). 2203-2213. 10.1109/TMM.2014.2360798.

[15] Huang, Zhengwei Dong, Ming Mao, Qirong Zhan, Yongzhao, Speech Emotion Recognition Us- ing CNN. MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia (2014). 801-804. 10.1145/ 2647868. 2654984.

[16] E. Chaudary, S. A. Khan andW. Mumtaz, EEG-CNN-Souping: Interpretable emotion recognition from EEG signals using EEG-CNN-souping model and explainable AI, Comput. Electr. Eng., Volume 123, Page 110189, April 2025, doi: 10.1016/j.compeleceng.2025.110189.

[17] R. M. Mehmood and H. J. Lee, EEG based Emotion Recognition from Human Brain using Hjorth Parameters and SVM, Int. J. Bio-Sci. Bio-Technol., Volume 7, Issue 3, Page 23-32, June 2015, doi: 10.14257/ijbsbt.2015.7.3.03.

[18] Downey, S., Russell, M.J., A Decision Tree Approach to Task Independent Speech Recognition, Proceedings of the Inst Acoustics Autumn Conf on Speech and Hearing (1992).

[19] Chen, G., Hou, R., A New Machine Double-Layer Learning Method and Its Application in non-Linear Time Series Forecasting, Proceedings of the 2010 International Conference on Mechatronics and Automation (ICMA) (2007).

[20] Rani, P., Sarkar, N., Smith, C.A., Kirby, L.D., Anxiety detecting robotic system towards implicit human-robot collaboration, Robotica 22, 85–95 (2004).

[21] Tianqi Fan, Sen Qiu, Zhelong Wang, Hongyu Zhao, Junhan Jiang, Yongzhen Wang, Junnan Xu, Tao Sun, Nan Jiang, A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition, 01 June 2023 Publication History, https://doi.org/10.1016/j. compbiomed.2 023. 106938.

[22] Kevin G Montero Quispe, Daniel M S Utyiama, Eulanda M dos Santos, Horácio A B F Oliveira, Eduardo J P Souto, Applying Self-Supervised Representation Learning for Emotion Recognition Using Physiological Signals, Sensors (Basel). 2022 Nov 23;22(23):9102. doi: 10.3390/s22239102.

[23] Purnata Saha, Ali K. Ansaruddin Kunju, Molla E. Majid, Saad Bin Abul Kashem, Mohammad Nashbat, Azad Ashraf, Mazhar Hasan, Amith Khandakar, Md Shafayet Hossain, Abdulrahman Alqahtani, Muhammad E.H. Chowdhury, Novel multimodal emotion detection method using Electroencephalogram and Electrocardiogram signals, https://doi.org/10.1016/j.bspc.2024.106002.

[24] S. Kusal, S. Patil, K. Kotecha, R. Aluvalu and V. Varadarajan, AI Based Emotion Detection for Textual Big Data: Techniques and Contribution, Big Data Cogn. Comput., Volume 5, Issue 3, Page 43, September 2021, doi: 10.3390/bdcc5030043.