Review Of Emotion Recognition Technology Methods Based on Deep Recognition

Zhizhou Lu *

Nanjing University of Posts and Telecommunications, Nanjing, 210023, Jiangsu, China

* Corresponding Author Email: 2897694110@qq.com

Abstract. With the rapid advancement of artificial intelligence, sensor technology, and big data analytics, the way humans interact with machines is shifting from an "instruction-based" mode to a "perception-based" one. As emotions are the core driving force behind human behaviors and decisions, the objective and quantitative detection of emotions has become a focus of interdisciplinary research. Studies have shown that emotions can be detected and analyzed through four methods: facial expression recognition, speech emotion recognition, text sentiment analysis, and physiological signal recognition. Reducing errors in emotion recognition is of great help to the development and widespread application of human-computer interaction. The development and research on emotion detection is not only an inevitable outcome of technological development but also a key to addressing practical needs. This paper will analyze deep learning models such as CNN, LSTM, and SENN, and summarize their advantages and disadvantages. It breaks the traditional perception that "emotions cannot be quantified," enabling machines to move from "understanding language" to "understanding the human heart," and ultimately promoting the harmonious coexistence of technology and human society.

Keywords: emotion recognition; reducing recognition errors; human-computer interaction; deep learning technology.

1. Introduction

The emergence and development of emotion recognition technology are the result of various factors working together, including advancements in artificial intelligence technology, increasing demand for human-computer interaction, and the integration of multiple disciplines.

The reason of why we need to promote the development of emotion recognition technology is multifaceted. The first reason is the growing demand for human-computer interaction. Emotion recognition is ready to make significant contributions to artificial intelligence and human-computer interaction.[1] Contemporary users increasingly expect emotionally intelligent interactions beyond functional exchanges. [1,2] Emotion recognition enables machines to identify human emotional states[3], thereby achieving more natural and intelligent human-computer interaction. For example, if a smart speaker detects anxiety in the user's tone, it will proactively offer soothing music or suggestions. In other words, the promoted machines can perceive human joys, sorrows, anger, and happiness, and then adjust interaction strategies.

Besides, we can obviously find out that as emotion recognition technology developing, a new dimension of insight is provided for various industries, Promote the transformation of service models from "generalization" to "personalization" and "precision". Emotion recognition technology has many applications in daily life. For instance, an emotion recognition system can be used into a car's onboard driving system. This can help prevent accidents caused by a driver's stressed mental state.[4]

This paper aims to gather and evaluate significant emotion recognition techniques developed in recent years. It discusses three main types of emotion recognition and detection methods, which perform emotion recognition through text, images, and speech. The three methods are shown in Figure 1. It also collects innovative recognition technologies and optimized models, compares and analyzes the pros and cons of various models, aiming to provide a systematic contribution to the development of emotion recognition technology.

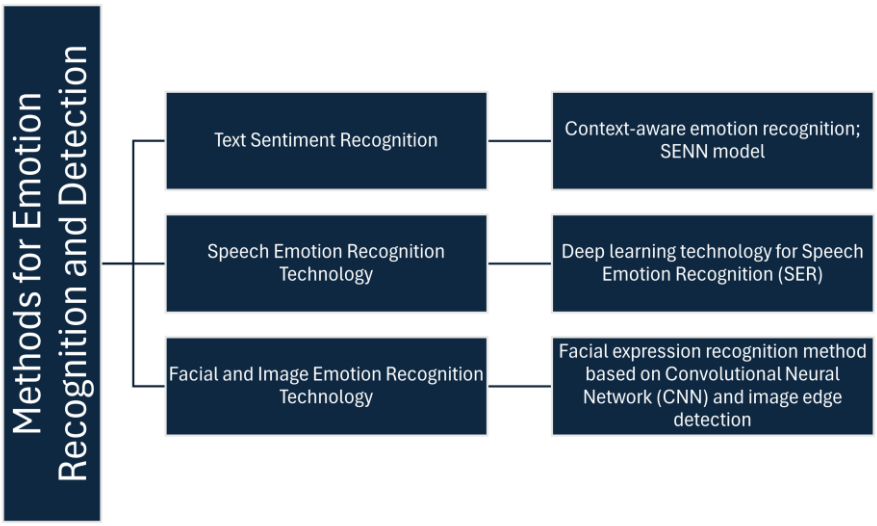

Figure 1. three methods of emotion recognition

The structural framework of this paper is shown below:

Point out the defects of current text sentiment recognition, and introduce context-aware emotion recognition and the SENN model. Compare traditional facial emotion recognition technologies with deep learning-based methods, and highlight the advantages of deep learning in facial emotion recognition (FER). Analysis begins with traditional speech emotion recognition (SER) technologies and provides a comprehensive comparison of the pros and cons of various deep learning-based SER methods.

2. Optimization Based on Text Sentiment Recognition Technology

The sentiment of text is complex, with a large number of ambiguous expressions. For example, the emotional tendency of neutral words depends on the context. The CCIM module proposed by Dingkang Yang eliminates background interference by constructing a causal graph (X-S-C-Z-Y variables) [5]. This idea is consistent with the goal of "improving the robustness of emotion recognition" focused on in this paper — especially in facial expression recognition in multi-person scenarios, it can effectively prevent the model from mistakenly incorporating the expression features of background characters into its judgment. Besides, Erdenebileg Batbaatar introduced a new neural network architecture named Semantic-Emotion Neural Network (SENN). This model uses both semantic and emotional data through the use of pre-trained word embeddings.[1]

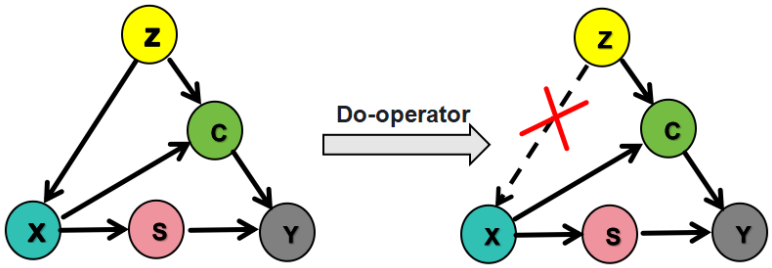

Figure 2. the CAER causal graph

Existing emotion recognition models rely on multi-modal information from subjects and backgrounds. However, these methods fail to consider the bias in background information, which may cause the models to over-rely on certain contextual features and affect the accuracy of emotion recognition. By utilizing Context-Aware Emotion Recognition (CAER) and through a Contextual Causal Intervention Module (CCIM), As shown in Figure 2, this method constructs a causal graph for the CAER task, involving variables such as input images (X), subject features (S), contextual features (C), confounding factors (Z), and predictions (Y).[5]

It clarifies the causal relationships between variables, effectively eliminates the adverse impact of contextual bias, and perceives the emotional state of the target person more accurately. Meanwhile, the Contextual Causal Intervention Module (CCIM) module eliminates the influence of confounding factors based on backdoor adjustment and existing Context-Aware Emotion Recognition (CAER) methods rely on the like- lihood P (Y |X). This process is formulated by Bayes rule:

$$P(Y|X)=\sum_{z} P(Y|X,S=f_s(X),C=f_c(X,z))P(z|X) \tag{1}$$

This formula is adapted from [5].

In the formula, two generalized encoding functions are f s(·) and f c(·) , which are used to obtain the subject feature S and context feature C from the input image X respectively; the confounding factor Z will introduce observational bias through the probability P(z|X), affecting the accuracy of the model's judgment on emotion Y. To eliminate the interference of Z, the model is enabled to accurately predict emotion Y based solely on input X, ensuring every context's semantics make equally to emotion prediction, thereby letting model's recognition accuracy higher.[5] Kang Yang uses backdoor adjustment to stratify Z, achieving causal intervention P (Y | do(X)) and training the model with real causal effects.

Word embedding is widely used in many natural language processing tasks. Training a single model with semantic or emotional word embedding yields good results. However, when embedding words with similar semantic functions are confused in emotional states. This makes it impossible to effectively encode and learn semantic and emotional relationships in short texts.

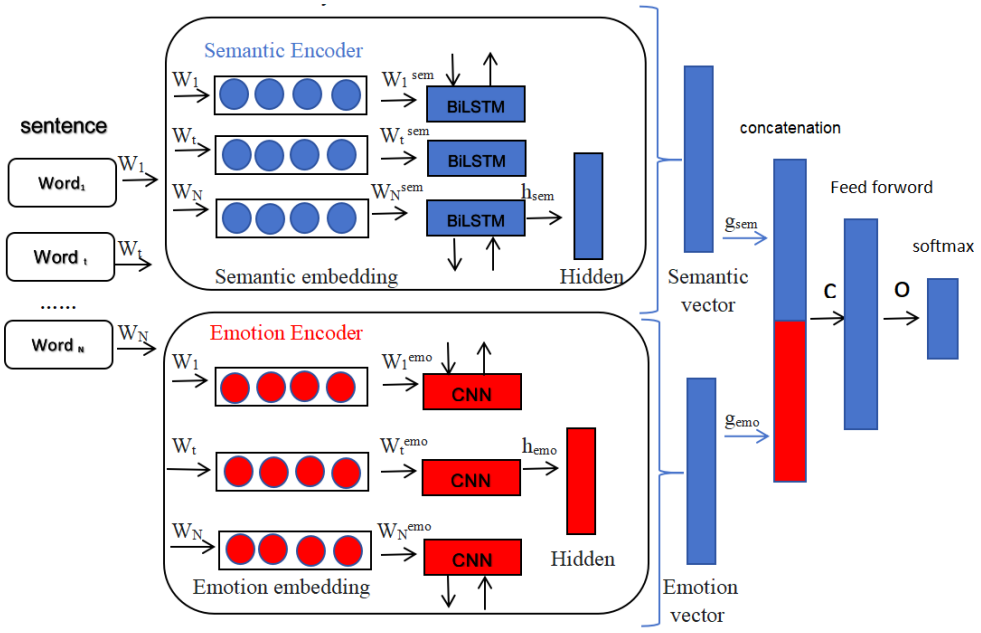

Figure 3. SENN model architecture

The emergence of the SENN model breaks the limitations of single representation and emotional context in traditional research. SENN adopts a dual-path architecture: a **BiLSTM-based semantic encoder** and a **CNN-based emotional encoder**, enabling parallel extraction and fusion of features [1], and its structure is shown in Figure 3. It abandons the disadvantages of manual feature models, such as poor adaptability and time-consuming processes, and introduces a dual-network structure. By processing the same input text simultaneously, it extracts semantic and emotional information respectively and merges the outputs, thereby achieving more accurate emotion recognition.[1] Utilizing BiLSTM-based semantic encoder and CNN-based emotional encoder, and the output of the subnetwork is used to identify emotions in the text.

This architecture processes text through dual networks in parallel, extracting semantic and emotional features separately before fusing them, thus addressing the limitation of traditional models

that rely on a single type of feature. What's more, SENN can improve performance by fine-tuning pre-trained word embeddings.[1]

3. Optimization Based on Facial and Image Emotion Recognition Technology

Unlike text, which relies on contextual semantics, facial expression recognition is more dependent on the fine extraction of visual features. Therefore, the CNN architecture in deep learning, through the combination of edge detection and convolutional layers, demonstrates advantages in retaining texture information that traditional methods can hardly match.

With the rapid development of computer technology and the continuous growth of data volume, computer vision has become an emerging research direction. Among them, face detection, recognition and expression analysis technologies are one of the most active research directions in the current computer vision community.[6]

Traditional facial expression recognition requires first locating the face and facial features from the image, then extracting spatial and temporal features from these regions, and finally outputting the recognition results according to the features by pre-trained classifiers such as Support Vector Machine (SVM), AdaBoost, and Random Forest.[7]

Unlike traditional methods that rely on handcrafted features, deep learning has become a leading approach in machine learning, achieving top-tier results in numerous computer vision studies, thanks to the availability of big data. [8]

3.1 FER Based on Deep Learning

A method for recognizing facial expressions using Convolutional Neural Networks (CNN) and image edge detection: First, perform image preprocessing; then, CNNs gradually extract facial texture features through hierarchical convolution [9], which is similar to a progressive detail magnification process; subsequently, simplify the features and process and compress the extracted features.Finally, use the Softmax classifier to classify and recognize the expressions of the test sample images.[9] CNN structure for facial expression recognition is shown in Figure 4.[9]

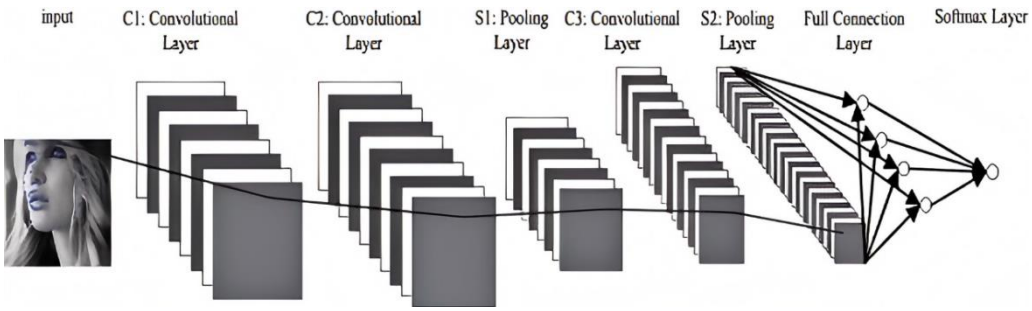

Figure 4. CNN architecture for facial expression recognition[9]

Suci Dwijayanti proposed a real-time application scheme combining face and emotion recognition. Specifically, a convolutional neural network architecture is used to simultaneously develop face and emotion recognition systems. This model is compared with well-known architectures such as AlexNet and VGG16 to determine which is more suitable for implementation in humanoid robots. Table 1 [10] shows the accuracy obtained from the test data.

Table 1. Emotion recognition Adapted from [10]

No.	Emotion	Model A (AlexNet) 500 epochs	Model B (VGG16) 500 epochs	Model C (Proposed model) 500 epochs
1	Surprise	Recognized	Recognized	Recognized
2	Angry	Unrecognized	Recognized	Recognized
3	Neutral	Recognized	Recognized	Recognized
4	Sad	Recognized	Recognized	Recognized
5	Smile	Recognized	Recognized	Recognized
Accuracy		64%	82%	71%

After 500 rounds of training, the accuracy rate of VGG16 reaches 82%, which is higher than that of AlexNet and the Proposed model, indicating that the VGG16 architecture is more adaptable to the speech emotion recognition task. The accuracy rate of AlexNet is only 64%, which may be due to the relatively simple network structure and insufficient ability to extract and distinguish speech emotion features.[10] The accuracy rate of the Proposed model is lower than that of Model B, and there may be room for optimization in its network design (such as structural complexity, feature fusion method, etc.). All in all, in Dwijayanti et al.'s humanoid robot experiment [10], VGG16 showed the optimal recognition performance (82% accuracy). VGG16 excels in recognizing faces and emotions, making it suitable for implementation in humanoid robots. [10]

3.2 Image recognition technology

NVIDIA's Jetson Nano device has three functions: detect human faces, recognize human faces, and identify facial expressions. Among them, face detection is done by the deep learning-based DNN face detector in OpenCV. This detector uses the ResNet architecture and is much more accurate than previous models. With the support of the aforementioned hardware, even if the lighting and shooting angles change, the calculation results of the OpenCV framework library can still maintain reliable accuracy. As for face recognition, the deep metric learning method of OpenCV and ResNet-34 architecture is used.[6] Facial expression recognition is achieved by analyzing the eye and mouth regions:

Extract features of key parts

Fuse to generate a new image

Match the seven basic expressions

(The seven basic expressions include: Fear, Anger, Disgust, Happy, Sad, Surprise, Neutra).

Its advantage lies in that Jetson Nano is a powerful platform with low power consumption, which can easily perform intensive computing of algorithms and contribute to high video processing frames.[6]

4. Optimization Based on Speech Emotion Recognition Technology

Speech emotion recognition has evolved from a specialized area to an essential component of human-computer interaction (HCI). [2,11,12] Voice interaction facilitates more natural communication between humans and machines, which is different from the way traditional devices are used for input to make machines understand verbal content. Moreover, voice interaction makes it easier for human listeners to respond and comprehend.

Process speech signals to identify underlying emotions, understand the emotions present in speech, and synthesize the emotions required in speech based on expected information. It’s important to establish an appropriate emotional language database. By analyzing the impact of text information on emotional expression, there are three types of databases for better emotion recognition: actor-based (simulated) emotional speech databases, induced emotional speech databases, and natural emotional speech databases.[13]

Pattern recognizers for language emotions:

1. Linear classifier
2. Nonlinear classifier[13]

In speech emotion classification, the selection logic between linear classifiers and nonlinear classifiers is similar to that in speech emotion recognition (both depend on feature correlation, task complexity, etc.). However, compared with "emotion," "affect" emphasizes more delicate, complex, and context-dependent psychological states such as "relief," "grievance," and "jealousy." Nonlinear correlations between features are more common, so the selection of classifiers needs to focus more on capturing "subtle nonlinear patterns."

Speech emotion recognition using deep learning technology:

Traditional SER technologies:

To achieve speech emotion recognition, traditional SER relies on manually designed features (such as spectrum, prosody) and statistical model classification [13, 14], and its process includes: preprocessing → feature engineering → model inference. The core of this process lies in its reliance on manually designed features. It requires manual design of temporal features, resulting in limited modeling capabilities. Moreover, it is more sensitive to noise and individual differences, with restricted generalization performance.[13] The simplified structure of traditional speech emotion recognition is Figure 5 [13] shown below.

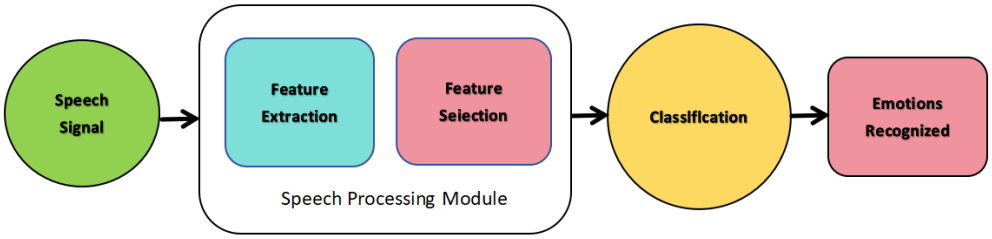

Figure 5. Traditional Speech emotion recognition [13]

However, the feature extraction of traditional SER relies on manual experience, and this limitation has promoted the application of deep learning models. Deep learning technology utilizes several key functions in the field of speech emotion recognition technology, as shown in Figure 6. Analysis and comparison of the respective advantages of Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RvNN), Deep Belief Network (DBN), etc. in emotion recognition.

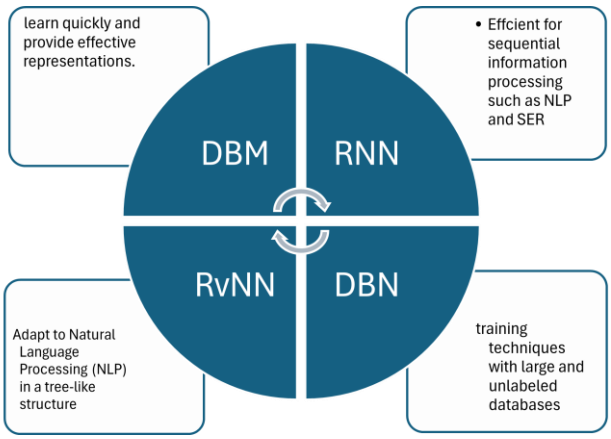

Figure. 6 Overview of Deep Learning Techniques and Their Key Features Adapted from Khalil et al. [14]

Deep Boltzmann Machine (DBM): DBM can be combined with other models, using unsupervised training to identify various speech emotions. The main advantage of DBM is that it it learns things quickly and can also provide useful information representation. However, DBM also has some disadvantages, such as limited effectiveness in certain cases.

Recurrent Neural Network (RNN): RNNs are suitable for speech emotion recognition due to their short-time frame segmentation of acoustic features.[15] However, When RNNs process long sequence data, the gradients gradually decay as they propagate over time, resulting in a decreased ability to capture long-distance information.[14]

Recursive Neural Network (RvNN): RvNN is mainly used in natural language processing, and its structure can handle different situations. Studies have found that such networks can not only classify sentences in natural language but also be applied to natural image processing.[14]

Deep Belief Network (DBN): DBN does not require manual feature design and can automatically learn multi-level abstract features from raw data. For example, it can extract spectrum, prosody, and emotion-related features layer by layer from speech signals. Deep Belief Networks (DBNs) have two main advantages. The first is that during pre-training, they can perform unsupervised learning on large-scale unlabeled databases. And they don't need the manual data labeling. The second is that they can approximate the required output weights of variables during the inference process.[14] Some limitations still exist, such as the inference process of Deep Belief Network (DBN) is limited to bottom-up transmission.

5. Conclusion

Although the optimization of these three types of emotion recognition technologies focuses on different data modalities, they share the same core goal: enhancing the precision and robustness of emotion perception. From a technical perspective, these optimizations have driven innovations in methods such as multimodal fusion, causal reasoning, and lightweight models. In terms of application value, their outcomes have supported the transformation of service models from "generalized" to "personalized" and "precision-oriented." Although deep learning has improved recognition accuracy, current technologies still face three major challenges: cross-modal emotional conflicts, insufficient support for low-resource languages, and energy consumption limitations of real-time systems. We hope that future research and development can overcome these challenges. We also hope that they will play a more critical role in fields such as education, medical care, retail, and public security in the future, and ultimately achieve the deep integration of technology and human emotions.

References

[1] Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-Emotion Neural Network for Emotion Recognition from Text. IEEE Access, 7, 111866-111878.

[2] Hossain, M. S., & Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49, 69-78.

[3] Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Information Fusion, 53, 209-221.

[4] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International Journal of Speech Technology, 15, 99-117.

[5] Yang, D., Chen, Z., Wang, Y., Wang, S., Li, M., Liu, S., Zhao, X., Huang, S., Dong, Z., Zhai, P., & Zhang, L. Context De-Confounded Emotion Recognition. IEEE, 19005-19015.

[6] Sati, V., Sánchez, S. M., Shoeibi, N., Arora, A., & Corchado, J. M. (2021). Face Detection and Recognition, Face Emotion Recognition Through NVIDIA Jetson Nano. In P. Novais, G. Vercelli, J. L. Larriba-Pey, F. Herrera, & P. Chamoso (Eds.), Ambient Intelligence – Software and Applications (pp. 1239). Springer.

[7] Ko, B. C. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors, 18(2), 401.

[8] Kahou, S. E., Michalski, V., & Konda, K. (2015). Recurrent neural networks for emotion recognition in video. In Proceedings of the ACM on International Conference on Multimodal Interaction (pp. 467-474). Seattle, WA, USA: ACM.

[9] Zhang, H., Jolfaei, A., & Alazab, M. (2019). A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing. IEEE Access, 7, 159081-159089.

[10] Dwijayanti, S., Iqbal, M., & Suprapto, B. Y. (2022). Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network. IEEE Access, 10, 89876-89886.

[11] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61, 90-99.

[12] Chen, M., Zhou, P., & Fortino, G. (2016). Emotion communication system. IEEE Access, 5, 326-337.

[13] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International Journal of Speech Technology, 15, 99-117.

[14] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access, 7, 117327-117345.

[15] Chernykh, V., & Prihodko, P. (2017). Emotion recognition from speech with recurrent neural networks [Preprint]. arXiv:1701.08071