Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

A Survey of Deep Learning-based Facial Expression Recognition Research

Chengxu Liang, Jianshe Dong *

School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
* **Corresponding author**: Jianshe Dong (Email: dongjianshe@tute.edu.cn)

Abstract: Facial expression is one of the ways to convey emotional expression. Deep learning is used to analyze facial expression to understand people's true feelings, and human-computer interaction is integrated into it. However, in the natural real environment and various interference (such as lighting, age and ethnicity), facial expression recognition will face many challenges. In recent years, with the development of artificial intelligence, scholars have studied more and more facial expression recognition in the case of interference, which not only promotes the theoretical research, but also makes it popularized in the application. Facial expression recognition is to identify facial expressions to carry out emotion analysis, and emotion analysis can be analyzed with the help of facial expressions, speech, text, video and other signals. Therefore, facial expression recognition can be regarded as a research direction of emotion analysis. This paper focuses on the perspective of facial expression recognition to summarize. In the process of facial expression recognition, researchers usually try to combine multiple modal information such as voice, text, picture and video for analysis. Due to the differences between single-modal data set and multi-modal data set, this paper will analyze static facial expression recognition, dynamic facial expression recognition and multi-modal fusion. This research has a wide range of applications, such as: smart elderly care, medical research, detection of fatigue driving and other fields.

Keywords: Expression Recognition; Deep Learning; Multimodality.

1. Introduction

In recent years, facial expression recognition has been one of the important research topics in computer vision, through the recognition and analysis of various facial expressions, to determine a person's emotional state. This will not only help us better understand human communication, but also expand the scope of human-computer interaction to include more emotional elements of people. Facial expression recognition originated from psychological research. In 1872, Darwin [1] proposed for the first time that human expression evolved from the expression features of animals, and also elaborated the correlation between humans and animals. Since then, expression recognition technology began to rise, and research on expression continues until now. In the 1970s, Ekman and Friesen [2] defined human expressions into six types through research, including happiness, surprise, sadness, fear, disgust, and anger, and first proposed the facial behavior coding system, the most important of which is to use computers to recognize facial expressions. In recent years, with the rapid development of artificial intelligence and deep learning, more and more experts and scholars pay more attention to facial expression recognition.

The advent of the data era has promoted the explosive growth of multimodal data, and many researchers have carried out the work of establishing multimodal data sets to provide data support and research basis for future experiments. Multimodal emotional information can be obtained from different emotional expression modes, including video, voice, text, body posture, walking style and facial expression. Multimodal emotion recognition uses clues from various modes in the same data segment to identify emotions, and uses the complementarity between modes to eliminate ambiguity in emotion recognition [3]. Early expression recognition mainly focused on the expression recognition of single-mode data. Although the research effect is getting better and better, the emotion information contained in a single mode is not comprehensive, and the complex emotions expressed by human beings cannot be accurately recognized. With the development of multimedia, people will use multimedia platforms to share daily life and express complex emotions through rich channels such as images, videos and texts.

Cai et al. [4] combined speech and facial expression features, used CNN and LSTM to learn the emotional features of speech, designed multiple small-scale nuclear convolution blocks to extract facial expression features, and finally used DNN to integrate the two. Li et al. [5] used several multimodal fusion strategies to combine various features of acoustic, visual and textual modes. Mittal et al. [6] combinational cues from multiple simultaneous modes such as face, text, and speech. It can be seen that the fusion of information of multiple modes can achieve information supplement, improve the accuracy of prediction results, improve the robustness of prediction models, and make the final results more reliable.

Facial expression recognition is widely used in many fields, such as medicine, education, games, security and entertainment. In medicine, for example, doctors could use the technology to help identify what patients are really feeling when they can't accurately express their pain. In the field of education, teachers can analyze students' facial expressions to determine whether they truly understand what is being taught; In the entertainment industry, facial expression recognition can create more interactive and immersive experiences.

However, facial expression recognition also has its challenges and issues, including accuracy, privacy concerns, and cultural differences. Despite these challenges, we must also acknowledge that facial expression recognition offers great potential to enrich and enhance human-computer

interaction.

2. Single Mode Facial Expression Recognition

Facial expression recognition based on single mode refers to emotion classification through the analysis and recognition of facial expression, which mainly relies on the data of the single mode of facial expression for analysis. First, we need to collect the data set, then preprocess the image, and then use machine learning or deep learning algorithms to recognize and determine the category of facial expressions. Through these steps, we can effectively understand and interpret the emotional information conveyed by facial expressions. Figure 1 shows the flow chart of facial expression recognition based on single mode.

Figure 1. Single-mode facial expression recognition flow chart

2.1. Correlation Data Set

This paper lists common unimodal expression recognition data sets, and provides descriptions of their acquisition methods, sample numbers and annotated categories. Some of the data sets listed in the table were collected under laboratory shooting. The common features of these data sets are small data size and clear frontal face images, whose annotations have been repeatedly confirmed by psychological experts. Therefore, it is generally considered to be reliable, such as CK+ and JAFFE. However, there are also data sets, such as RAF-DB and FER2013, which are large-scale data sets collected in uncontrolled environments, and the quality of these data sets is relatively low, which will be greatly affected by the subjective feelings of the tagger. Common unimodal data sets are shown in Table 1.

Table 1. Common unimodal data sets

Dataset	Access	Sample quantity	Expression category
FER2013[7]	Network collection	35887	7
CK+[8]	Laboratory shooting	593	8
JAFFE[9]	Laboratory shooting	213	7
SFEW2.0[10]	Video capture	1766	7
Multi-PIE [11]	Laboratory shooting	755370	6
BU-3DFE [12]	Laboratory shooting	2500	7
Oulu-CASIA [13]	Laboratory shooting	2880	6
RaFD[14]	Laboratory shooting	1608	7
KDEF[15]	Laboratory shooting	4900	7
RAF-DB [16][17]	Network collection	29672	7+12
BAUM-1s [18]	Laboratory shooting	1222	8
RML[19]	Laboratory shooting	720	6
DFEW[20]	Video capture	16372	7

Current data sets are scarce in quantity and quality, and the data volume is small enough to train well on large deep network structures. The current lack of large-scale facial expression datasets with occlusive types and head pose labeling also affects the ability of deep networks to address gaps within larger classes and efficiently identify facial features.

2.2. Recognition Network

The facial expression recognition method based on deep learning is applied to the facial expression recognition of static images and dynamic videos. Due to the relative scarcity of video data sets and the difficulty in processing the relationship between frames, there are few methods to train video data sets on deep learning models, most of which are facial expression recognition of static images. The single-mode facial expression recognition method is shown in Table 2.

Despite the exploration and research of these methods, the accuracy rate of facial expression recognition based on single mode is generally low, and it is still in the research stage and cannot be widely used in real life. Therefore, we should continue to study and improve the recognition method, constantly optimize and improve the accuracy, so that it can be better applied to practical scenarios.

Table 2. The single-mode facial expression recognition method

Network model	Reference
CNN+LSTM+Transformer	[21]
VGG+SE	[22]
VGG19+CBAM+IGN+DEN	[23]
AREN+CSAF+ Zonal loss	[24]
ResNet+CSAM+GAP	[25]
CBM-Net	[26]
FRR-CNN	[27]
SqueezeNeXt	[28]
ShuffleNetV3	[29]
3DResNet18	[30]
ResNet-VIT	[31]

3. Multimodal Facial Expression Recognition

Multi-modal facial expression recognition refers to the use of multiple information sources or multiple modal data for facial expression recognition and classification. Facial expression recognition based on single mode only uses the information of the visual mode of facial expression for analysis. However, facial expressions are often accompanied by other modes of information, such as text, voice, movement, and EEG signals. Therefore, multi-modal facial expression recognition uses the information relationship between different modes to improve the recognition accuracy and enhance the robustness.

The methods of multimodal facial expression recognition mainly include two aspects: modal fusion and multimodal learning. Modal fusion is the fusion of features of different modes, which can be either feature level fusion, decision level fusion, or hybrid. Multimodal learning is to synthesize the information of different modes by means of joint modeling or joint training.

The use of multi-modality can more comprehensively understand the emotional state of people and provide more accurate recognition results. Although multimodal facial expression recognition faces challenges such as data set

acquisition, modal fusion, and model optimization, it has broad application prospects and can provide more possibilities and solutions for sentiment analysis, human-computer interaction, virtual reality and other fields. Figure 2 shows the flow chart of facial expression recognition based on multimode.

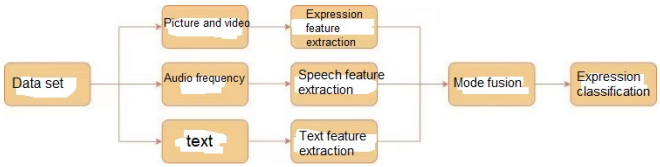

Figure 2. Multimodal facial expression recognition flow chart

3.1. Section Headings

Several common multimodal data sets are summarized. In the following multimodal data sets, most of them have expression pictures or videos, assisted by one or more modes such as audio, text and EEG signals, so as to improve the recognition accuracy. The following data sets include laboratory collection, talk show collection, news video collection, natural environment recording, film and television clips, etc. Among them, the data mode conforms to the representative meaning: video (V), physiological signal (PS), audio (A), text (T), body movement (BM), facial movement (FM), image (I), electroencephalogram(E) etc. Common multimodal data sets are shown in Table 3.

Table 3. Common multimodal data sets

Dataset modality	Dataset	Access	Sample distribution	Expression category
V, PS	DEAP [32]	Laboratory shooting	Thirty-two test subjects	9
V, A, T	CASIA [33]	Talk show collection	9,600 different sound fragments	6
V, A, BM, FM, T	IEMOCAP [34]	Laboratory shooting	12 hours of audio-visual data	Category + Dimension
V, A	SAVEE [35]	Laboratory shooting	480 recordings	7
T, I, A	News Rover Sentiment [36]	News video capture	929 videos	3
V, BM, A	AFEW [37]	Natural environment recording	1,809 videos	7

3.2. Recognition Network

There are many kinds of facial expression recognition methods based on multi-mode, among which the most important is the fusion between modes, modal fusion is divided into three ways: feature fusion, decision fusion and hybrid fusion. Effective multi-modal fusion can share and complement information between different modes, and thus improve the accuracy of single mode recognition.

The process of feature fusion is to integrate the corresponding features obtained from different modal data through feature extraction. This method can effectively learn the correlation and complementarity between different modal features. Decision fusion is to use the output data obtained after the deep learning model is trained on different modal features as the input of the regression model in the next stage. Hybrid fusion is a combination of feature-level fusion and decision-level fusion. The network model is very difficult. Although it combines the advantages of feature fusion and decision fusion, it is relatively poor in practical application.

Table 4. The multimodal facial expression recognition method

Dataset modality	Network model	Reference
V+A	LBP+ Random Forest model	[38]
	VGGNet-19	[39]
	Deep learning algorithm +Gabor	[40]
V+E	Histogram equalization +LBP	[41]
	BCN	[42]
RGB+3D(I)	DFDN	[43]

The current facial expression recognition methods applied in bimodal and multi-modal are summarized as follows. There are many methods also studying multi-modal facial expression recognition. In this field, multi-modal recognition is challenging and innovative, and it will be widely used in the future. The multimodal facial expression recognition method is shown in Table 4.

4. Conclusion

With the continuous improvement of computer processing power, deep learning network and fusion algorithm, expression recognition based on multi-modal data will be rapidly developed, but there are still some shortcomings in the progress, such as:

(1) The multi-modal facial expression data set is seriously insufficient, and the distribution of data categories is unbalanced. At present, the data of each expression in the existing expression database is relatively small, and they are very deliberate, and the expression is not natural, there are certain differences with the expression in the natural situation, it is difficult to become a very accurate and effective data, and the dynamic sequence image is seriously lacking. Category distribution happy, happy recognition rate is high, anger, contempt recognition rate is low.

(2) Research sites are mostly laboratories, lacking training in real situations. Most of the research on expression recognition is carried out under ideal conditions. However, due to the natural environment will block objects, block faces, different brightness at different times, and other circumstances such as the surrounding environment, will have a greater impact on the facial expression recognition results, and eventually lead to the actual results and experimental results are different.

(3) There are differences in the faces of different races. Because each person's nationality, age, growth conditions and other factors will affect the correctness of the identification. And there are differences in the habits of people of different races, which makes it difficult to use a unified model to classify faces, increasing the difficulty of recognition.

(4) There are still problems in the optimization of effective fusion methods between modes. Can not integrate the

information between multiple modes well, multi-angle, multi-faceted analysis of the facial expression at this time closer to which category

Acknowledgments

I would like to thank the school and my tutor for their guidance and training of my research, and the anonymous reviewers for their review of this paper.

References

[1] Darwin C. The expression of the emotions in man and animals[M]//The expression of the emotions in man and animals. University of Chicago press, 2015.

[2] Ekman P , Friesen W V . Facial action coding system: A technique for the measurement of facial movement[J]. a technique for the measurement of facial action, 1978.

[3] Han W,Chen H,Gelbukh A, et al.Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis [C]. Proc.of the 2021 International Conference on Multimodal Interaction.2021.6-15.

[4] Cai L, Dong J, Wei M. Multi-modal emotion recognition from speech and facial expression based on deep learning[C].Proc. of the Chinese Automation Congress (CAC). 2020.5726-5729.

[5] Li R, Zhao J, Hu J,et al.Multi-modal Fusion for Video Sentiment Analysis[C]. Proc. of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. 2020.27-34.

[6] Mittal T,Bhattacharya U,Chandra R,et al.M3ER:Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues[C].Proc. of the AAAI Conference onArtificial Intelligence. 2020.1359-1367.

[7] GOODFELLOW I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[C]//20thInternational Conference on Neural Information Processing, Daegu, Korea, 3-7 November, 2013. UK, Neural Networks, 2015, 64: 59-63.

[8] LUCEY P, COHN J F, KANADE T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]// Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern RecognitionWorkshops. San Francisco, 2010. Los Alamitos, IEEE: 94-101.

[9] LYONS M, AKAMATSU S, KAMACHI M,et al..Coding facial expressions with gabor wavelets[C]// Third IEEE International Conference onAutomatic Face and Gesture Recognition, Nara Japan, April 14-16 1998. Los Alamitos, IEEE Computer Society ,1998:200-205.

[10] Levi G, Hassner T. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns [C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. New York: ACM, 2015: 503-510.

[11] GROSS R, MATTHEWS I, COHN J, et al. Multipie[J]. Image and Vision Computing,2010.28(5): 807-813.

[12] YIN L, WEI X, SUN Y, et al. A 3d facial expression database for facial behavior research[C]//Automatic face and gesture recognition, FGR 2006 7th international conference, Southampton, UK,10-12 April 2006. Piscataway, IEEE, 2006: 211-216.

[13] ZHAO G, HUANG X, TAINI M, et al. Facial expression recognition from near-infrared videos[J]. Image and Vision Computing, 2011, 2(9): 607-619.

[14] LANGNER O, DOTSCH R, BIJLSTRA G, et al. Presentation and validation of the radboud faces database[J]. Cognition and Emotion, 2010, 24(8): 1377- 1388.

[15] LUNDQVIST D, FLYKT A, OHMAN A.The karolinska directed emotional faces (kdef)[M/CD].CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Sweden,1998.

[16] LI S,DENG W, DU J. Reliable crowdsourcing and deep locality preserving learning for expression recognition in the wild[C]//in IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Venice, Italy, 21-16 July, 2017. Piscataway, IEEE, 2017, 2584-2593.

[17] LI S,DENG W. Reliable crowdsourcing and deep locality preserving learning for unconstrained facial expression recognition [J]. IEEE Transactions on Image Processing, 2018.

[18] Wang Y J, Guan L and Venetsanopoulos A N. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia, 14 (3): 597-607 [DOI: 10. 1109 / TMM. 2012. 2189550].

[19] Zhalehpour S, Onder O, Akhtar Z and Erdem C E. 2017. BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 8(3): 300-313 [DOI: 10. 1109 / TAFFC. 2016. 2553038.

[20] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2881–2889, 2020. 1, 5, 6.

[21] Chen G, Zhang S Q and Zhao X M. 2022. Video sequence-based human facial expression recognition using transformer networks. Journal of Image and Graphics,27(10):3022-3030[DOI:10. 11834 / jig. 210248].

[22] GAN YCHEN J, YANG Z,et al. Multiple attention net-work for facial expression recognition[J]. IEEE Access. 2020. 8: 7383-7393.

[23] Guo Jingyuan, Dong Yishan, Liu Xiaowen, Lu Shuhua. Attentional mechanism and Involution operator improved facial expression recognition [J/OL]. Computer engineering and applications.

[24] Chen Gongguan, Zhang Fan, Wang Hua, Fan Hui, Zhang Caiming. Facial expression recognition in region-enhanced attention networks [J/OL]. Journal of Computer-Aided Design and Graphics.

[25] Guo Xin-Gang, Cheng Chao, Shen Ziqi. Facial expression recognition based on convolution network attention mechanism [J/OL]. Journal of jilin university (engineering science). https://doi.org/10.13229/j.cnki.Jdxbgxb20221345.

[26] Liu Cheng-Guang, WANG Shan-Min, LIU Qingshan. Category balance modulation of facial expression recognition [J/OL]. Computer science and exploration. https://kns. cnki. net/kcms/detail//11.5602.TP.20230203.1654.004.html.

[27] ZHANG J, ZHENG Y,QI D.Deep spatio -temporal residualnetworks for citywide crow d flows prediction[C] / /Proceed-ings of the thirty -one the association for the advance of arti1ficial intelligence. San Francisco: AAAI Association , 2017:1655-1661.

[28] Cholami A, Kwon K, Wu B, et al. SqueezeNext: Hard-ware-aware neural network design [EB].arXiv: 1803. 10615. 2018.

[29] Howard A, Sandler M , Chu G, et al. Searching for Mobile-NetV3[EB]. arXiv:1905.02244,2019.

[30] Kensho Hara, Hirokatsu K ataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet In CVPR, pages 6546- 6555, 2018.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770-778, 2016.

[32] KOELSTRA S, MUHL C, SOLEYMANI M, et al.. Deap: A database foremotion analysis using physiological signals[J]. IEEE transactions on affective computing, 2011, 3(1): 18-31.

[33] LI Y,TAO J H,CHAO L L, et al.. CHEAVD: a Chinese natural emotional audio-visual database[J]. Journal of Ambient Intelligence and Humanized Computing, 2017, 8(6):913-924.

[34] BUSSO C, BULUT M ,LEE C, et al..IEMOCAP: Interactive emotional dyadic motion capture database[J], Journal of Language Resources and Evaluation, 2008, 42(4), 335-359.

[35] WU M , SU W J, CHEN L F, et al.. Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition [J]. IEEE Transactions on Affective Computing, 2020.

[36] ELLIS J G, JOU B, CHANG S F. Why we watch the news: a dataset for exploring sentiment in broadcast video news[C]. Proceedings of the 16th international conference on multimodal interaction, Istanbul Turkey, Nov 12-16,2014. New York: ACM, 2014:104-111.

[37] DHALL A,GOECKE R,LUCEY S,et al..Collecting large, richly annotated facial-expression databases from movies [J]. IEEE Multi Media,2012,19(3):34-41.

[38] WEI F G, ZHANG S D, FU X H. audio-visual bimodal emotion recognition based on emotional tone[J]. Computer Applications and Software, 2018, 35(8): 238-242.

[39] SONG G J, ZHANG S D, WEI F G. Research on audio-visual dual-modal emotion recognition fusion frame-work[J]. Computer Engineering and Applications, 2020, 56(6):140-146.

[40] ZHANG L. Multimodal emotion recognition based on face and speech and the application in reasoning of robot service tasks [D]. ShanDong University,2021.

[41] SHEN J. Bimodal emotion recognition system based on EEG and facial expression[D]. Nanjing: Nanjing University of Posts and telecommunications,2020.

[42] ZHAO Y F, CHEN D Y. Expression EEG multimodal emotion recognition method based on the bidirectional LSTM and attention mechanism[J]. Computational and Mathematical Methods in Medicine, 2021 (2021): 9967592, 1-12.

[43] Xinwang Li. Human sentiment analysis with multimodal information fusion[D]. Guangzhou: Guangdong University of Technology, 2022.