Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230760

A survey of facial expression recognition in the wild

Chengjun Duan

Northeastern University Department of Computer and Communication linyi, China

2754605340@qq.com

Abstract. With the continuous development of facial expression technology, especially the development of deep learning, and the establishment of in-the-wild datesets in recent years, field face recognition has become a hot research field in the wild of facial expression recognition. Different from the traditional facial expression recognition(FER), in-the-wild facial facial expression recognition, the problem of recognition accuracy caused by illumination, occlusion or low image resolution. Therefore, in order to solve these problems, new methods have been put forward continuously in recent years. In this paper, we first continue to summarize the widely used datesets, and then summarize the paper methods of facial expression recognition in the field proposed in the past two years.

Keywords: facial expression recognition, deep learning, in-the-wild dateset, convolutional neural network(CNN).

1. Introduction

Expression is an important thing used to convey information in human life. Therefore, the technology of recognizing human facial expressions can be used in many practical applications, such as human-computer interaction[1], automatic driving[2], medical treatment[3], and so on, which plays a significant role in promoting human social development. Therefore, with the development of science and technology, facial expression recognition has become a hot research field. In addition, with the maturity of machine learning technology in recent years, deep learning has been more and more applied to the field of facial expression recognition. For example, the development of convolutional neural network has basically replaced the traditional facial expression recognition method with its excellent performance in the field of facial expression recognition.

Over the past 20 years, datesets for facial expression recognition have been built. Most of the images in the first datesets were taken exclusively in the laboratory. The datesets and face images obtained in the laboratory are accurate and clear. For example, the CK+ dateset[4]. However, with the increasing improvement of CNN technology applied to the recognition of laboratory datesets in recent years, and the actual facial expression recognition will always encounter the interference of uncontrollable influencing factors such as illumination, pose, occlusion and so on, which leads to the failure of the trained datesets in the laboratory to perform properly. Attention has turned again to identifying field datesets. For example, the RAF-DB dateset[5]. However, most of the images in-the-wild dateset are collected without constraints, so it is necessary to develop new convolutional neural networks for training and recognition. Based on the above background, this paper makes the following contributions:

DOI: 10.54254/2755-2721/6/20230760

1.We introduce the commonly used facial expression recognition datesets. The datesets were compared from five aspects: sample number, sample form, image pixel, collection source and expression classification.

2. We collate and compare the experimental data of all the papers described in this film, and through the comparison of experimental results, we draw some conclusions about facial expression recognition in the wild.

The rest of the paper is organized as follows. Section II briefly introduces the datesets commonly used in current papers. Section III describes The latest deep learning methods in facial expression recognition in the past two years.In section Ⅳ, the experimental data in the paper are analyzed and compared Finally, conclusions are drawn in Section Ⅴ.

2. Dateset

The traditional facial expression dateset is basically obtained by manual screening after the participants make corresponding expressions in the restricted environment such as laboratory. However, in the unrestricted case, the sample is closer to the face image of the real scene, which highlights the authenticity of the face. In this section, we briefly introduce the popular face datesets used in the field of facial expression recognition, and organize and compare these datesets from five aspects(As shown in Fig. 1): sample numbers, sample forms, image pixels, collection source, and expression classification.

2.1. The dateset collected in the experimental environment

Some textThe Multi-PIE dateset[6] (Multi Pose, Illumination, Expressions dateset) is a large dateset of facial expressions from different poses and light angles. The dateset contains 750,000 photos taken by 337 volunteers from various angles and under various lighting conditions, and labels the images with six expressions: disgust, neutral, scream, smile, squint and surprise.

The IJB-A dateset[7] (IARPA Janus Benchmark-A dateset) is A dateset for face detection and recognition published by the National Institute of Standards and Technology (NIST) at the 2015 CVPR, which opened the curtain of the face challenge around the dateset. The dateset contains 49,759 individuals and a total of 24,327 images. The face data in IJB-A is completely collected in an unconstrained environment, including both static images and video clips.

The AR dateset[8] (Active Record dateset) contains 4,000 color face images of 126 people, 56 women and 70 men, and all of the photos are from the front. The dateset contains four basic expressions: natural, happy, annoyed, ashamed and surprised. It also includes face images affected by different angles of illumination and partial occlusion.

The Oulu-CASIA dateset[9] (Oulu-CASIA NIR&VIS facial expression dateset) includes 2,880 images of 80 participants taken under three different lighting conditions: low light, normal light, and dark light. There are six basic expressions.

The CK+ dateset[10] (Cohn-Kanade plus dateset) facial expression dateset consists of 593 images. The pictures were taken by 123 people ranging in age from 18 to 30, more than half of whom were of European and American ethnicity. All of the images were tagged with actions, and 327 of them were tagged with emoticons.

Figure 1. Sample in the CK+ dateset[22].

The JAFFE dateset[11] (Japanese Female Facial Expression dateset) dateset consists of 213 images, all gray images. The images contained 10 different Japanese women, each with seven expressions, including six basic facial expressions and one neutral facial expression, each with about three to four

DOI: 10.54254/2755-2721/6/20230760

gray images. However, because the number of images in the dateset is too small, it cannot be used for more demanding experiments.

2.2. The dateset collected in the natural world or online

The AffectNet dateset[12] is a dateset of facial expressions with intensity tags. It manually annotates 450, 000 of the more than a million face images and classifies them into eight expression categories, making it the largest dateset to date with emotion intensity and emotion category tags.

Figure 2. Partial sample of the AffectNet dateset[23].

The FER2013 dateset[13] (Facial Expression Recognition dateset) is a facial expression dateset published by the Kaggle website in the 2013 Facial Expression Contest. The dateset consists of 35,887 images, all of which are grayscale images, containing seven facial expressions that are numerically numbered in order: 0 is angry, 1 is disgust, 2 is fear, 3 is happy, 4 is sorrow, 5 is surprise, 6 is neutral. The FER+ dateset is an extension of the original FER dateset in which images have been relabeled to one of eight emotion types: neutral, happy, surprised, sad, angry, disgusted, fearful, and contemptuous.

Table 1. Commonly used facial expression recognition dateset (The dateset in the Fig.1 are sorted according to the number of samples contained).

dateset	Sample Numbers	Sample Forms	Collection Source	Expression Classification
Multi-PIE	750 000	Image	laboratory	6
AffectNet	450 000	Image	natural environment	8
FER2013	35 887	Image	natural environment	7
RAF-DB	29 672	Image	natural environment	7
IJB-A	24327	Image or Video	laboratory	N/A
AR	4000	Image	laboratory	4
Oulu-CASIA	2880	Image Sequence	laboratory	7
SFEW	1766	Image	natural environment	7
CK+	593	Image Sequence	laboratory	8
JAFFE	213	Image	laboratory	7

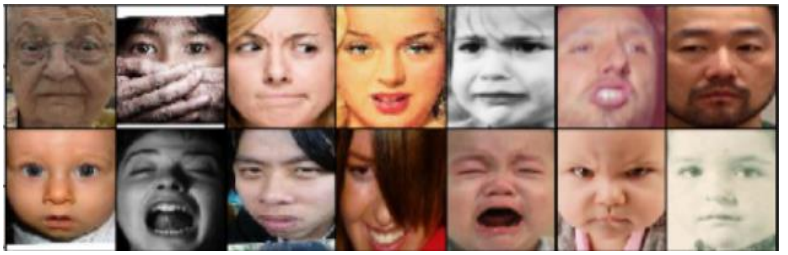

Figure 3. Partial samples in the FER2013 dateset[23].

DOI: 10.54254/2755-2721/6/20230760

Figure 4. Partial samples in the FER+ dateset[23].

The RAF-DB dateset[5] (Real world Affective Faces dateset) is the first dateset that takes photos in the wild, which contains 29672 face images of different ages, genders and skin colors, which is very consistent with the characteristics of expression images in Real life. In addition, the dateset contains 6 basic emotion tags and 12 composite emotion tags, and each image has its own tag to annotate it.

Figure 5. Partial samples in the RAF-DB dateset[22].

The SFEW dateset [14] (Static Facial Expressions in the Wild dateset) is a dateset for Facial expression recognition. The most commonly used version is SFEW 2.0. SFEW 2.0 is divided into three groups: Train (958 samples), Val (436 samples), and Test (372 samples). Each image was classified into one of seven expression categories: anger, disgust, fear, neutral, happiness, sadness and surprise. The labels of the training and validation sets are public, but the labels of the test sets remain private.

3. Methods

In this section, we summarize the research methods in the field of facial expression recognition in the past two years according to the years of publication.

3.1. Lightweight convolutional neural network architecture

The first paper[15] is written by Muhammad Naveed Riaz et al. In this paper, a convolutional neural network architecture called Expression Net (eXnet) is proposed to solve the problems of insufficient computing power and large memory footprint of convolutional neural networks in the field of face recognition. The method is mainly divided into three parts. The first part is the preliminary extraction of features, which includes two convolutional layers, and each convolutional layer is followed by batch normalization and rectifier unit (ReLU). The second part is the further extraction of features. Based on [16], two identical ParaFeat modules are set up for parallel operation. The final step is to use two fully connected global average pools (gaps) for feature classification. In addition, regularization is used to stabilize the training process.

In the second paper[17], Darshan Gera et al. proposed a network architecture called Imponderous Net. The network architecture can achieve excellent facial expression recognition performance on the basis of lightweight. The network architecture consists of three main modules: transfer learning, attention mechanisms, and local and global contexts. The first module is characterized by using the output features of the pool layer after the max-feature-map in Light CNN29 [18] as facial features. The second module uses a lightweight attention mechanism called Efficient Channel Attention (ECA) [19] to improve the recognition performance while reducing the complexity of the overall architecture. The last module mainly uses ECA in four local patches to avoid the loss of information in the occluded area.

The paper[20] is written by Fabio Valerio Massoli et al. The author studies the performance degradation of untrained CNN in recognizing low-resolution images caused by shooting problems. In this paper, a Multi-resolution Approach to Facial Expression Recognition (MAFER) is proposed. This method uses two random extraction methods to simulate the multi-resolution datesets. One is to determine whether the image should be down-sampled, and the other is to specify a specific resolution. Then, in the training of convolutional neural network, the author adopts two methods. The first method

DOI: 10.54254/2755-2721/6/20230760

is to use the method in [21] for training. The second is an improved lightweight training method based on [21], which sets an upper limit for increasing the probability of image down-sampling.

The paper[22] was written by Ping Liu et al. In this paper, the author proposes a simple and effective method called Point Adversarial Self Mining (PASM), which uses the method of mining knowledge from training samples to generate training data of CNN, thereby improving the performance and versatility of convolutional neural networks. The main steps are as follows: firstly, a medium-scale convolutional neural network is trained with the given sample, and then the contribution of different regions in the sample to the classification model is analyzed. Then, the most sensitive position in the sample is found by the point adversarial method and the patch centered on it is screened. In addition, this paper also uses the idea of iteration, PASM can strengthen the universality and performance of the network after each iteration.

3.2. Convolutional neural network for interference factors in field environment

In the paper [23], the author Thanh - Hung, vo Et al. proposes a method called Pyramid with Super-resolution

(PSR) network structure. This method can be used to identify images acquired from the field under the influence of attitude, orientation, etc. This method is mainly divided into six modules, namely, Spatial Transformation Network (STN), scaling, low-level feature simulator, high-level feature simulator, fully connected and finally connected module. Subsequently, this paper proposed a Priori Distributed label Smoothing (PDLS) function, which can help the model learn how to increase the score of correct labels while reducing the score of wrong labels in the back propagation process. And this function can also be applied to additional prior knowledge about the confusion of each expression in the field of facial expression recognition. This function can be divided into two parts: one is the one-heat distribution, the other is the prior distribution, and the FER+ dateset is used for training.

The article[24] was written by Amir Hossein Farzaneh et al. This paper proposed a method, which is the high-level overview of Deep Attentive Center Loss (DACL) method. It consists of attention network and coefficient center loss. The main process of this method is as follows: firstly, inputting the image , features will be generated by convolutional neural network, and the final D-dimensional depth feature vectors of soft-max loss and sparse center loss are extracted from the feature pool layer. It should be noted that the soft maximum loss and sparse center loss are linear combinations. Secondly, the convolutional features are fed into the attention network to estimate the attention weight.

The article[25] was written by Xiaoliang Zhu et al. In this paper, the authors propose a cascade neural network that can recognize facial expressions in dynamic video sequences in real environments. The network consists of three modules: spatial feature extraction module, temporal feature module and mixed attention module. In the module of spatial feature extraction, the author firstly uses the Residual Network (ResNet) to extract spatial features, and then transmits the extracted features to the mixed attention module. In the hybridism module, self-attention, spatial attention and specific feature fusion methods are used to learn attention weight. These data are weighted to spatial features to form mixed attention features. The time feature extraction module uses the Gated Cyclic Unit (GRU) to extract time features and uses weighted mixed attention feature vectors. This module is connected to other modules through the full connection layer, and then connected to the SoftMax layer to output the classification results.

The article[26] was written by Hanting Li et al. In this paper, a submerged vision converter based on pure transformation is proposed. It can be divided into two key modules. The first is mask generation network, which is proposed based on the traditional GAN (Generative Adversarial Network) for generating real-world images. However,in this paper, the MGN (Mask Generation Network) is used to improve the classification accuracy of facial expressions by filtering out the interference of face images, such as background. In training the network, the authors use an additional ViT (Vision Transformer) as a discriminator, and then train the MGN by alternately training the generator and discriminator, and finally obtain a mask that can cover the effective region of facial expression. The second module is

DOI: 10.54254/2755-2721/6/20230760

dynamic remarking, which is enabled to correct the mislabeled images in the dateset so as to improve the quality of the training dateset.

3.3. Convolutional neural network for facial expression uncertainty

The article[27] was written by Kai Wang et al. In paper, a simple and effective network：the Self-Cure Network (SCN) is proposed to solve the problem of uncertainty caused by ambiguous facial expressions in facial expression recognition. The implementation of the network structure consists of three steps: firstly, the features are extracted through CNN network after inputting images, and then the self-attention importance weighting module is used to learn the weight of each image so as to obtain the loss-weighted sample importance. Images with uncertain facial expressions will be assigned low importance weights. Then, the sorting regularization module divides the images into two groups: high importance weight and ground importance weight based on the principle of descending order, and then uses the loss function to regularize the two groups of images. The final noise remarking module will attempt to re-label the sample by the predicted probability.

The paper[28] was written by Amir Hossein Farzaneh et al. In this paper, a new loss function called Discriminant Distribution-Agnostic loss (DDA loss) is proposed. The main process of this method consists of the following steps: After the image is convolutional feature extracted through CNN network, the features will be collected into the embedding space, and the loss function will map the deep features to the label of the expression. The deep features embedded in the space will be learned by two ways: center loss and discriminant distribution agnostic loss. Finally, in the field of facial expression recognition, it can generate highly recognizable depth features, so as to overcome the uncertainty problems encountered in the process of recognition.

4. Experimental results

In this section, we collate and summarize the experimental results of all the above papers, and then compare the accuracy of different papers' methods under the unified dateset. Since the datesets used for testing are different in different papers, we only selected the five datesets that were adopted more than three times in the above papers. They are four in-the-wild datesets: the FER2013 dateset, the RAF-DB dateset, the FER+ dateset and the AffectNet dateset. It should be noted that the AffectNet dateset used to test the performance of the paper method is classified by seven expressions. And a dateset collected in a laboratory setting: the CK+ dateset. And if there is no special explanation, the accuracy data in tables are the maximum performance that the paper method can achieve.

Table 2. Comparison of the accuracy of the method in [15], [20],[22]on the FER2013 dateset.([20]^ represents the fine-tuned paper method on the Aff-wild2 dateset [29]).

Authors	Accuracy
Muhammad Naveed Riaz et al. [15]	73.54
Fabio et al. [20]^	73.45
Ping Liu et al. [22]	73.59

DOI: 10.54254/2755-2721/6/20230760

Table 3. Comparison of the accuracy of the method in [15], [20], [22], [23], [24], [25], [27] and [28] on the RAF-DB dateset.

Authors	Accuracy
Muhammad Naveed Riaz et al. [15]	86.37
Fabio et al. [20]	88.43
Ping Liu et al. [22]	88.68
THANH-HUNG et al . [23]	88.98
Amir Hossein Farzaneh et al. [24]	87.78
Xiaoliang Zhu et al.[25]	88.62
Kai Wang et al.[27]	88.14
Amir Hossein Farzaneh et al.[28]	86.90

Table 4. Comparison of the accuracy of the method in [17], [23], [25]and [27] on the FER+ dateset.

Authors	Accuracy
Darshan Gerna [17]	88.17
THANH-HUNG et al . [23]	89.75
Xiaoliang Zhu et al. [25]	89.22
Kai Wang et al.[27]	89.35

Table 5. Comparison of the accuracy of the method in [17], [23], [24], [25], [27]and [28] on the AffectNet dateset.

Authors	Accuracy
Darshan Gerna [17]	62.06
THANH-HUNG et al . [23]	63.77
Amir Hossein Farzaneh et al. [24]	65.20
Xiaoliang Zhu et al.[25]	64.57
Kai Wang et al.[27]	60.23
Amir Hossein Farzaneh et al.[28]	62.34

DOI: 10.54254/2755-2721/6/20230760

Table 6. Comparison of the accuracy of the method in [15], [17], [25] on the AffectNet dateset.

Authors	Accuracy
Muhammad Naveed Riaz et al. [15]	96.75
Darshan Gerna [17]	87.09
Xiaoliang Zhu et al.[25]	98.46

The datesets in Tables 1-4 are all the test results of the paper method on the in-the-wild dateset, and Table 5 is the test results of the paper method on the dateset composed of data collected from the laboratory. From the above tables, first of all, we can see that the performance of the paper method on the dateset collected from the laboratory is better than that of the in-the-wild dateset. For example, the accuracy of the method in paper [25] on the CK+ dateset is nearly 35 percent higher than its accuracy on the AffectNet dateset. And we can see that the performance of the proposed method on the RAF-DB dateset and the FER+ dateset is also better than that on the FER2013 dateset and the AffectNet dateset for the same in-the-wild dateset. For example, the accuracy of the method in paper [15] on the RAF-DB dateset is nearly 16 percent higher than that on the FER2013 dateset. Finally, we find that the performance of some paper methods on in-the-wild datesets even exceeds their accuracy on datesets collected under restricted conditions. For example, the accuracy of the methods in paper [17] on the RAF-DB dateset is slightly higher than the performance of these two methods on the CK+ dateset.

5. Conclusions

In this paper, we summarize the new methods proposed in the field of face recognition in the past two years. And through the comparison of their experimental data, the performance of the paper's method and the degree of difficulty in identifying each dateset are compared.The above papers are basically the research on unimodal data, so in the future research, we can consider the combination of multi-modal information, which is the combination of language, text, posture and other modal information, to more comprehensive and accurate facial expression recognition.

References

[1] Chowdary M K, Nguyen T N, Hemanth D J. Deep learning-based facial emotion recognition for human–computer interaction applications[J]. Neural Computing and Applications, 2021: 1-18.

[2] Kim J H, Poulose A, Han D S. The extensive usage of the facial image threshing machine for facial emotion recognition performance[J]. Sensors, 2021, 21(6): 2026.

[3] Pham L, Vu TH, Tran T A. Facial Expression Recognition Using Residual Masking Network [C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021:4513-4519.

[4] 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]//2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010: 94-101.

[5] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[6] Jayabharathi P, Suresh A. convolutionalal Neural Network model in Different dateset for Pose Face Recognition[C]//2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE, 2022: 1-7.

[7] Liu Y, Chen J. Unsupervised face frontalization for pose-invariant face recognition[J]. Image and Vision Computing, 2021, 106: 104093.

DOI: 10.54254/2755-2721/6/20230760

[8] Mahmoud S, Saif O, Nabil E, et al. AR-Sanad 280K: A Novel 280K Artificial Sanads dateset for Hadith Narrator Disambiguation[J]. Information, 2022, 13(2): 55.
[9] Chen C, Winkler S. Generating near-infrared facial expression datesets with dimensional affect labels[J]. arXiv preprint arXiv:2206.13887, 2022.
[10] Gill R, Singh J. A Deep Learning Model for Human Emotion Recognition on Small dateset[C]//2022 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2022: 1-5.
[11] Ni J. Performing seven expression class classification on SFEW dateset by Casper Algorithm[J].
[12] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31.
[13] Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep learning approaches for facial emotion recognition: A case study on FER-2013[M]//Advances in hybridization of intelligent methods. Springer, Cham, 2018: 1-16.
[14] Riaz M N, Shen Y, Sohail M, et al. Exnet: An efficient approach for emotion recognition in the wild[J]. Sensors, 2020, 20(4): 1087.
[15] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutionals[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
[16] Gera D, Balasubramanian S. Imponderous Net for Facial Expression Recognition in the Wild[J]. arXiv preprint arXiv:2103.15136, 2021.
[17] Wu X, He R, Sun Z, et al. A light CNN for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11): 2884-2896.
[18] Wang Q, Wu B, Zhu P, et al. Supplementary material for ‘ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA. 2020: 13-19.
[19] Massoli F V, Cafarelli D, Gennaro C, et al. MAFER: A Multi-Resolution Approach to Facial Expression Recognition[J]. arXiv preprint arXiv:2105.02481, 2021.
[20] Massoli F V, Amato G, Falchi F. Cross-resolution learning for face recognition[J]. Image and Vision Computing, 2020, 99: 103927.
[21] Liu P, Lin Y, Meng Z, et al. Point adversarial self mining: A simple method for facial expression recognition in the wild[J]. arXiv preprint arXiv:2008.11401, 2020.
[22] Vo T H, Lee G S, Yang H J, et al. Pyramid with super resolution for in-the-wild facial expression recognition[J]. IEEE Access, 2020, 8: 131988-132001.
[23] Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 2402-2411.
[24] Zhu X, Ye S, Zhao L, et al. Hybrid attention cascade network for facial expression recognition[J]. Sensors, 2021, 21(6): 2003.
[25] Li H, Sui M, Zhao F, et al. MVT: mask vision transformer for facial expression recognition in the wild[J]. arXiv preprint arXiv:2106.04520, 2021.
[26] Jiang Y, Chang S, Wang Z. Transgan: Two pure transformers can make one strong gan, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.
[27] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6897-6906.
[28] Farzaneh A H, Qi X. Discriminant distribution-agnostic loss for facial expression recognition in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 406-407.
[29] Kollias D, Zafeiriou S. Aff-wild2: Extending the aff-wild database for affect recognition[J]. arXiv preprint arXiv:1811.07770, 2018.