A Review of Deep Learning-Based Micro-Expression Classification

Yifei Zhang

School of Future and Technologies, Beijing Institute of Technology, Beijing, China

* Corresponding Author Email: 1120220566@bit.edu.cn

Abstract. In daily life, it is important to discern the emotions of others. People always rely on facial expressions and body language to discern the emotional tendencies of others, so that they can choose appropriate ways to communicate and interact with others. Macro expression has a large range of action, while micro expression has a small range of action and is not easy to detect. The former can be falsified while the latter are generated spontaneously. Therefore, analyzing micro-expressions can more accurately determine a person's true emotions, which can then assist criminal investigation or provide psychological therapy. However, even trained individuals find it challenging to correctly identify micro-expressions because of their fleeting and nuanced characteristics, hence, machine learning techniques are frequently employed to handle such tasks. This article looks into deep learning and conventional machine learning algorithms for micro-expression classification. Sentiment analysis methods based on machine learning usually detect facial changes based on texture features and optical flow features. Deep learning methods can be used to complete end-to-end classification through modules such as CNN, GCN and transfer learning. Finally, it was found that to further improve classification accuracy, enhancing the quality of datasets and attempting to apply self-supervised methods to train deep models could be beneficial.

Keywords: micro-expression; machine learning; deep learning; CNN; GCN.

1. Introduction

The classification of facial expressions is one of the important ways of the psychological analysis of characters, which can be divided into macro expressions and micro expressions according to the size of facial movements. Macro-expressions involve large movements and are non-spontaneous, meaning they can be faked. Micro-expressions have small movement ranges and extremely short duration. They are spontaneous and unconscious emotional disclosures. Therefore, a person's potential emotions can be detected by analyzing micro-expressions, which has wide applications in fields such as case investigation and psychotherapy.

Micro-expression recognition can be studied from two aspects: expression detection and expression classification. The core purpose of the former is to answer whether there are micro-expressions in the image sequence, while the latter is to determine the category of micro-expressions under the assumption that there are micro-expressions in the sequence. In addition, the detection task also involving advanced tasks, such as determining the occurrence timing (Onset), peak timing (Apex), and disappearance timing (Offset) of micro-expressions. From a machine learning perspective, both tasks are classification tasks, so some researchers do use the same logic to train different models that solve the two tasks.

The characteristics of micro-expressions, with small magnitude of movement and short duration, pose many challenges for manual detection and classification: The time span of micro-expressions is typically less than one second, which results in the camera does not capture the micro-expression apex frame, thus allowing people ignore the existence of micro-expressions. And the small range of movements means that the accuracy of trained person in classifying micro-expressions is only 47%. Also, it is difficult to distinguish between macro-expressions and micro-expressions, thus observers might be deceived by the expressions fabricated by another person.

In order to solve various problems of manual recognition, traditional machine learning methods based on texture and optical flow methods of images try to extract features for classifying micro-expressions. Zhao's [1] research team at the University of Oulu in Finland uses LBP (Local Binary

Pattern) and its extended algorithm LBP-TOP to capture facial texture information. Wang's [2] team at the Chinese Academy of Sciences used the MDMO (Main Directional Mean Optical flow) method to reveal the underlying manifold structure.

With the development of deep learning, image and video processing techniques based on deep learning have made tremendous progress. In particular, feature extraction methods based on CNNs (convolutional neural networks) can unearth deep semantic information from images. Therefore, some researchers use CNNs for micro-expression classification. Combined with the attention mechanism, some scholars have proposed GCN-CNN processing methods; some researchers have also applied transfer learning to adapt well-established macro-expression classification models to address micro-expression tasks.

This article will start from introducing commonly used micro-expression datasets, review traditional machine learning and deep learning micro-expression recognition methods, and compare the advantages and disadvantages of each method to boost the future development of micro-expression classification tasks.

2. Micro-expression Datasets

In micro-expression classification tasks, the evaluation metrics typically used are consistent with those used in macro-expression classification tasks: accuracy, precision, and F1 score. In order to train a model with high performance, datasets are crucial. Due to micro-expressions' spontaneous and subtle nature, their collection process is difficult. Currently, widely used micro-expression datasets include: Polikovsky [3] established by the University of Tsukuba University in Japan, USF-HD [4] established by the University of South Florida in the United States, CASME [5] (the Chinese Academy of Sciences micro-expression) established by the Chinese Academy of Sciences, SAMM [6] established by Manchester Metropolitan University in the United Kingdom, Shandong MMEW established by the university and SMIC (Spontaneous micro-expression corpus) established by Oulu University in Finland. The SMIC dataset contains three sub-datasets HS, VIS and NIR collected by different recording devices. Among them, the datasets from Tsukuba University and the University of South Florida record the micro-expressions imitated by the subjects after watching the video. Therefore, strictly speaking, the expressions in these two datasets are non-spontaneous, which differs from spontaneous micro-expressions. For micro-expression datasets, the induction method, frame rate, and labeling method are particularly important.

2.1. Induction Method

Since the core value of the micro-expression classification task is to judge a person's emotions through his or her spontaneous and subtle facial configuration changes, the induction method of a dataset is highly important.

In the SMIC dataset, the team from the University of Oulu showed subjects videos that could induce emotional fluctuations and expected them to suppress their emotions. If a subject's emotions were guessed by the researchers, as a punishment, the subject needed to fill out a lengthy questionnaire. This punitive mechanism ensures the spontaneity of the subject's micro-expressions, making the dataset highly reliable.

A similar reward and punishment mechanisms is also reflected in the CASME [5] dataset. CASME [5] consists of 8 types of micro-expressions and 195 samples, during the collection process, the researchers also played videos to the subjects. If the subjects can suppress their emotions without being noticed, they can receive cash rewards. SAMM and MMEW also have similar mechanisms to ensure spontaneity. However, in contrast, some datasets lack spontaneity guarantees, such as Polikovsky [3]. Using them as research data may affect the performance of the model in practical applications.

2.2. Frame Rate

Micro-expressions typically last between twenty-fifth of a second and one-third of a second. Therefore, ordinary cameras may miss the apex of micro-expression, so the shooting frame rate of the dataset should not be too low. But on the other hand, a frame rate that is too high may cause the shutter opening time to be too short to allow enough amount of light to enter, making the image darker and difficult to identify.

In order of increasing frame rate, the datasets can be arranged as follow: CASME [5] (60 frames per second), SMIC (100 frames per second), Polikovsky [3] and CASME II (200 frames per second). But in fact, datasets with frame rates close to those of everyday cameras also hold significance, such as USF-HD [4]. This is because the frame rate of the micro-expression video that needs to be recognized in actual applications is not as high as the above-mentioned datasets. Therefore, datasets captured at 25 frames per second can to some extent reflect the practical performance of models.

2.3. Data Labeling Method

The labeling of micro-expression datasets can be divided into types: natural emotion description and FACS (Facial action coding system). Natural emotion description uses everyday words to mark the subject's positive or negative emotions, such as happiness, sadness, surprise, angry, disgust, fear, etc. FACS encodes local facial movements and is widely used in the field of expression recognition. In FACS, each code corresponds to an AU (Action Unit). For example, AU1 represents inner brow raiser and AU10 represents upper lip raiser. Using the FACS method is to label the code of the action unit that the subject appears on the image sequence, so a micro-expression can be regarded as a combination of some AUs.

3. Micro-expression Classification

The micro-expression classification task refers to classifying the categories of micro-expressions from a sequence of images in which micro-expressions are assumed to exist, such as joy, sadness, surprise, etc. From the perspective of machine learning, this is a classification task. According to different feature extraction methods, micro-expression classification methods include two main types: traditional machine learning-based and deep learning-based approaches.

Among these, traditional methods can be divided into two categories: texture-based image features and optical flow features. The former pays more attention to the spatial characteristics of the image, while the latter, due to the characteristics of optical flow, focuses more on the time dimension and examines the characteristics of pixels changing over time.

Methods based on deep learning can be divided into three categories: CNN method based on classic image classification tasks; GCN-CNN method based on local-global perspective, and based on transfer learning. The first type starts from the basic image classification task, using convolutional neural network to extract features make predictions as a overall network. The second type, building upon the first type, considers the local feature AU, uses GCN to extract AU features, and combines the extracted features of GCN and CNN. The third type takes the similarity between macro-expressions and micro-expressions into account and attempts to use transfer learning based on macro-expression classification models.

3.1. Classification Method Based on Traditional Machine Learning

3.1.1 Texture-based classification method

LBP, as a texture analysis method, has been widely used in computer graphics. Its main idea is to extract texture features of images based on comparison of gray values in the neighborhood. In order not to ignore the timing information of micro-expressions, LBP can be promoted to LBP-TOP (Local Binary Pattern from Three Orthogonal Planes), apply the LBP method on the three 2D planes X-Y, X-T, and Y-T, and concatenate the extracted features in sequence. PFISTER et al. [7] first proposed this method, but due to the need to calculate LBP features on three planes, the feature dimensions

extracted using the LBP-TOP method are tremendous. To address this issue, subsequent methods such as STCLQP and LBP-SIP were proposed.

Piao et al. [8] extract texture features of images based on statistical methods. Piao regards expressions as random variables, uses the SSA (Statistical Shape Analysis) method to study the distribution of random variables in high-dimensional space, and finally uses the Naive Bayes method to predict the expressions in the test set. However, before using the SSA method, manual annotation of facial expression landmarks is required. This process is time-consuming and labor-intensive, and the limitations of human expression recognition restrict the upper limit of classification accuracy.

3.1.2 Optical flow features based classification method.

Optical flow, like texture, is also a commonly used concept in computer graphics, which represents the instantaneous speed of pixels in an image. In order to further extract the temporal characteristics of micro-expression sequences, Xu et al. [9] used the optical flow field as the core method to describe the motion characteristics of micro-expressions and proposed the FDM (Facial dynamics map) . First, extract the optical flow field between adjacent frames of the image sequence and remove errors caused by potential image translations. Then, further divide the optical flow field into space-time cubes, and use iterative methods to determine the main direction of pixel displacement, thereby obtaining the facial movement characteristics of micro-expression. However, the time complexity of extracting the space-time cubes of the optical flow field is high, making the method unable to predict micro-expressions in real time.

To improve computational efficiency, Tu et al. [10] proposed an improved algorithm based on PCA. After compressing the feature dimensions, utilize SVM as the classifier. This approach significantly reduces the running time of the algorithm while barely reducing classification accuracy.

3.2. Classification Method Based on Deep Learning

3.2.1 CNN method derived from classic image classification tasks

From the mentioned traditional methods, it can be found that the method of extracting micro-expression sequence features and the method of classification prediction lack integrity and consistency, which results in affecting the accuracy of classification to a certain extent.

If you can get key frames of the image sequence, such as apex frames. Then the method for handling image classification task can be directly applied to the micro-expression classification task, thereby extracting features and predicting expression categories in one neural network, ensuring the integrity and consistency of the two processes. In fact, the micro-expression recognition task can provide possible key frames.

Li et al. [11] used the VGG-16 framework and the LBP method to extract image features at the same time, and before the fully connected layer, using preset parameters as weights, the features extracted by LBP and CNN were merged together to predict the class of images. This method treats the features of each part of the face equally. However, the contribution of most areas to the micro-expression category is not significant, and the facial movements of the apex frames are not amplified, which may affect the prediction accuracy of the model on a larger test set. Since it involves the adjustment of weight hyperparameters, repeated attempts are required, which also results in low training efficiency.

In order to deal with these problems, Li et al. [12] also based on VGG-16, first used the Euler motion amplification algorithm to amplify the subtle movements of the face. Then the global features of the face and the local features of the area of interest are extracted respectively, and finally the features are fused to classify expressions. Experimental results show that based on the global-local feature extraction method, the classification method that only processes apex frames is not inferior to the method that processes the complete image sequence.

3.2.2 GCN-CNN method based on Local-Global perspective.

Similar to the work of Li et al. [12], to encapsulate and emphasize the detailed features of micro-expressions, many classification methods start from a local-global perspective, extract key local

features of the face, and fuse them with the overall features of the whole face as the core basis for classification. Yao et al. [13] used CNN to extract global features and used SVM to find local features in the three areas of eyes, eyebrows, and mouth. After fusing the two, they used MLP to predict image categories. This method is not reasonable enough for local feature area selection and insufficient feature mining.

In order to improve this situation, the concept of AU can be utilized. Each AU can be regarded as a node on the graph, and the simultaneous occurrence of two AUs is regarded as an edge between corresponding nodes on the graph. By using GCN as the method for extracting local features, traditional methods such as LBP and SVM can be replaced to improve classification accuracy. Lo et al. [14] proposed MER-GCN, using 3D convolution to extract sequential features of image sequences, and establishing a graph model in the above way. After using GCN and CNN to extract features separately, the important areas of the face are highlighted through dot product operation, thus the fusion features are obtained. However, 3D convolution causes the model to have a large quantity of parameters and require a long time to train.

Zhao et al. [15] located 68 key points of the face, used a weighted attention mechanism for AU to extract local features, extracted dependencies between nodes through GCN, and finally assembled global features for micro-expression recognition. Wang et al. [16] defined 16 Regions of Interest and extracted the features of each region. However, because the corresponding relationship of AU is manually defined, the accuracy of the model is limited by human domain knowledge.

3.2.3 Transfer learning method

Actually, existing datasets are still insufficient because it is challenging to gather and identify micro-expression data accurately. This proposes challenges in model training, making it more likely to result in insufficient generalization or overfitting. It is easy to notice that macro-expressions and micro-expressions are quite similar from the perspective of facial muscle movement. Therefore, an attempt has been made to address this issue by utilizing transfer learning methods, which means starting from macro-expression classification models.

Wang et al. [17] first trained the model on 4 macro-expression datasets, and then used the micro-expression dataset to fine-tune the parameters. The training process on the two datasets is separated and does not reflect the relationship between them well. Kang et al. [18] built two CNNs with shared parameters based on the attention mechanism. The correlation loss measured by the log-Euclid distance is combined with the classification loss to balance the two to obtain better prediction results.

4. Summary

This article first introduces widely used datasets in the area of classification of micro-expressions and points out that the induction method of a dataset is particularly important, which divides datasets into two categories: spontaneous and non-spontaneous. In addition, the frame rate and annotation method of a dataset should also be considered. Then, this article introduces popular approaches to micro-expression classification tasks in deep learning and classical machine learning. Traditional methods mostly use basic concepts of computer graphics to extract features, while deep learning methods are highly related to CNN.

Current research shows that the model accuracy obtained by deep learning methods is higher. In order to further develop deep learning methods, future research can.

(1) Establishing larger datasets for training, thereby boosting the generalization ability of the model.

(2) Exploring different training methods, such as self-supervised methods. Current models are trained using supervised methods, which rely heavily on data volume. Introducing unsupervised methods could reduce this dependence.

(3) Building multimodal models for micro-expression classification or emotion recognition, taking body language into consideration to improve the model prediction accuracy.

References

[1] MOILANEN A, ZHAO GY, PIETIKÄINEN M. Spotting rapid facial movements from videos using appearance-based feature difference analysis. Washington D. C. , USA: IEEE Press, 2014, 1722-1727.

[2] LIU Y, et al., A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. IEEE Transactions on Affective Computing, 2016, 7(4): 299-310.

[3] POLIKOVSKY, S., Y. KAMEDA and Y. OHTA, Facial Micro-Expression Detection in Hi-Speed Video Based on Facial Action Coding System (FACS). IEICE Transactions on Information and Systems, 2013, E96.D(1): 81-92.

[4] SHREVE, M., et al. Macro- and micro-expression spotting in long videos using spatio-temporal strain. IEEE, 2011.

[5] YAN W J, LI X, WANG S J, et al. CASMEII: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One, 2014, 9(1): 86041.

[6] Davison, A.K., et al., SAMM: A Spontaneous Micro-Facial Movement Dataset. IEEE Transactions on Affective Computing, 2018, 9(1): 116-129.

[7] PFISTER T, LI X B, ZHAO G Y, et al. Recognising spontaneous facial micro-expressions. Washington D. C. USA: IEEE Press, 2011, 1449-1456.

[8] PIAO LZ, et al., Facial Expression Recognition Using Statistical Shape Analysis. Journal of Ningbo University, 2009, 22(02): 196-201.

[9] XU, F., J. Zhang and J.Z. Wang, Microexpression Identification and Categorization Using a Facial Dynamics Map. IEEE Transactions on Affective Computing, 2017, 8(2): 254-267.

[10] TU L and LIU BY, PCA-based Modification of Dynamics Map Feature Extraction in Micro-Expression Recognition. Communications Technology, 2019, 52(02): 337-342.

[11] LI Y, et al., Progress and challenges in facial action unit detection. Journal of Image and Graphics, 2020, 25(11): 2293-2305.

[12] LI, Y., X. HUANG and G. Zhao, Joint Local and Global Information Learning with Single Apex Frame Detection for Micro-Expression Recognition. IEEE Transactions on Image Processing, 2021, 30: 249-263.

[13] YAO LS, XU GM, et al., Facial Expression Recognition Based on Local Feature Fusion of Convolutional Neural Network. Laser & Optoelectronics Progress, 2020, 57(04): 338-345.

[14] LO, L., et al. MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks. IEEE, 2020.

[15] ZHAO JH and YANG QX, Micro-expression Recognition Based on AU-GCN and Attention Mechanism. Computer and Modernization, 2023(03): 48-53.

[16] WANG, S., et al., Micro-Expression Recognition Using Color Spaces. IEEE Transactions on Image Processing, 2015, 24(12): 6034-6047.

[17] WANG C Y. Micro-attention for micro-expression recognition. Neurocomputing, 2020, 410: 354-362.

[18] KANG J, LI SY, Transfer Learning Method for Facial Expression Recognition Base on Attention Mechanism. Computer Engineering and Design, 2021, 42(03): 797-804.