Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241545

The investigation of traditional models and machine learning models in dynamic facial expression recognition

Xiyu Wu

The Department of Mathematics and Statistics, South-Central Minzu University, 430074, China

202221101052@mail.scuec.edu.cn

Abstract. In everyday life, dynamic facial expressions are merely continuous human responses to external events. However, in human-computer interaction, rapidly recognizing changes in facial expressions from video streams is a relatively complex process. This complexity renders Dynamic Facial Expression Recognition (DFER) a critical research task in the domains of computer vision and image processing. This paper analyses the correlations and contrasts between static and dynamic facial expression research, highlighting key issues in the study of dynamic facial expressions, such as dynamic feature extraction and frame extraction. After that, it enumerates significant algorithms in both traditional models and deep learning models, providing an analysis of the advantages and disadvantages of these two major approaches. At the same time, it investigates the reasons behind the transition of research models for DFER from traditional methods to deep learning approaches. The paper focuses on two notable models from each approach: Histogram of Oriented Gradient (HOG) for processing raw images, Support Vector Machine (SVM) for data classification in traditional models. Convolutional Neural Network (CNN) for spatial feature extraction and Long Short-Term Memory (LSTM) for temporal feature extraction in deep learning models. These models are discussed in detail concerning their strengths and weaknesses, operational processes, and performance outcomes. In the concluding section, the author summarizes the main factors influencing research in this field and the current challenges encountered. By focusing on future research directions, the paper also presents a review of recent methodologies and offers insightful research directions for further investigation.

Keywords: Traditional model, machine learning, dynamic facial expression recognition.

1. Introduction

Facial expression is one of the important forms of non-verbal communication in interpersonal communication. It is of great significance to establish good interpersonal relationships and promote effective communication. Since DFER has practical importance in public safety, human-robot interaction, psychological health monitoring and other fields, it has recently received increased attention.

In the 20th century, Ekman and Friesen identified six primary emotions that were disgust, anger, fear, sadness, happiness, and surprise. In later studies, contempt was included in the list of the primary emotions. According to the feature representations, Facial Expression Recognition systems can be divided into two main classes: dynamic sequence FER and static image FER. In the study of dynamic

DOI: 10.54254/2755-2721/87/20241545

fields [1], they considered the time static characteristics of dynamic texture and temporal texture, and their local information and spatial position.

There are some algorithms in traditional models used for FER such as geometric features, template matching, Eigenfaces features algorithms, Hidden Marcov method, singular value decomposition method. While recognition of geometric features is sensitive to the positions of feature points, template matching is sensitive to head posture and scale changes. Additionally, the Eigenfaces method is sensitive to changes in lighting and micro-expressions, while singular value decomposition is sensitive to noise and outliers in the data. All of them will affect the accuracy of features extraction. Meanwhile, these algorithms cost a lot when processing high -dimensional data, which is short of the performance requirements in actual use.

With the new forms of information technology, in order to help computer systems recognize and solve different problems from the surrounding environment, scholars gradually devote themselves to the study of Machine Learning [2] and Deep Learning [3]. The effectiveness of machine learning depends on the integrity of input data, and the feature extraction of facial expressions will be affected by lighting, posture, and obstruction etc. They might lead to prejudice feature selection that may lead to incorrect discrimination between classes. Therefore, using traditional methods including machine learning usually requires manual features extraction, which greatly reduces efficiency. For example, Local Binary Patterns (LBP), LBP on three orthogonal planes (LBP-TOP), Histogram of Oriented Gradient (HOG), Scale-invariant Feature Transform (SIFT), Support Vector Machines (SVM). Relatively, features extraction is automatically implemented in Deep Learning. Convolutional Neural Network (CNN) is an extremely popular approach, which can automatically detect the most distinctive features without any manual supervision. It can also reduce the number of training network parameters to a certain extent, help the network enhance the generalization and avoid overfitting. Artificial intelligence uses this similar layered architecture to simulate the process of the core sensory region of the human brain.

Benefit from the rapid development of DFER, the objective of this paper is to conduct an exhaustive review of the research on DFER through both traditional and machine learning models. At the same time, it introduces the latest research progress. Finally, the challenges that must be addressed to make DFER research applicable to real-world situations are discussed.

2. Method

2.1. Traditional model-based DFER

2.1.1. Histogram of oriented gradient (HOG)-based model

Local objects are characterised by their form and appearance through the distribution of edge directions or local intensity gradients using HOG, a shape descriptor. The main purpose of this program is to detect objects, but it can also be used to intuitively model the shape of facial muscles through edge analysis. Moreover, the input data will not be a factor affecting the parameter configuration of HOG. However, if it is desired to achieve higher detection performance, it is required to adjust the HOG parameters to a fine scale inverse, more direction boxes, and medium-sized, strongly normalized, overlapping descriptor blocks. Existing study has shown that the highest FER performance can be achieved if the parameters are configured to a unit size of 7 pixels and 7 direction boxes [4]. HOG can be subdivided into rectangular HOG (R-HOG) and circular HOG (C-HOG) according to the geometric shape of the descriptor block. In DFER, the frame image is pre-processed using the HOG method, which typically involves converting the image to grayscale. The Sobel operator is then applied to determine the horizontal and vertical gradients. Subsequently, the image is split up into a number of descriptor blocks. In every block, the histogram of gradient directions is computed and the frequency of each gradient direction is documented. To generate the feature vector, the normalized gradient histograms are concatenated in the end.

DOI: 10.54254/2755-2721/87/20241545

2.1.2. Support vector machines (SVM)-based model

The capability to be used for classification and regression prediction can be obtained from SVM, a generalized linear classifier. It is suitable for pattern classification and regression-based applications. Due to its strong statistical basis and effectiveness, it is capable of using linear function hypothesis space in high-spatial feature area. By building a hyperplane which is associated with decision planes in higher dimensions, SVM performs classification. This hyper-plane refers as decision planes, which can make a distinguish between two different groups of data. Data in higher-dimensional spaces is categorized by constructing a hyperplane using a suitable non-linear mapping. The investigation of SVM involves examining the support vectors that determine the decision boundary and yield a significant marginal separation between the classes. SVM distinguishes between classes by recognizing different expression types with the maximum marginal distance [5]. In the study of dynamic facial expressions, the feature vector extracted by HOG can be input into the SVM library, and then the SVM became a useful tool to classify the observed facial emotions. SVM has been enhanced in different ways in recent decades, including Lagrangian SVM, twin SVM, Least Square SVM, Quantum SVM and many others improvements.

2.2. Deep Learning model-based DFER

2.2.1. Convolutional Neural Network (CNN)-based model

Using CNN, the characteristic of countenance can be extracted through a feedforward neural network. The biggest difference from conventional feature extraction methods is that it does not require manual feature extraction and can respond to various features automatically. Its overall architecture includes input layer, folding layer, pooling layer and completely connected layer. In the study of DFER, the biggest difficulty is how to extract effective facial features in the video. Thanks to the convolution layer of CNN, it can continuously abstract the original frame image and extract effective features layer by layer. The multiple convolution cores in the convolution layer can ensure that CNN extracts multiple feature descriptions of facial expressions in each frame during the learning process. Recently, a study proposed a method to combine CNN and HOG to extract more comprehensive dynamic facial expression features [6]. If DFER is performed, ordinary CNN can only obtain the spatial relationship of the input data but not the temporal relationship. To overcome this limitation, the concept of 3D convolutional neural network (3DCNN) was proposed. Large deep CNN can use pure supervised learning. It is vital to note that the depth of CNN is very important for the realization of expression recognition. If a single convolutional layer is removed, its network performance will decrease. Especially when used on video sequences, this urgently requires very large and deep convolutional networks [7].

2.2.2. Long Short-Term Memory (LSTM)-based model

LSTM is an impactful tool for sequentially encoding spatiotemporal features. It was created to alleviate the gradient vanishing or exploding problem encountered by traditional recurrent neural networks when dealing with long-term dependency problems. It replaces the hidden layer of the traditional RNN with a composite unit containing input nodes, input gates, internal states, forget gates, and output gates [8]. It is within the realm of possibility for LSTM to bridge minimum delays of over 1000 discrete-time steps without sacrificing short-time delay capabilities by impelling a steadfast error flow through a Constant Error Carousel (CEC) within a particular unit. The processing and prediction of time series data can be effectively handled by using LSTM in the study of DFER, which effectively handles the temporal dependency of facial expression changes. At the same time, since LSTM has the effect of improving the robustness of the model to noise and uncertainty, LSTM can also effectively cope with the challenges brought by illumination changes and facial occlusion to DFER. After that, LSTM can be expanded to Bilateral LSTM (Bi-LSTM) [9]. Prediction can be achieved through the use of both past and future information by Bi-LSTM, but it has higher computational complexity and memory requirements than unilateral LSTM. Recently, a study proposed a method to use 3D-CNN and LSTM to extract the

DOI: 10.54254/2755-2721/87/20241545

provisional relationship between consecutive frames in video sequences, and found that the facial recognition rate was improved to a certain extent [10].

3. Discussion

In daily life, facial expressions are not static but dynamic, which makes video-based facial expression recognition became a mainstream trend. Although traditional models and deep learning models can roughly finish facial expression recognition, they are still facing many limitations and challenges. Conventional models have limited feature representation capabilities and rely on manual feature extraction, which may lead to an inaccurate capture of tiny shifts in facial expressions and longitudinal data present in dynamic facial expression sequences. While spatial feature dimensionality has been reduced via the use of Principal Component Analysis (PCA) [11], their ability to address the complexity of dynamic features is still limited. Therefore, evolving towards more complex models is necessary.

Deep learning models leverage large datasets and complex layer structures to learn rich feature representations, which are more conducive to research in DFER compared to traditional models. However, they also introduce new challenges. For instance, their internal mechanisms are more complex, making the decision-making process difficult to interpret, thereby reducing model interpretability and credibility. While the use of Grad-CAM improves model interpretability and transparency [12], the black-box spontaneous of deep learning models remains a significant obstacle in sensitive areas, which is widely accepted.

There are still many issues worth exploring in the topic of DFER research. For example, the processing time for single-frame data is lengthy, and the efficiency of storing and managing large amounts of video data is low. In the future, the Apache Spark framework may serve as an effective tool for distributed management and processing of large datasets [13]. External factors such as lighting changes, occlusion, and different shooting angles also pose challenges because they can significantly affect feature recognition accuracy. Developing more robust models that generalize well under different conditions is crucial. Furthermore, diversity in facial expressions due to factors such as gender, age, and ethnicity is often lacking in current datasets, restricting the practical application of models in real-world scenarios. Therefore, creating ideal dynamic facial expression datasets that include more attributes like gender, age, and ethnicity is necessary. Currently, only combining Transfer Learning (TL) can alleviate the problem of data imbalance and scarcity [14, 15]. Face video data may also contain sensitive information such as personal identity, behaviour patterns, and emotional states, raising serious ethical and privacy concerns regarding data collection and usage. Ensuring data anonymity and secure storage is crucial to prevent misuse of information and protect personal privacy. Facial expressions are just one component of human expression behaviour in reality. The emergence of social media and numerous digital platforms emphasizes the importance of Multimodal Sentiment Analysis (MSA) methods [16] that analyze human opinions on something from text, audio, images, etc.

Future research should focus on developing more efficient and interpretable deep learning models, improving robustness to external variations, and optimizing frameworks for processing large-scale video data. Additionally, diversifying datasets, integrating multimodal emotional data, and creating more accurate and reliable DFER systems are essential. Safeguarding the privacy of adopted facial data is also crucial while facilitating the smooth progress of this research.

By addressing these challenges, DFER technology can make significant strides in more practical and widespread applications, ultimately advancing fields such as human-computer interaction, psychological health monitoring, and public safety.

4. Conclusion

This work conducts a comprehensive survey on traditional models and machine learning models used for DFER, with a focus on HOG, SVM, CNN, and LSTM algorithms. Even with the advancements made possible by the use of both traditional and deep learning models, a number of obstacles still exist. These include limitations in expressing dynamic facial features, sensitivity to external factors, limited data storage, and inefficient processing. Future research should focus on building datasets that include

DOI: 10.54254/2755-2721/87/20241545

diverse attributes such as gender, age, and race. Developing more robust recognition models to address the effects of occlusion, lighting, and angle variations on DFER is also essential. Ensuring the privacy and security of video data used in research is crucial for protecting personal identity and sensitive information. Utilizing multimodal emotion analysis models, which combine data from text, audio, and images, can enhance the practicality of facial expression recognition systems in real-world applications. Efforts to tackle these challenges can enhance the accuracy, reliability, and applicability of DFER in various domains.

References

[1] Jung H Lee S Yim J Park S & Kim J 2015 Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Santiago: IEEE) pp 2983–2991

[2] Kim S An G H & Kang S -J 2017 Facial expression recognition system using machine learning Proceedings of the International SoC Design Conference (ISOCC) (Seoul: IEEE) pp 266–267

[3] Hinton G E & Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks Science vol 313 (Washington D.C.: American Association for the Advancement of Science) pp 504-507

[4] Carcagnì P Del Coco M Leo M et al. 2015 Facial expression recognition and histograms of oriented gradients: a comprehensive study SpringerPlus vol 4 (Berlin: Springer) p 645

[5] Chandra M A & Bedi S S 2021 Survey on SVM and their application in image classification International Journal of Information Technology vol 13 (Mumbai: Bharati Vidyapeeth) p 1-11

[6] Pan X 2020 Fusing HOG and convolutional neural network spatial–temporal features for video-based facial expression recognition IET Image Processing vol 14 (London: Institution of Engineering and Technology) p 176-182

[7] Mao L Chen S & Yang D 2021 Guided convolutional neural network video pedestrian action classification improvement method Journal of Wuhan University (Information Science Edition) vol 46 (Wuhan: Wuhan University Press) p 1241-1246

[8] Bai M & Goecke R 2020 Investigating LSTM for micro-expression recognition Companion Publication of the International Conference on Multimodal Interaction pp 7-11

[9] Siami-Namini S Tavakoli N & Namin A S 2019 The Performance of LSTM and BiLSTM in Forecasting Time Series Proceedings of the IEEE International Conference on Big Data (Big Data) (Los Angeles: IEEE) pp 3285-3292

[10] Hasani B & Mahoor M H 2017 Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (Honolulu: IEEE) pp 2278-2288

[11] Kang J Lin X & Wu X 2015 Face recognition algorithm based on Laplace pyramid dimension reduction Journal of Shaanxi University of Science and Technology (Natural Science Edition) vol 35 (Xi'an: Shaanxi University of Science and Technology) p 165-168174

[12] Selvaraju R R et al. 2017 Grad-cam: Visual explanations from deep networks via gradient-based localization Proceedings of the IEEE international conference on computer vision (Venice: IEEE)

[13] Li H Tan Y & Wu F 2017 Massive video face extraction and recognition parallel framework design and optimization Computer Applied Research vol 34 (Beijing: Science and Technology of China Press) pp 3811-3815

[14] Liang Z Liu D & Sun Y 2022 Micro-expression recognition method combining migration learning and separable three-dimensional convolution Computer Engineering vol 48 (Beijing: Science Press) pp 228-235

[15] Qiu Y Hui Y Zhao P Wang M Guo S Dai B Dou J Bhattacharya S & Yu J 2024 The employment of domain adaptation strategy for improving the applicability of neural network-based coke quality prediction for smart cokemaking process Fuel vol 372 (Amsterdam: Elsevier) p 132162

DOI: 10.54254/2755-2721/87/20241545

[16] Yan J Lu G Li H & Wang S 2018 Dual-modal emotion recognition based on face expression and speech Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition) vol 38 p 60-65