Electronic Engineering and Informatics
G. Izat Rashed (Ed.)

doi:10.3233/ATDE240135

Real-Time Face Expression Recognition Monitoring Using Deep Learning

Yuan XU[1]
School of Artificial Intelligence, Hohai University, Nanjing, China

Abstract. Face expression recognition plays a crucial role in emotion recognition, human-computer interaction and other fields. The aim of this paper is to implement a deep learning-based monitoring system for real-time facial expression recognition. In this paper, a public expression dataset ExpW is used, which is a widely used public face expression dataset. The dataset contains face images from various scenarios in the real world, covering a rich variety of expressions, encompassing seven major expression categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The YOLOv5 target detection algorithm is selected on the deep learning framework for the training and testing of the training set, which is a high-performance algorithm for fast and accurate implementation of the target detection task through a simplified network structure and optimized algorithm design. The experimental evaluation results can be concluded that the system in this paper achieves accurate expression recognition in real-time scenarios. The results of this paper show that real-time face expression recognition monitoring based on deep learning has potential applications in emotion recognition and human-computer interaction.

Keywords. Face expression recognition; YOLOv5; Real-time monitoring

1. Introduction

1.1 Introduction to the background of the study

Psychologist Mehrabian found that 7% of human information is transmitted through language and 55% through facial expressions [1]. Facial expression recognition is an important research topic in computer vision, through the recognition of facial expression can promote the understanding of human psychological state and emotion, expression as an important basis for judging human subjective emotions and has a wide range of application value in the field of public security and psychotherapy [2]. In recent years, many people have begun to use deep learning to automatically recognize facial expressions, which greatly improves the feasibility of the application of expressions [3]. The addition of real-time monitoring in facial expression recognition can be used in a variety of practical scenarios. Traditional facial expression recognition methods usually rely on hand-designed features and classifiers, but there are many challenges in real-time monitoring scenarios.

The aim of this paper is to implement a deep learning based real-time face expression recognition monitoring system. The system can accurately recognize and quickly

[1] Corresponding author: Yuan XU, School of Artificial Intelligence, Hohai University, e-mail: 1906040120@hhu.edu.cn

respond to face expressions in real-time scenes in a short period of time, and at the same time, the system has a certain degree of robustness [4], which enables it to accurately recognize face expressions in complex real-world environments. By constructing such a deep learning-based real-time face expression monitoring system, it can bring potential application prospects for various fields. For example, in terms of emotion recognition, the system can be applied in the fields of emotion monitoring, psychological research and user experience evaluation [5]. Meanwhile, the system can also play an important role in human-computer interaction, such as in the fields of virtual reality, augmented reality and intelligent robotics, to realize a more natural and intelligent human-computer interaction experience.

1.2 Overview of Existing Results in Face expression recognition

Traditional machine learning methods have achieved some results in face expression recognition. For example, classification and recognition of face expressions can be achieved using feature extraction and classifier based methods such as LBP features and SVM classifiers [6]. Also, there are some approaches using feature point based methods such as Active Shape Models (ASM) and Active Appearance Models (AAM) for modeling and recognition of face expressions [7]. Convolutional neural network (CNN) is one of the widely used methods in deep learning for face expression recognition. By using convolutional, pooling and fully connected layers, CNNs can automatically learn the feature representation of face expressions. Some classical CNN models, such as AlexNet, VGGNet and ResNet, have achieved better performance in face expression recognition tasks [8]. Facial static or dynamic information is widely used in face expression recognition. By utilizing video sequences or facial motion information, changes in facial expressions can be captured more accurately. Some methods, such as 3D model-based face deformation modeling and optical flow-based methods, can extract facial dynamic information and use it for face expression recognition [9]. For solving the problem of cross-dataset and cross-domain in face expression recognition, some research works focus on solving the differences between different datasets and propose some domain adaptive methods such as deep domain adaptive networks (DAN) and multi-task learning methods [10].

Chapter 1 briefly introduces the research background of deep learning based face expression recognition and the research objectives of the real-time face expression recognition monitoring system, and reviews the existing results of others. Chapter 2 describes the methodology and process of this paper, briefly describes the face training set and the test set used, outlines the selection of the deep learning model and the strategy of data training and model optimization. Chapter 3 describes the experimental setup and the model performance evaluation and other metrics, and briefly analyzes the experimental results. Chapter 4 summarizes the results of the study and develops an outlook for future work.

2. Materials and Methods

2.1 Dataset description

A personally improved ExpW dataset is used as the training set for this paper. The original ExpW (The Expression in-the-Wild Database) expression dataset has a total of

(91,793) expressions, which are divided into seven expressions: angry, disgust, fear, happy, sad, surprise, and neutral, including 3671 angry, 3995 The ExpW emoji dataset consists of seven types of emoji, including 3671 for angry, 3995 for disgust, 1088 for fear, 30537 for happy, 10559 for sad, 7060 for surprise, and 34883 for neutral.

Among the ExpW expression dataset there are certain problems such as face tilting, recognizing irrelevant data as facial data, corresponding facial data labeling non-owner face data, so the ExpW expression dataset is improved by adopting Zhao Dongyu from Peking University, the original dataset is preprocessed, the aspect ratio of the face is fixed to 112 × 112, and the relabeled face data includes 3585 sheets of angry, 3861 disgust, 1053 fear, 29243 happy, 10039 sad, 6882 surprise, and 32642 neutral. With the traditional ratio of the number of training set, validation set, and test set as 6:2:2, Fer2013 (35886 sheets) is chosen here as the validation set, which has a fixed size of 48×48 to facilitate the validation of the training results.

The RAF-DB dataset (29,672 sheets) is chosen as the test set, which is a natural real-world image, and each image is different in terms of age, gender and race of the subjects, head pose, lighting conditions, occlusion, post-processing operations, etc., so that the framing ability of the face data and the ability of emotion recognition can be effectively measured in real-time after the training of the YOLOv5 network and facilitates the better evaluation of the real-time monitoring system of facial expressions. Evaluate the performance of the real-time face expression monitoring system.

2.2 Deep learning models

In this paper, the YOLOv5 model was chosen as the base model for this paper. The structure of this model consists of multiple inputs, Backbone network, Neck network, outputs, activation function, and the Mish activation function, which is an alternative activation function to ReLU that uses a single neural network and avoids the complex multi-stage process. This makes YOLOv5 easy to implement and deploy, and allows it to run efficiently on devices with limited computational resources. YOLOv5 also uses a method called "anchor-based", which predicts the position and size of a target from a predefined set of anchor points. These anchors correspond to the different scales of the input image, thus enabling YOLOv5 to detect targets at different scales.In the Yolov5 framework, there are some special structures and modules that are used to enhance the performance and efficiency of the model, such as the CBL module, the Focus module, the CSP module, and so on.The following is a brief introduction to some important structures.The CBL is the basic convolutional block in Yolov5, which consists of a series of operations: convolutional layers are used for feature extraction, batch normalization is used for accelerating training and stability, and the LeakyReLU activation function is used for introducing nonlinearities. LeakyReLU activation function is used to introduce nonlinearities.The CBL structure is widely used throughout the model to help enhance feature representation and accelerate convergence.The Focus structure is used to enhance detection of small targets. It converts the input feature maps into feature maps with a higher number of channels before proceeding to the next step. This helps the model to focus more on the localization and classification of small targets.The SAM module is used to enhance the model's ability to perceive spatial information. It calculates the mean and maximum values of the channel dimension of the feature map and then combines the information from both, enabling the model to better focus on important feature regions, thus improving detection accuracy.CSPDarknet53 is a feature extraction network which is an improvement of the Darknet53 network.CSPDarknet53 introduces the CSP (Cross

Stage Partial connections) structure, which enhances feature transfer and representation by splitting the feature map into two parts and performing inter-channel connections.DetectHead is the detection part in Yolov5 for the final target detection output. It contains a series of convolutional and fully connected layers that are used to predict the location, class and confidence score of the target. The Yolov5 framework is shown as Figure 1.

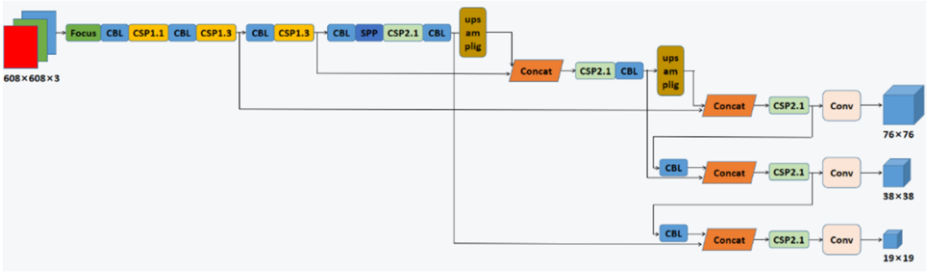

Figure 1. Yolov5's network structure

Overall, YOLOv5 employs a simplified network structure and an efficient design that is fast, accurate and easy to implement, making it more suitable for accomplishing real-time target detection tasks such as face recognition.

2.3 Data training and model optimization strategies

In this paper, a training set is used to train the model and a validation set is used for parameter tuning. In order to prevent overfitting and improve the generalization ability of the model, this paper employs an early stopping strategy and a learning rate decay technique. Specifically, the early stopping strategy is used to monitor the loss value on the validation set during the training process. If the loss on the validation set no longer decreases, this paper stops training and selects the model with the best performance as the final model.

In addition, this paper introduces a learning rate decay technique to optimize the training process of the model. Learning rate decay is a method to gradually reduce the learning rate, which helps the model to converge better and avoid falling into local optimal solutions. In this paper, the learning rate is dynamically adjusted according to the loss changes during the training process to improve the training effect of the model.

In order to monitor and tune the performance of the model more intuitively, this paper uses wandb (Weights & Biases) to visualize the training process. wandb is a powerful experimental tracking and visualization tool, which helps this paper to monitor the training metrics, the changes of the loss function, and the performance of the model in real time. By visualizing the training process, this paper can better understand the training status of the model and make timely model selection and parameter tuning. The above strategies and techniques help to ensure that this paper selects a model with the best performance and provides accurate and reliable results for subsequent test set evaluation [11].

2.4 Model evaluation formula

Based on the trained YOLOv5 model, the total loss function used in this face expression

monitoring system is summed by the following three loss functions: classification loss Classification loss: used to measure the accuracy of the model for each category, the cross entropy adopted for this loss as a measure is as follows:

$$H(p,q) = -\sum_i p_i \log_2(q_i) \quad (1)$$

$p_i(x)$ denotes the true label of the face expression dataset and $q_i(x)$ denotes the model prediction result.

ii) localization loss Localization loss: this loss is used for the error between the prediction box and the real box, which is convenient to measure the prediction accuracy of the model on the position of the target box, and this loss adopts the mean-square error as a measure, as follows:

$$MSE = \frac{1}{n}\sum_i (y_i - \widehat{y_i})^2 \quad (2)$$

y_i denotes the true position of the face in the dataset and $\widehat{y_i}$ denotes the model predicted position.

iii) Confidence loss: The object loss is used to measure the prediction accuracy of whether the target box of the model contains the object or not, and the loss adopts the binary cross entropy as a measure, as follows:

$$BCE(y,\hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y}) \quad (3)$$

y_i denotes the true value and $\widehat{y_i}$ denotes the model predicted position.

3. Results

3.1 Experimental setup and assessment indicators

The experimental platform of this paper is 64-bit Windos10 operating system, the processor is Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz 2.40 GHz,NVIDIA RTX 1050 graphics card, using Pytorch deep learning framework, the implementation environment is pytorch1.8,python3.7. training The number of rounds epochs was set to 300. In order to effectively evaluate the experimental effect, this experiment adopts two indexes of detection accuracy (P) and mean average detection accuracy (mAP) as the main judging criteria.

3.2 Results of model performance evaluation

After training and testing, the model in this paper achieved an accuracy of 71.9% on the test set. In the results, it shows that the model performs well on the expression categories such as happy and sad, but the performance is slightly lower on the expression categories such as fear and disgust. The training set, validation set and test set were divided according to 6:2:2. The data sets are shown in Figure 2 for detection accuracy and mAP0.5 curve in YOLOv5s model.

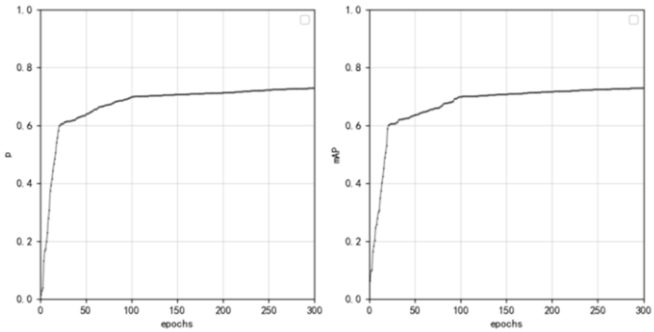

Figure 2. Algorithm Detection Accuracy Curve and mAP Curve Change

The detection results of this algorithm on the test set RAF-DB are shown in Table 1.

Table 1. Comparison of experimental data

Facial expression	YOLOv5	
	P	mAP
All	0.72	0.76
angry	0.813	0.866
disgust	0.634	0.666
fear	0.647	0.671
happy	0.852	0.908
sad	0.625	0.583
surprise	0.786	0.837
neutral	0.683	0.788

3.3 Performance test results

In order to evaluate the performance of a real-time deep learning-based face expression monitoring system, this paper applies the system to real scenes using a camera and conducts comprehensive performance tests. The experimental results show that the system is able to accurately recognize face expressions in real-time situations and maintains stability under different lighting, postures and other complex conditions.

By testing in real scenarios, this paper verifies the reliability and robustness of the deep learning-based real-time face expression monitoring system. The system is able to respond to and capture face images in real time, instantly perform feature extraction and expression classification on the images. Whether in indoor or outdoor environments, the system is able to recognize face expressions effectively, and even in the presence of interfering factors such as lighting changes, posture changes, and other natural occlusions, the system's performance still exhibits good stability and accuracy.

4. Conclusion

The aim of this paper is to implement a deep learning based monitoring system for real-time facial expression recognition. By using public datasets for training and testing, the system implemented in this paper achieves more accurate expression recognition in real-

time scenarios. The results show that deep learning-based real-time face expression recognition monitoring has potential applications in emotion recognition and human-computer interaction. The above test results provide strong support for the practical application of deep learning-based real-time face expression monitoring systems. For example, in the field of emotion recognition and psychological research, the system can be used to monitor and record people's emotional changes in real time, thus providing more accurate and objective emotional data. In the field of user experience evaluation and intelligent interaction, the system can be applied to automatically recognize and respond to users' expressions to achieve a more intelligent and natural human-computer interaction experience. In summary, the performance test results of the deep learning-based real-time facial expression monitoring system in different real-world scenarios show its reliability and robustness. This provides strong support for the application of the system in the fields of emotion recognition, user experience evaluation and human-computer interaction, and provides a valuable reference for future research and application. Through further improvement and optimization, the system is expected to play a greater role in practical applications and bring more innovation and development to related fields.

Although the system in this paper achieved good performance in real-time scenarios, there is still some room for improvement. Future work can explore more complex deep learning models, such as attention mechanisms and generative adversarial networks, to further improve the accuracy and robustness of expression recognition. In addition, the construction of human-computer interaction interfaces can be considered to be combined into real-time facial expression monitoring for more comprehensive emotion recognition capabilities.

References

[1] David Matsumoto, "More evidence for the universality of a contempt expression," Motivation and Emotion, (1992).
[2] Xue Y., Mao X., Guo Y., "The research advance of facial expression recognition in human computer interaction," Journal of Image and Graphics 14(05), 764-772 (2009).
[3] Zeng Z, Pantic M, Roisman G I, et al, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," IEEE Trans Pattern Anal Mach Intell 31(1), 39-58 (2009).
[4] Chi L., Li M., Wu X., "A motion objects detection algorithm with real-time and robust properties," Computer Applications and Software 32(02), 132-134 (2015).
[5] Wu Y., LIU W., Zhang K., "Study of intelligent tutoring system based on affective recognition," Computer Engineering and Design 29(9), (2008).
[6] Yang, H., Kwon, J., & Cho, N. I., "Facial expression recognition using local directional pattern and ensemble of classifiers," Pattern Recognition Letters 32(10), 1410-1417 (2011)..
[7] Cootes, T. F., Edwards, G. J., & Taylor, C. J., "Active appearance models," IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681-685 (2001).
[8] Simonyan, K., & Zisserman, A., "Very deep convolutional networks for large-scale image recognition," arXiv preprint, 1409.1556 (2014).
[9] Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J., "A 3D facial expression database for facial behavior research," In Proceedings of the 7th IEEE International Conference on Automatic Face and Gesture Recognition, 211-216 (2006).
[10] Long, M., Cao, Y., Wang, J., & Jordan, M. I., "Learning transferable features with deep adaptation networks," In Proceedings of the 32nd International Conference on Machine Learning, 97-105 (2015).
[11] Wang X., Wu W., Zeng Z., "Study of Face Detection Algorithm based on Video," Electronic Science and Technology 33(02), (2020).