A facial and vocal emotion recognition system based on deep learning

Aiden Gu

Shanghai American School, Shanghai, China

aidenjiliangu@outlook.com

Abstract. The traditional teaching mode, mainly characterized by "cramming" instruction, often overlooks the absorption of knowledge by students during the teaching process. Even in recent years, the emerging "flipped classroom" model heavily relies on teachers' experience and expertise, making it challenging to quantify students' classroom learning states. Among young teachers lacking teaching experience, the replication of excellent teaching models faces difficulties. With the rapid development of artificial intelligence technology, deep learning and other AI techniques are increasingly applied to people's lives. Additionally, emotion computing technology is also becoming more sophisticated. In the field of education, the use of AI technology holds the potential to help teachers systematically and comprehensively assess teaching quality, promoting the sharing of high-quality educational resources nationally and globally. Existing emotion detection technologies often focus on single techniques such as image recognition or natural language processing, resulting in low accuracy in discerning human emotions and making it difficult to accurately identify emotional changes in complex scenarios such as classrooms. Furthermore, there is a lack of comprehensive emotional statistical analysis methods for audio and visual data. In light of these, we have designed and implemented a multi-modal emotion recognition and analysis system. Through testing in classroom settings, this system has the potential to assist classroom teaching and enhance teaching quality.

Keywords: Deep Learning, Affective Computing, Video Processing, Audio Processing, Multi-Model.

1. Introduction

Yu Gu, Eric Postma, and Hai-Xiang Lin present that vocal emotion recognition aims to identify the emotional states of speakers by analyzing their speech signals [1]. Jordan J, Azhar Aulia, Naoyuki Kubota, and Ahmad Lotfi use Machine Learning to make an affective computing study, which is based on computer vision [2]. Each human face is extracted from an image before computing the emotional states.

Multimodal learning has gained immense popularity due to the explosive growth in the volume of image and textual data in various domains. Vision-language heterogeneous multimodal data has been utilized to solve a variety of tasks, including classification, image segmentation, image captioning, question-answering, etc. Priyankar Bose, Pratip Rana, and Preetam Ghosh revisit the various attention mechanisms on image-text multimodal data from its inception in 2015 till 2022 and considered a total of 75 articles for the survey [3].

Artificial intelligence, as a discipline, began to take shape in the mid-20th century, with Alan Turing proposing the concept of the first computer program. This can be viewed as the precursor to early AI. In the evolution of artificial intelligence, various schools of thought emerged, including symbolic AI, expert systems, machine learning, and more. Today, deep learning plays an increasingly vital role in fields such as image recognition, pattern classification, and natural language processing. With the advancement of technology, human emotions and sentiments have also become quantifiable.

We propose a method for multimodal emotion recognition and statistical analysis in scenarios such as classrooms and have systematically implemented this approach. This method integrates facial keypoint detection techniques from visual image processing and speech analysis techniques. Emotion computation and statistical analysis are performed using a multimodal approach. Experimental results demonstrate that this system can accurately assess classroom teaching effectiveness.

The scientific contributions of this work relate to the exploration of computer vision model topologies and vocal model technologies, which are used to predict emotional states. Following 10 evaluation experiments, the final model is able to predict emotional states of valence with relatively high accuracy and low loss. We open-source our model and make it available to the research community for future work.

The remainder of this article is as follows: Section 2 presents a review of the literature on related work in the field. Section 3 explains the methodology followed by a summary of the experiments carried out in this study, before the results are presented in Section 4. Finally, future work is discussed and this study is concluded in Section 5.

2. Methodology

The aim of the study is to propose a method for multimodal emotion recognition and statistical analysis in scenarios such as classrooms, which will help improve the efficiency of students' studies.

2.1. DataSet

We trained the facial keypoint recognition model using an open-source face_landmark dataset. The facial feature detection includes 150 key points for facial features such as cheeks, eyebrows, eyes, mouth, nose, and facial contours. The audio feature detection includes pitch, sound intensity, chroma, sound quality, and spectrogram. We also collected and annotated audio and video data ourselves. The self-annotated audio and video samples are mainly categorized into seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

2.2. Data Preprocessing

We need to extract key information from the collected audio and video files, remove redundant information, and compress data that remains unchanged for a long time. Additionally, we added some noise data to avoid overfitting. For video data, we preprocessed it using the Histogram of Oriented Gradients (HOG) method and identified faces in the video. For audio data, we performed preprocessing steps such as log-mel-spectrogram extraction.

2.3. Deep Learning Architecture

Face recognition is an important computer vision task affected by several factors. These factors include the face pose of the input image, features used to describe the image, illumination conditions, and facial expression. Shinfeng D. Lin, and Paulo Linares Otoya propose a pose-invariant face recognition framework based on large pose detection and facial landmark description, which gives us a lot of inspiration [4].

Our deep learning framework is implemented based on TensorFlow. The video deep learning architecture consists of 34 fully connected layers, 36 regularization layers, 4 pooling layers, and 36 activation layers. As for the audio deep learning architecture, it includes 4 fully connected layers, 4 regularization layers, 4 max-pooling layers, and 4 activation layers. The architecture also utilizes SGD (Stochastic Gradient Descent) optimization.

2.4. System Architecture

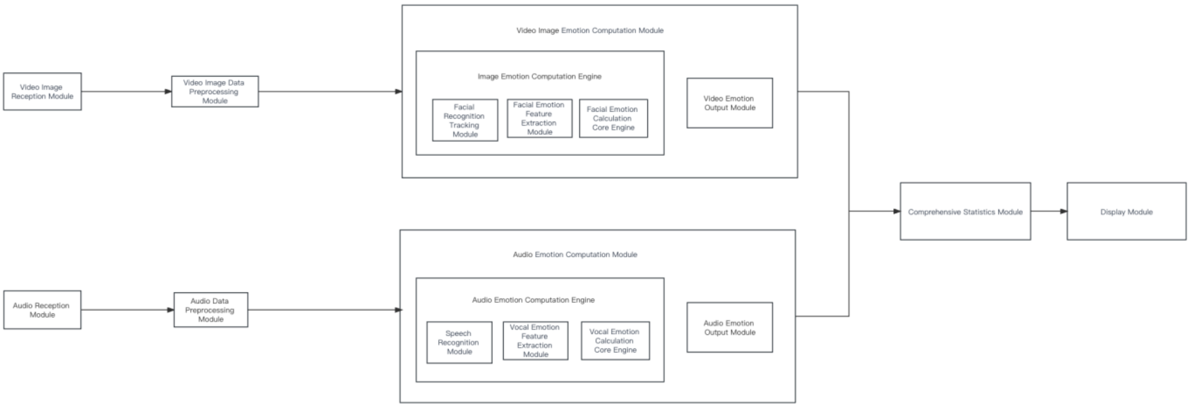

Video Image Reception Module
Video Image Data Preprocessing Module
Video Image Emotion Computation Module
Image Emotion Computation Engine
Facial Recognition Tracking Module
Facial Emotion Feature Extraction Module
Facial Emotion Calculation Core Engine
Video Emotion Output Module
Audio Reception Module
Audio Data Preprocessing Module
Audio Emotion Computation Module
Audio Emotion Computation Engine
Speech Recognition Module
Vocal Emotion Feature Extraction Module
Vocal Emotion Calculation Core Engine
Audio Emotion Output Module
Comprehensive Statistics Module
Display Module

Figure 1. Architecture Diagram of Multimodal Audio-Visual Emotion Analysis and Statistical System

Sun X, Ma S, and Li Y use multimodal deep learning to detect malicious traffic with noisy labels [5]. This is of great help for designing deep neural network architectures.

The multimodal audio-visual data analysis system for the classroom scenario consists of five main modules, which are described as follows:

Multimodal Data Reading and Preprocessing Module: This module receives multimodal data such as video images and audio and performs parsing and preprocessing. It includes specific sub-modules for video image processing and audio processing. The video image processing module is responsible for enlarging and normalizing faces. The audio processing module focuses on extracting acoustic prosodic features from speech signals, such as logarithmic extraction of linear predictive cepstral coefficients and splitting the audio spectrogram using a rolling window.

Video Image Emotion Calculation Module: This module performs emotion analysis calculations on the preprocessed video image data. It consists of an image emotion calculation engine and a video image emotion output module. The image emotion calculation engine includes a face recognition and tracking module for identifying and tracking facial data in visual images, including eyes, eyebrows, nose, mouth, and chin. The facial feature extraction module extracts facial key points from the facial data and obtains emotion feature values based on the keypoint sequence. The facial core emotion calculation engine takes the seven typical emotion parameters mentioned earlier as input, predicts the current emotional state based on the extracted emotion feature values, determines the target object's emotional state, and outputs confidence levels by comparing them with a pre-constructed map of psychological behavior.

Audio Emotion Calculation Module: This module performs emotion analysis calculations on the preprocessed audio data. It includes an audio emotion calculation engine and an audio emotion output module. The emotion calculation engine consists of a speech recognition module, an emotion feature extraction module, and an audio core emotion calculation engine. The speech recognition module converts the preprocessed speech signal into a digital signal sequence. The sentence feature extraction module extracts sentence features from the digital sequence. The audio core emotion calculation engine takes the seven typical emotion parameters as input, constructs a machine learning model, extracts speech features, predicts the current emotional state, and inputs the classified speech features into the emotion model to obtain the emotional state of the speech signal and output confidence levels.

Statistical Analysis Module: This module performs inductive analysis by statistically integrating the results of emotion calculations, resulting in comprehensive analysis results.

Visualization and Reporting Module: This module visualizes the results obtained from the statistical analysis module for display and reporting purposes.

2.5. Training Process

After multiple rounds of training, by effectively learning and adjusting the weight matrices and biases during the training process, the system achieves a 95% accuracy in classifying and recognizing emotions more accurately. This is accomplished by using backpropagation to iteratively adjust the matrices, reducing the gap between predicted results and true emotion labels. Multiple rounds of training and backpropagation are commonly used methods in deep learning, allowing for the iterative optimization of model parameters. These methods continuously minimize the difference between predicted results and true emotion labels.

3. Evaluation and results

Here, we have designed an experimental process to validate and analyze the effectiveness of the multimodal audio-visual emotion detection system. We utilize video streams and audio streams collected in a classroom scenario as the multimodal input data. The analysis performed by the emotion detection system yields comprehensive statistical results, which are then expressed and visualized. We propose the following 8 steps as shown in figure 2:

Step 1: The device receives video image data input.

Step 2: The device receives audio data input.

Step 3: Video image preprocessing.

Step 4: Audio preprocessing.

Step 5: Emotion analysis and prediction using the video image emotion calculation engine.

Step 6: Emotion analysis and prediction using the audio emotion calculation engine.

Step 7: Comprehensive statistical analysis of the emotion prediction results.

Step 8: Visualization of the comprehensive emotion prediction results through graphical representation.

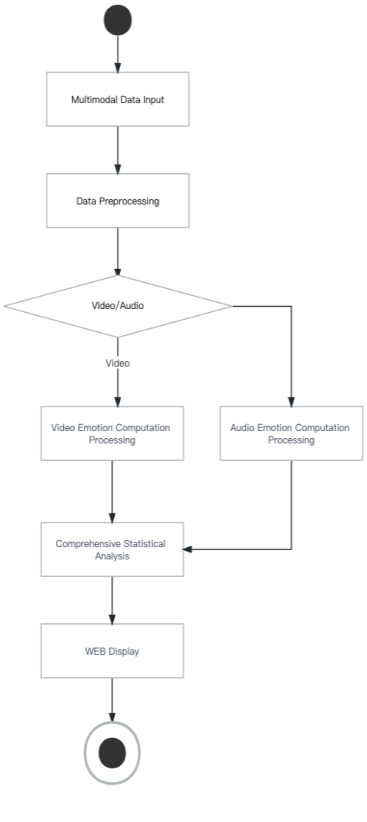
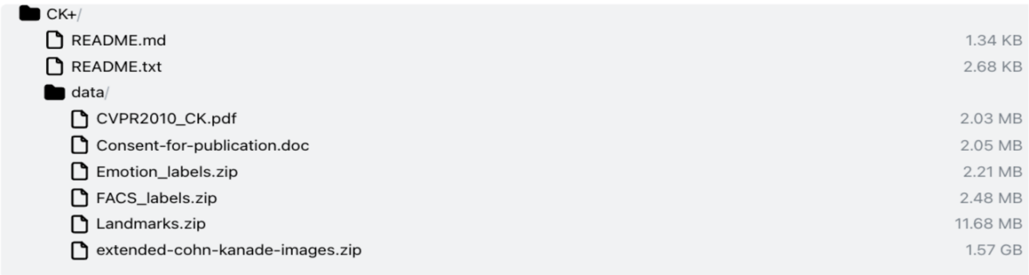

Multimodal Data Input
Data Preprocessing
Video/Audio
Video
Video Emotion Computation Processing
Audio Emotion Computation Processing
Comprehensive Statistical Analysis
WEB Display

Figure 2. State Machine Diagram for Multimodal Emotion Recognition and Statistical Analysis System

3.1. Experiments

In the experimental validation section, we use the CK+ dataset, which stands for the Extended Cohn-Kanade dataset. T. Kanade, J. F. Cohn and Yingli Tian proposed this dataset consisting of 593 video sequences [6]. The CK+ dataset is widely regarded as the most commonly used facial expression classification database and has been used in major facial expression classification methods.



```
CK+/
  README.md                                  1.34 KB
  README.txt                                 2.68 KB
  data/
    CVPR2010_CK.pdf                          2.03 MB
    Consent-for-publication.doc              2.05 MB
    Emotion_labels.zip                       2.21 MB
    FACS_labels.zip                          2.48 MB
    Landmarks.zip                           11.68 MB
    extended-cohn-kanade-images.zip          1.57 GB
```

Figure 3. CK+ dataset directory structure graph

The videos in the CK+ dataset come from 182 different subjects, ranging in age from 18 to 50 years, with varying genders and ethnicities. Each video demonstrates facial transitions from a neutral expression to an extreme expression, recorded at a rate of 30 frames per second (FPS) and a resolution of 640 x 490 pixels.

We synthesized the video frame data into .mp4 format videos to meet the requirements of our system. We obtained a total of 593 video clips, each labeled as one of the seven emotion categories: anger, contempt, disgust, fear, happiness, sadness, and surprise.

By combining the labels provided in the CK+ dataset, we conducted evaluations of the recognition performance. Each experiment ultimately outputs the proportions of various emotions, similar to the example shown in table 1. Through statistical analysis, we found that the percentage of correctly identified single emotions in each output was 98.65%.

Table 1. Proportions of various emotions analyzed in an emotion detection experiment

Facial Emotions		
#	Emotion	Probability
1	Anger	9
2	Disgust	54
3	Fear	0
4	Happy	0
5	Sad	0
6	Surprise	0
7	Neutral	36

3.2. Application Scenario

In the current experimental setup, we deploy video-audio capture devices such as cameras and microphones to capture clear facial expressions and sounds of teachers and students as the primary objective. After completing the collection of video and audio data, the corresponding data files can be uploaded to the backend of the multimodal audio-visual emotion detection system. The system analyzes the emotional state changes of teachers and students, and outputs comprehensive statistical analysis results in the form of graphs. Through the experiment, we can promptly understand the interaction between teachers and students in the classroom, compare the emotional state changes of both sides, assess the level of student engagement, and predict the quality of classroom teaching. In more severe instances, if a student's negative emotions are excessively high, we can offer alerts and reminders to the teacher, effectively functioning as a classroom alarm system

In online teaching scenarios utilizing platforms such as Tencent Meeting and Classin, the system can also be activated in the background to perform real-time multimodal emotion analysis.

After small-scale deployment and usage, the multimodal emotion detection and statistical analysis system has shown good performance in both offline and online scenarios. It provides effective alerts when students are either too active or too passive, leading to improved attentiveness and learning efficiency as reported by users. However, it is important to acknowledge that environmental noise can have a negative impact on the system. In offline classrooms, background noise or suboptimal camera angles can significantly affect the accuracy of emotion recognition. In online classrooms, it is inevitable that people may leave the camera or mute the microphone, which can also impact the system's effectiveness.

4. Discussion

Regarding the proposal of this system, we acknowledge that there are several positive impacts on the overall improvement of teaching quality. However, concerns related to ethics and morality do require further discussion.

Firstly, the promotion of classroom norms.

In the context of using this system, the classroom, which was previously more informal, becomes more controllable. It enables teachers to monitor the attention levels of all students, rather than just a select few. Furthermore, it allows for prompt feedback to be provided to teachers regarding the learning states of students, enabling them to make flexible adjustments to the pace of the classroom.

Secondly, cost savings in human resources.

After each day's classes, the system can generate classroom situation reports in batches, significantly reducing the cost of manual analysis and relieving the workload of supervisory teams.

Thirdly, ethical and moral concerns.

When collecting audio and video data comprehensively, further discussions are needed to ensure the protection of personal privacy for students and teachers. Robust legal protections are necessary to prevent the unauthorized sale and misuse of such personal data.

Fourthly, the quantifiability of learning.

In supervised scenarios, it remains a topic of discussion whether students can concentrate better and transform knowledge into what they truly need. Or are students merely performing the role of "classroom actors" and appearing attentive without genuine engagement?

Fifthly, the accuracy of emotion recognition rates.

It is important to note that accuracy is only an evaluation result based on training and testing data and may not fully represent the system's performance in real-world scenarios. In practical applications, our system may face additional challenges, such as different environmental conditions and individual differences. Therefore, to ensure the system's robustness and accuracy in real-world scenarios, more extensive testing and evaluation, as well as necessary optimizations and adjustments, may be required.

5. Conclusion

The proposed multimodal emotion computation and statistical analysis system integrates speech recognition and visual image technologies to enable simultaneous emotion recognition from multiple channels. This system accurately analyzes the emotions of the target subjects and evaluates students' learning states during the teaching process.

From an implementation perspective, the system includes the following steps:

1. Pretraining of deep neural network models.
2. Reading of multimodal video images and audio input data.
3. Preprocessing using techniques like Histogram of Oriented Gradients to detect faces in the video.
4. Preprocessing of audio data, such as Log-mel-spectrogram feature extraction.

5. Using a pre-trained model and an emotion computation engine to predict emotions from facial expressions in the video and emotions expressed in the audio.

6. Integration of emotion statistical analysis results from video images and speech, and visualization of the findings.

By analyzing the proportions of various emotions and the overall positive/negative emotions in teachers and students during teaching sessions, this system provides a comprehensive evaluation of the teaching quality. It serves as a reference for teaching quality committees, promoting intelligent teaching management decision-making and personalized educational implementation.

There are certain limitations that need to be taken into consideration. Since this paper proposes a computer prototype system, we provide the hardware and software specifications used in our implementation.

Acknowledgement

Words cannot express my gratitude to my advisor Lucas Lu, a graduate of the Computer Science School at Fudan University who gave me meaningful guidance on every aspect of this project. This journey would not have been possible without his expertise. I am also grateful to my schoolmates, especially the five sophomores and juniors who participated in the interviews and also took the MBTI personality test. Lastly, I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

References

[1] Yu Gu, Eric Postma, and Hai-Xiang Lin. 2015. Vocal Emotion Recognition with Log-Gabor Filters. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC '15). Association for Computing Machinery, New York, NY, USA, 25–31. https://doi.org/10.1145/2808196.2811635

[2] J. J. Bird, A. A. Saputra, N. Kubota and A. Lotfi, "Affective Computing in Computer Vision: A Study on Facial Expression Recognition," 2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter), Phuket, Thailand, 2022, pp. 84-88, doi: 10.1109/IIAI-AAI-Winter58034.2022.00027.

[3] P. Bose, P. Rana and P. Ghosh, "Attention-Based Multimodal Deep Learning on Vision-Language Data: Models, Datasets, Tasks, Evaluation Metrics and Applications," in IEEE Access, vol. 11, pp. 80624-80646, 2023, doi: 10.1109/ACCESS.2023.3299877.

[4] S. D. Lin and P. Linares Otoya, "Large Pose Detection and Facial Landmark Description for Pose-invariant Face Recognition," 2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII), Hualien, Taiwan, 2022, pp. 143-148, doi: 10.1109/ICKII55100.2022.9983525.

[5] Sun X, Ma S, Li Y, Wang D, Li Z, Wang N, Gui G. Enhanced echo-state restricted Boltzmann machines for network traffic prediction. IEEE Internet of Things Journal, 2020, 7(2): 1287–1297

[6] T. Kanade, J. F. Cohn and Yingli Tian, "Comprehensive database for facial expression analysis," Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 2000, pp. 46-53, doi: 10.1109/AFGR.2000.840611.