Article

CH-CC: A Chinese Multimodal Classroom Atmosphere Analysis Dataset Based on Teachers' Behavior and Voice

Zicen Liao [1,*]

[1] Z.Liao5@newcastle.ac.uk
* Correspondence: Z.Liao5@newcastle.ac.uk; Tel.: (+86 17353264882)

Abstract: Previous studies in multimodal sentiment analysis have utilized limited datasets, particularly for the analysis of classroom atmosphere in teaching scenarios. In this paper, we present a multimodal dataset for the analysis of classroom atmosphere, based on the behavior and voice of teachers in teaching scenarios. The dataset includes refined video clips divided into five categories according to the classroom atmosphere: boring, slightly boring, normal, slightly attractive, and attractive in both visual and audio modalities, and text modality can be added through further annotation. In this paper, we propose four visual models, three audio models, and one visual-audio dual-modality model to be tested on our dataset. These models include ResNet-LSTM, CLIP-LSTM, 2DCNN-LSTM and 3DCNN as visual unimodals, Mel-Spectrogram-LSTM, Opensmile-LSTM and Wav2vec-LSTM as audio unimodals, 3DCNN-Mel-Spectrogram-LSTM as the dual-modality model. The results of our tests show that the visual models and dual-modality model have high accuracy on the CH-CC dataset, while the performance of the audio models is relatively low. The results indicate that the CH-CC dataset is feasible and reliable, and that the visual modality plays a major role in the analysis of this dataset.

Keywords: multimodal dataset; sentiment analysis; classroom atmosphere; intelligent education

1. Introduction

Multimodal atmosphere analysis aims to extract, analyze and interpret the sentimental state of the environment through text, video and audio information. With advances sensor technologies, new forms of intelligent education have become possible [1], and using deep learning to analyze classroom atmosphere through different modalities has become essential for intelligent educational technologies to help students improve their learning efficiency. However, there are few open-source datasets for studying classroom atmosphere, and there has been little focus on the multimodal tasks related to the impact of teachers on classroom atmosphere. Therefore, we propose a Chinese dataset for analyzing classroom atmosphere based on teachers' behavior and voice and have conducted evaluations of its performance.

Early sentiment analysis tasks mainly used datasets that only had one type of data, such as text from Yelp and SST [2] datasets, images from IASP and GAPED [3] datasets, and audio from RECOLA [4] and SWEA datasets. These datasets couldn't mimic real-world scenarios that often involve multiple types of information. With the advancement of multimodal, there are now more datasets like UR-FUNNY [5] and CMU-MOSEI [6] that provide video/image, audio, and text data all at once. These multimodal datasets are more realistic and allow models to perform more accurate sentiment analysis.

Due to the fact that most current research primarily uses everyday natural interactions between people as the data basis for sentiment analysis, and class data is difficult to obtain permission for, there are very few datasets in the field of multimodal sentiment analysis that focus on classroom atmosphere. Additionally, these datasets have not yet been standardized.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

Based on this, we collected classroom videos from different teachers teaching multiple subjects at a Chinese middle school. After simplifying the processing, the videos were divided into 667 segments that lasted 10 seconds each, and 20 volunteers were invited to score the video segments based on classroom atmosphere, with a minimum score of 1 and a maximum score of 10. According to the average score of each video, all the videos were divided into 5 categories: 1-2 Boring, 3-4 Slightly Boring, 5-6 Normal, 7-8 Slightly attractive, 9-10 Attractive.

After this, we tested multiple visual and audio models, as well as a visual and audio multimodal model on this dataset and obtained good accuracy, proving the dataset's availability.

This paper introduces a multimodal Chinese classroom atmosphere dataset based on teacher behavior and speech for the first time, which guarantees the students' information security while providing effective data support for intelligent education in the field of classroom atmosphere analysis and also provides necessary pre-exploration for future research on using intelligent education systems to help improve teaching quality.

2. Related Work

In this section, we briefly review previous research on multimodal datasets and models, and contrast it with our own work.

2.1. Related Datasets

With the development of sentiment analysis and intelligent education, researchers built up multiple datasets for the atmosphere or classroom behaviors, such as BNU-LCSAD [7], Emotic [8], TAD-08 [9], etc. Their specific information is shown in table 1：

Table 1. Information of BNU-LCSAD, Emotic and TAD-08 datasets.

Dataset	Content	Size	Modality	Categories
BNU-LCSAD	Students' behavior	1538 videos	Visual	9 categories of students' classroom behavior
Emotic	Facial expression and environmental atmosphere	23266 pictures	Visual	26 categories of emotions and three categories of atmosphere
TAD-08	Teachers' behavior	2048 videos	Visual	8 categories of teachers' classroom behavior

In these datasets, BNU-LCSAD and TAD-o8 are visual single-modality datasets for classroom behavior, using classroom videos as samples, respectively labeling student or teacher behavior. Emotic is a visual single-modality dataset for facial expressions and environmental atmosphere, which labels the atmosphere into three categories. Researchers can use these datasets to conduct visual modality classification research on teacher or student behavior in the classroom, or to study factors that affect visual modality analysis of atmosphere.

2.2. Dataset used in this experiment

Compared to the three datasets mentioned earlier, the CH-CC dataset also provides data in the audio modality, and focuses on classifying the classroom atmosphere. Researchers can use this dataset to perform audio, visual single-modality, and audio-visual multimodal atmosphere analysis. In future research, researchers can also add text

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

annotations to the video, enabling tri-modal atmosphere analysis. The information of CH-CC dataset is shown in table 2:

Table 2. Information of CH-CC dataset.

Dataset	Content	Size	Modality	Categories
CH-CC	Teachers' teaching videos	667 videos	Visual Audio	5 categories of teaching atmospheres in the classroom

2.3. Related Models

2.3.1 Video processing models

Currently, for video classification tasks in visual modality, the following model schemes are available:

1) Recognition method combining CNN and LSTM. This method generally uses CNN or other 2D feature extraction networks to extract image features from the raw data, then inputting the features into LSTM to obtain the temporal relationship between features, and finally outputting from classifier to recognize and classify videos. This method is used in models such as Motion-Aware ConvLSTM [10] and STS-ALSTM [11].

2) Method using C3D networks. This method generally uses 3DCNN or other 3D feature extraction networks to directly extract image and temporal features from videos and output from classifier to recognize and classify videos. This method is used in models such as C3D[12], T3D[13] and R-M3D[14].

3) Two-Stream method. This method generally uses two CNNs or their derivatives to extract image features and temporal features respectively from raw data, then fusing them and outputting from the classifier to recognize and classify videos. This method is used in models such as TSN[15] and TRN[16].

2.3.2. Audio processing models

A spectrogram can visually represent the changes in sound signals over time. It can be obtained by dividing the input signal into frames, applying windowing, and performing a discrete Fourier transform. After processing, the spectrogram is further input into a Mel filter for dimensionality reduction to enable efficient training of the network. The Mel filter applies non-linear processing to linear frequency bands, increasing the weight of low frequencies and decreasing the weight of high frequencies to emphasize relevant frequency ranges. This processed spectrogram is called a Mel-Spectrogram, which effectively helps the network learn audio features.

Besides spectrograms, other standard parameter sets or pre-trained models can also be used to extract audio features. Due to the focus on certain audio features, such as the time and frequency domains, in audio-based emotion recognition, there may be thousands of low-level parameters involved. Using standard parameter sets or pre-trained models can efficiently extract these audio features that meet the requirements of sentiment analysis.

2.4. Models used in this experiment

The model used in this experiment begins with a simpler network structure in order to reduce complexity and improve training efficiency. On this foundation, it attempts to explore network structures that are more complex but generally have better performance on the dataset.

This experiment, based on the method of combining CNN and LSTM mentioned in section 2.2.1, selected three models: 2DCNN-LSTM, ResNet-LSTM, and CLIP-LSTM. Based on the method of using C3D network, one model, 3DCNN, was also selected. A total of four models were used as test models for visual modality.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

This experiment used Mel-Spectrogram extracted by the LibROSA toolkit [17], audio features extracted by the GeMAPS standard parameter set provided by the Opensmile [18] and Wav2vec [19] pre-trained model as inputs. These three models were used for audio modality testing, which processed the sequences through LSTM and finally performed classification.

The models in the experiment included a relatively basic network in relevant fields to reduce network complexity and improve training efficiency; based on it, it had a try of some network structures with generally better performance to explore the performance in this dataset. After testing with single modality models, this experiment combines the visually and audibly best single modality models, 3DCNN and Mel-Spectrogram-LSTM, into a dual modality model using the Two-Stream method mentioned in section 2.2.1, and tests its performance on the dataset.

3. CH-CC Dataset

We present a multi-modal Chinese classroom atmosphere dataset CH-CC based on teacher behavior and speech, which includes 667 refined video segments, each segment classified into five levels of classroom atmosphere from low to high. Categories and their sizes are shown in table 3:

Table 3. Categories and their sizes (Number of videos) in CH-CC dataset.

Total	Boring	Slightly Boring	Normal	Slightly Attractive	Attractive
667	35	41	203	356	32

3.1 Data Collection and Processing

We collected multiple classroom videos of different teachers during multiple teaching sessions by placing cameras behind the classroom, the main subject of the video is the teacher, and the teachers and students in the video are from the same middle school in Chongqing, China. We used Adobe After Effect software to remove the parts of the video where the teacher was not teaching and divided the video into 667 segments of 10 seconds each with a size of 1280*720, and these segments were labeled and classified according to the scores given by volunteers.

3.2. Annotation

We invited 20 volunteers to score the video segments based on the classroom atmosphere, the lowest score being 1 and the highest being 10, and each video segment was scored by at least 5 people. The final score for each video was obtained by removing the abnormal scores and taking the average score. We labeled all the videos into five categories based on the final score: 1-2 boring, 3-4 slightly boring, 5-6 normal, 7-8 slightly attractive, 9-10 attractive.

4. Experiments

4.1 visual models

The convolutional neural network is one of the most classic networks for processing visual mode classification tasks with simple structure and high training efficiency. Gautam et al. successfully implemented the efficient recognition of classroom teaching videos using 2DCNN-LSTM and 3DCNN [20]; Due to its simple structure and excellent performance in similar tasks, our experiment chose 2DCNN-LSTM and 3DCNN models to test on the CH-CC dataset.

ResNet [21] is a residual network designed based on convolutional neural networks. It can avoid the performance degradation caused by overly complex network structures, while the structure is similar to traditional 2DCNN. CLIP [22] is a currently very popular multi-modal processing model that can output both image and text features with excellent performance. Through it, we can explore the performance of currently efficient multi-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

modal models on this dataset. Therefore, this paper also uses ResNet-LSTM and CLIP-LSTM as visual single-modal models on the CH-CC dataset.

In summary, this paper uses 2DCNN-LSTM, ResNet-LSTM, CLIP-LSTM, and 3DCNN as visual modal test models. The structure of the model is shown in figure 1:

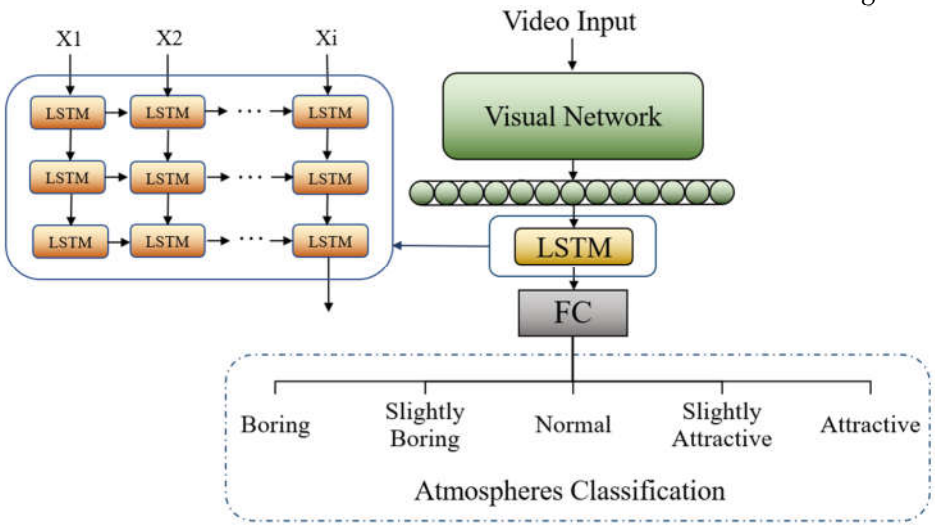


Figure 1. Structure of visual model. In the picture, the Visual Network represents four models: ResNet, CLIP, and 2/3DCNN. When the Visual Network is the 3DCNN model, the LSTM is not needed as 3DCNN has its own time processing.

The specific parameters of the model are shown in Table 4.

Table 4. Parameters of visual models.

Model	Hidden Layers	LSTM	Fc Layers	Optimizer	Learning Rate
ResNet-LSTM	ResNet-152 Default hidden layers	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-3
CLIP-LSTM	CLIP Default hidden layers	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-3
2DCNN-LSTM	32,64,128,256	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-4
3DCNN	32,48	No LSTM	Fc1=256 Fc2=256 Fc3=5	Adam	1e-4

4.2. Audio models

This paper uses Mel-Spectrogram extracted by LibROSA, features extracted by Opensmile standard parameter set GeMAPS and Wav2vec as audio features input to LSTM, and finally output from the classifier as audio modal test models, the structure of the model is shown in Figure 2: Where Audio Network represents LibROSA, Opensmile, and Wav2vec three feature extractors. The specific parameters of the model are shown in Table 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

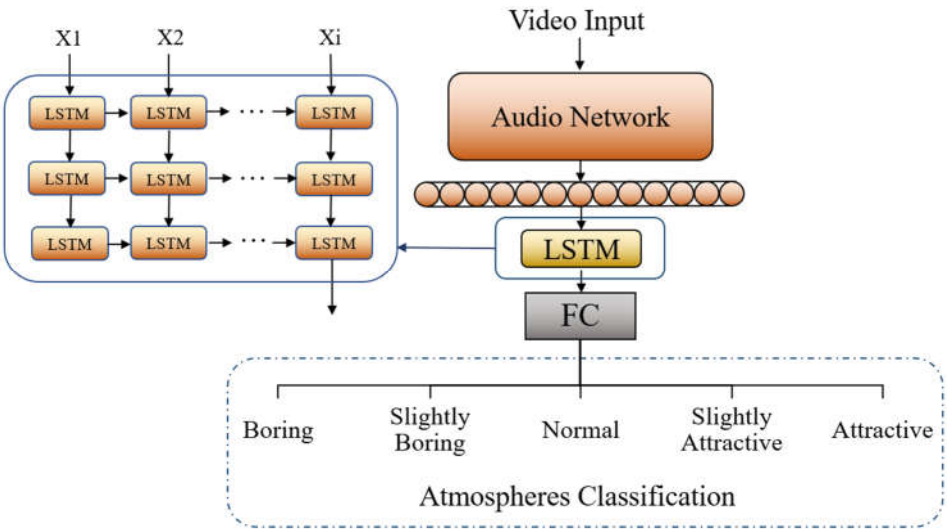


Figure 2. Structure of audio model. In the picture, Audio Network represents LibROSA, Opensmile, and Wav2vec three feature extractors.

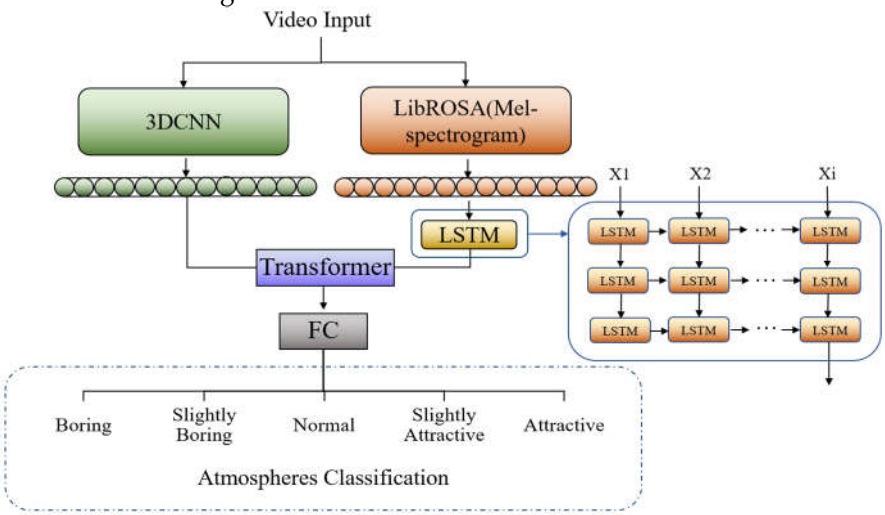
The specific parameters of the model are shown in Table 5.

Table 5. Parameters of audio models.

Model	Feature extracted	LSTM	Fc Layers	Optimizer	Learning Rate
Mel-spectrogram-LSTM	Mel-spectrogram	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-3
Opensmile-LSTM	25-dimensional features extracted from the GeMAPS standard parameter library	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-3
Wav2vec-LSTM	768 dimensional features extracted from the Wav2vec model	3 Layers; 256 hidden nodes	Fc1=256 Fc2=5	Adam	1e-3

4.3. Visual and audio dual-modality model

This experiment combines the visually and audibly single modality models with best performance, 3DCNN and Mel-Spectrogram-LSTM, into a dual modality model, and its structure is shown in Figure 3:



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

Figure 3. Structure of visual and audio dual-modality model.

Where the learning rate is uniform at 1e-3, the optimizer is Adam, and specific model parameters are shown in the Figure 4:

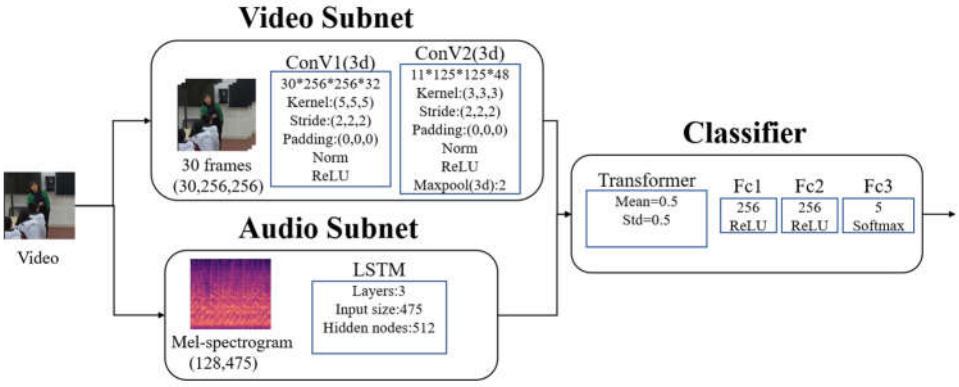


Figure 4. Specific parameters of visual and audio dual-modality model.

5. Results and discussion

5.1 Results and discussion for visual modality model

The results for visual modality models are shown in Figure 5:

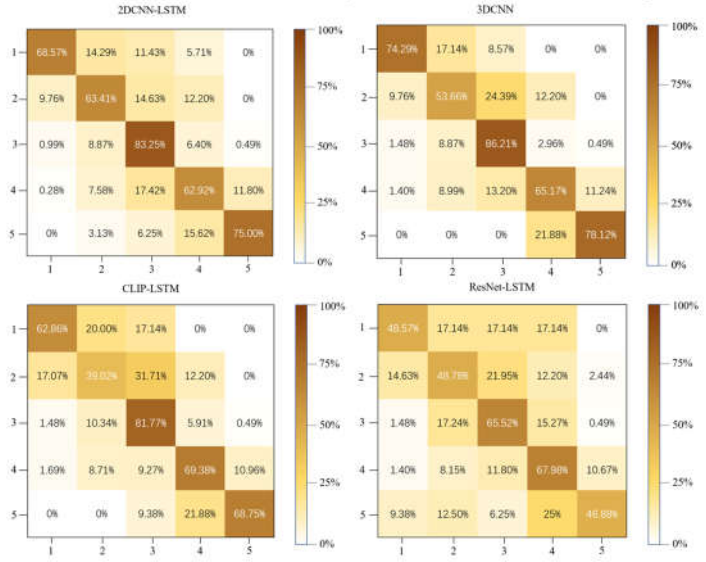


Figure 5. The confusion matrix shows the recognition rate of each visual model for each category. The vertical axis represents the category in which the recognized segments are labeled, and the horizontal axis represents the category in which the model classifies the recognized segments. 1 to 5 represent boring, slightly boring, normal, slightly attractive, and attractive, respectively.

From the confusion matrix results, it can be seen that the 4 models all have high accuracy on this test set, among which 3DCNN has the highest performance in the four models. This may be because 3DCNN can extract the feature representation of a video from 30 images at once, while the remaining three models can only extract the feature representation of a video from 13 images due to the large computation. This means that 3DCNN can analyze videos from a wider scale. At the same time, for the analysis of actions with shorter duration, LSTM's excellence is not enough to make up for the possible feature loss, while 3DCNN, because of its structure-built short-term time series analysis, performs better than the other three models in this task.

5.2. Results and discussion for audio modality model

The results for visual modality models are shown in Figure 6:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

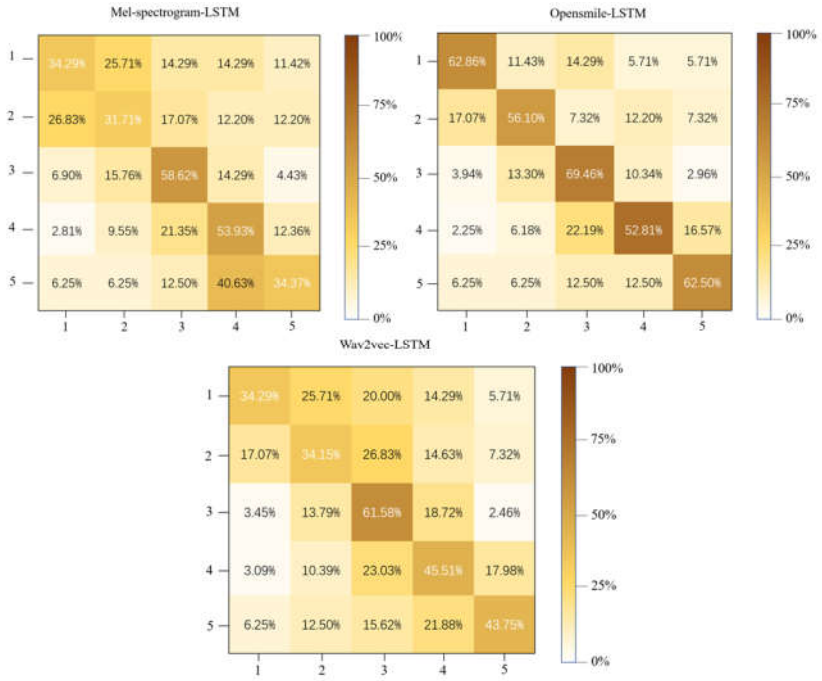


Figure 6. The confusion matrix shows the recognition rate of each visual model for each category.

In the process of testing the audio model, this paper uses the models proposed in the papers Lezhenin, E. [23]; Eyben, F. [18], Meng, L [24]. which have good test results, that is, using Mel-Spectrogram extracted by LibROSA, features extracted by Opensmile with standard parameters, and features extracted by Wav2vec as the audio feature input, and connecting the classifier as the audio modality test model. However, these methods have not achieved ideal results on this dataset. This may be because the similarity between this task and the tasks for which these models were used before is not high, and the audio noise of this dataset is relatively large. After adding the LSTM part in this experiment, the test results were effectively improved. This is because the structural characteristics of LSTM allow the audio features that are mutually related to be retained, and after considering the sequence, the model's classification of audio becomes more accurate.

5.3. Results for dual-modality model and overall discussion

The results for all models are shown in Table 6, the performance gap between the best and worst models for each single modality and between the bimodal and best single modality models results are shown in Table 7, and the confusion matrix results are shown in Figure 6:

Table 6. Testing results of all models.

Model	Unweighted Accuracy	Weighted Accuracy	Unweighted F1	Weighted F1
CLIP-LSTM	73.13%	64.46%	62.34%	72.02%
ResNet-LSTM	65.67%	66.67%	57.39%	65.71%
2DCNN-LSTM	70.15%	76.67%	72.90%	69.57%
3DCNN	72.16%	72.33%	75.10%	72.15%
Mel-spectrogram-LSTM	52.74%	32.32%	30.57%	51.86%
Opensmile-LSTM	59.70%	28.15%	25.66%	52.33%
Wav2vec-LSTM	49.25%	23.15%	22.18%	47.19%
Mel-spectrogram-3DCNN	77.15%	75.72%	76.36%	77.90%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 January 2023 doi:10.20944/preprints202301.0551.v1

Table 7. Results of performance gap between the best and worst models for each single modality and between the bimodal and best single modality models.

Model	Unweighted Accuracy ↑	Weighted Accuracy ↑	Unweighted F1 ↑	Weighted F1 ↑
Δ3DCNN (Compared with ResNet-LSTM)	6.49%	5.66%	17.71%	6.44%
ΔMel-spectrogram-LSTM (Compared with Wav2vec-LSTM)	3.49%	9.17%	8.39%	4.67%
ΔMel-spectrogram-3DCNN (Compared with 3DCNN)	4.99%	3.39%	1.26%	5.75%
ΔMel-spectrogram-3DCNN (Compared with Mel-spectrogram-LSTM)	24.41%	43.40%	45.79%	26.04%

As shown in Table 6 and 7, in the visual single modality, the 3DCNN model has a higher F1 score than other models, with the highest improvement reaching 17.71%; in the audio single modality, the Mel-spectrogram-LSTM model has a higher F1 score than other models, with the highest improvement reaching 8.39%. This represents that these two single modality models have a higher recall rate and precision rate in similar atmosphere classification tasks. CLIP, Resnet, Opensmile, and Wav2vec are feature extraction models that use pre-trained parameters, and their performance is relatively low. This may be due to the fact that the task of this dataset is not very similar to most mainstream tasks, and it is not suitable to use pre-trained model weights for mainstream tasks when extracting features. Therefore, they need to be trained separately. The four indicators of the dual-modality model are the highest, and the performance of the audio single-modality model has been greatly improved, which indicates that inputting the audio and visual modalities at the same time is more beneficial for completing similar atmosphere classification tasks. Such classification tasks should be dominated by visual modalities and supported by audio modalities.

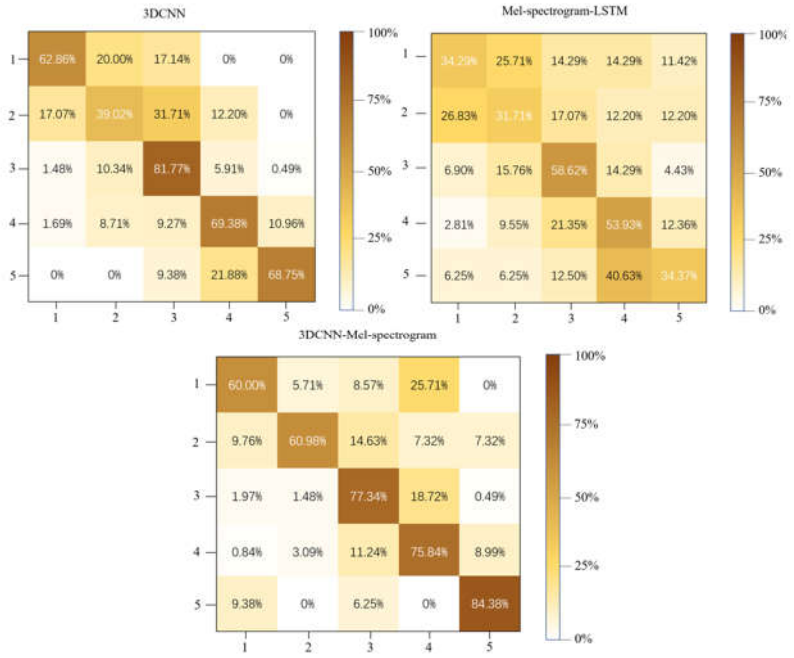


Figure 7. The confusion matrix of the best image model, best audio model, and best dual-mode model.

As can be seen from the results of the confusion matrix (Figure 7) and performance comparison table (table 6,7), the dual-mode model has a higher accuracy on this dataset, and is significantly better than the two single-mode models. This is because the dual-mode

 doi:10.20944/preprints202301.0551.v1

model uses the two-stream method, using one model to extract image features that contain time and using another model to extract audio features that contain time, and then fusing the two features together for classification. Compared to the single-mode model, the dual-mode model can consider more features for fine-grained classification of the atmosphere. The data in Figure 6 also proves the feasibility of using this dataset to train a dual-mode model.

The optimal results of the unimodal and bimodal models tested on the data set are shown in Table 8:

Table 8. Results of the unimodal and bimodal models tested on the dataset.

Task	Unweighted Accuracy	Weighted Accuracy	Unweighted F1	Weighted F1
A	59.70%	32.32%	30.57%	52.33%
V	73.13%	76.67%	75.10%	72.15%
M	77.15%	75.72%	76.36%	77.90%

As shown in the Table 8, the visual mode and audio-visual dual-mode model results are more accurate, and the accuracy of the model that only uses the audio single-mode is lower.

5. Conclusions

In tests on this dataset, models with 3DCNN or derivative structures performed better for image feature analysis. For audio feature analysis, using LSTM structures for time series analysis on Mel spectrograms significantly improved accuracy. Compared to using feature extraction models with pre-trained weights, this dataset's classification task requires more weight training for the feature extraction part of the model from scratch. Results from the comparison experiments showed that this dataset is feasible and effective for emotional calculation of classroom atmosphere for both single-mode and dual-mode models.

In the future, we will further explore tri-modality models' performance on CH-CC dataset through additional text annotation. Additionally, this dataset will be adding more samples and label content to meet the further research requirements of intelligent education in the field of guiding teachers to adjust classroom atmosphere and improve students' attention in class.

References

1. Timms, M.J. Letting Artificial Intelligence in Education Out of the Box: Educational Cobots and Smart Classrooms. *Int J Artif Intell Educ* **26**, 2016, pp. 701–712.
2. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631-1642.
3. Dan-Glauser, E.S., Scherer, K.R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behav Res* **43**, 2011, pp.468.
4. F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions", 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), in Proc. of IEEE Face & Gestures 2013, Shanghai (China), 2013, pp. 22-26.
5. Hasan, M. K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M. I., Morency, L. P., & Hoque, M. E. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor., 2019.
6. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236-2246.
7. Sun, B., Zhao, K., Xiao, Y., He, J., Yu, L., Wu, Y., & Yan, H. BNU-LCSAD: a video database for classroom student action recognition. In *Optoelectronic Imaging and Multimedia Technology VI*, 2019, pp. 417-424.
8. Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. EMOTIC: Emotions in Context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 61-69.
9. Gang, Z., Wenjuan, Z., Biling, H., Jie, C., Hui, H., & Qing, X. A simple teacher behavior recognition method for massive teaching videos based on teacher set. *Applied Intelligence*, *51*, 2021, pp. 8828-8849.

10. Majd M, Safabakhsh R A Motion-aware convLSTM Network for Action Recognition. Appl Intell 49(7), 2019, pp. 2515– 2521.
11. Liu Z, Li Z, Wang R, Zong M, Ji W Spatiotemporal Saliency-based Multi-stream Networks with Attention-aware LSTM for Action Recognition. Neural Computing & Application (11), 2020
12. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M Learning Spatiotemporal Features with 3D Convolutional Networks. In: IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
13. Diba A, Fayyaz M, Sharma V et al Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classiffcation. arXiv:1711.08200
14. Zong M, Wang R, Chen Z, et al. Multi-cue based 3D Residual Network for Action Recognition. Neural Comput Appl, 2020, pp. 1–15.
15. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D et al Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. arXiv:1608.00859
16. Zhou B, Andonian A, Oliva A, Torralba A Temporal Relational Reasoning in Videos. arXiv:1711.08496
17. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8, 2015, pp. 18-25.
18. Eyben, F., Wöllmer, M., & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* 2010, pp. 1459-1462.
19. Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
20. Gautam, A., Hazra, S., Verma, R., Maji, P., & Balabantaray, B. K. ED-NET: Educational Teaching Video Classification Network.
21. S. Pouyanfar, S. -C. Chen and M. -L. Shyu, "An efficient deep residual-inception network for multimedia classification," 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 2017, pp. 373-378.
22. Okazaki, S., Kong, Q., & Yoshinaga, T. A Multi-Modal Fusion Approach for Audio-Visual Scene Classification Enhanced by CLIP Variants. In DCASE, 2021, pp. 95-99.
23. Lezhenin, I., Bogach, N., & Pyshkin, E. Urban sound classification using long short-term memory neural network. In *2019 federated conference on computer science and information systems (FedCSIS)*, 2019, pp. 57-60.
24. Meng, L., Liu, Y., Liu, X., Huang, Z., Jiang, W., Zhang, T., ... & Liu, C. (2022). Multi-modal Emotion Estimation for in-the-wild Videos. arXiv preprint arXiv:2203.13032.