Computing and Informatics, Vol. 44, 2025, 1123–1143, doi: 10.31577/cai_2025_5_1123

MULTI-MODAL EMOTION RECOGNITION USING SITUATION-BASED VIDEO CONTEXT EMOTION DATASET

Guiping Lu, Honghua Liu, Kejun Wang,
Weidong Hu, Wenliang Peng

Key Laboratory of Short-Range Radio Equipment
Testing and Evaluation
Ministry of Industry and Information Technology
and Terahertz Science Application Center (TSAC)
Beijing Institute of Technology
Zhuhai, Guangdong, China
e-mail: 344088386@qq.com, Liu_544hz@163.com, heukejun@126.com,
hoowind@bit.edu.cn, 01059@bitzh.edu.cn

Tao Yang

School of Intelligent Science and Engineering
Harbin Engineering University
Harbin, Heilongjiang, China
e-mail: 2417882631@qq.com

Shan Lu

BMW Brilliance Automotive Ltd.
Shenyang, Liaoning, China
e-mail: lu34125shan@aliyun.com

Abstract. Current multi-modal emotion recognition techniques primarily use modalities such as expression, speech, text, and gesture. Existing methods only capture emotion from the current moment in a picture or video, neglecting the influ-

ence of time and past experiences on human emotion. Expanding the temporal scope can provide more clues for emotion recognition. To address this, constructed the Situation-Based Video Context Emotion Datasets (SVCEmotion) dataset in video form. Experiments show that both VGGish and BERTbase achieve good results on SVCEmotion. Comparison with other audio emotion recognition methods proves that VGGish is more suitable for audio emotion feature extraction on the dataset constructed in this paper. Comparison experiments with textual descriptions demonstrate that the contextual descriptions introduced in the SVCEmotion dataset for the emotion recognition task under wide time range can provide clues for emotion recognition, and that the combination with factual descriptions can substantially improve the emotion recognition effect.

Keywords: Multi-modal fusion, emotion recognition, transfer learning, dataset, deep learning

Mathematics Subject Classification 2010: 68-T45

1 INTRODUCTION

Most early research on emotion recognition has focused on the study of a single modality. However, unimodal approaches suffer from limited information and are easily affected by external factors. For example, facial expressions may be partially occluded, and speech signals can be distorted by noise, both of which can degrade recognition performance.

D'mello and Kory [1] in 2015 compared uni-modal and multi-modal performance on multiple databases by using a statistical approach, and experimentally demonstrated that multimodal expression recognition methods perform better than unimodal. After this, researchers began to conduct studies on multi-modal feature fusion recognition techniques. With the rise of deep learning and the enhancement of computer performance, sentiment analysis has gradually shifted from laboratory environments to real-life natural scenarios, making sentiment recognition techniques face more challenges. Meanwhile, since human emotions have a causal relationship with external factors, primarily specific experiences, expanding the temporal scope of emotion recognition research can enable more accurate emotion analysis. In addition to considering information from the present moment, it is also essential to account for past events experienced by the subject that are relevant to their current emotional state. This historical context serves as a priori information in sentiment analysis and can be treated as a situational context for feature extraction and learning the associations between contextual factors and the subject's emotions. In addition, the emotion recognition method that combines situational information and facial expression has received attention from researchers in the recent years, and researchers believe that the environment where the target subject is located has

a certain influence on his or her emotional state, so situational information can be used as auxiliary information to help the emotion recognition system analyse the subject's emotion more accurately.

In summary, the article is organized as follows: Construing a multi-modal audio-video contextual emotion dataset, including video, audio and text; Using pre-trained model VGGish [2] and BERTbase [3] to extracted sentiment features for both audio and text modalities. It is demonstrated through experiments on audio and text description data from the SVCEmotion dataset that both VGGish and BERTbase pre-trained models achieve good results. The comparative experiments on textual descriptions demonstrate that the situational descriptions introduced in the SVCEmotion dataset for the emotion recognition task under the wide time range can provide clues for emotion recognition, and that the recognition method combined with factual descriptions can significantly improve the emotion recognition effect, which establishes a certain foundation for carrying out multi-modal emotion recognition.

2 SITUATION-BASED VIDEO CONTEXT EMOTION DATASETS CONSTRUCTION

Human emotions are complex and varied. The emotions at a given moment may be a combination of several emotions. The emotions in most situations in daily life cannot be described by only one emotion. In order to be closer to the psychological state of real-life characters, the dataset constructed in this paper chooses multi-label for labelling emotions.

2.1 Emotion Definition

Humans have complex and diverse emotional states. As Clavel et al. [4] point out, fuzzy emotions (non-basic emotions) are common in daily life, so how to describe and define emotions is a complex problem.

Nowadays, the six basic emotions defined by Ekman and Friesen remain the mainstream categories studied in most existing datasets. However, some datasets have been expanded to include additional emotion categories [5].

IEMOCAP [6], collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California (USC), contains a total of 9 emotion labels of anger, sadness, happiness, disgust, fear, surprise, frustration, excitement, neutral state.

EMOTIC [7] is a context-based emotion database constructed by Kosti et al. It references other datasets and contains 26 discrete emotion categories, where each sample may have multiple labels assigned.

In this paper, the emotion categories defined in EMOTIC (26 categories) are used as the benchmark, supplemented with two additional emotion categories: "frustration" from IEMOCAP and "regret", which are commonly found in data labeling.

This results in a total of 28 emotion categories in SVCEmotion. The full list of emotion categories and their definitions is provided in Table 1.

Emotional Category	Definition
Affection	Fond Feelings; Love; Tenderness
Anger	Furious; Resentful
Annoyance	Impatient; Frustrated
Anticipation	State of Looking Forward
Aversion	Dislike; Repulsion
Confidence	Feeling of Being Certain
Depressed	Upset; Downhearted; Disappointed
Disapproval	Contempt
Disconnection	Distracted
Disquiet	Nervous; Worried
Doubt/Confusion	Distrustful
Embarrassment	Ashamed; Guilty
Engagement	Paying Attention to Something;
Esteem	Respect; Admiration
Excitement	Feeling Enthusiasm
Fatigue	Weariness; Tiredness; Sleepy
Fear	Scared; Afraid
Happy	Feeling Delighted
Pain	Discomfort
Peace	Placid
Pleasure	Feeling of Delight in the Senses
Regretful	Sense of Remorse
Sadness	Sorrow; Disappointed
Sensitivity	Vulnerable
Suffering	Distressed; Anguished
Surprise	Shocked
Sympathy	Compassionate
Yearning	Jealous; Envious

Table 1. Emotion categories and definitions

2.2 Data Annotation

With the aim of conducting the research of sentiment analysis over a wide time category and facilitating the capture of scene information, when collecting data, it is important to ensure the coherence of the data in chronological order.

In order to simulate the real environment, the data source selected cropped clips from Chinese and English films and television dramas. When choosing the themes, film and television dramas that do not differ much from the real aspects are selected, avoiding dramas such as costume dramas and fantasy dramas that have heavy filters

and do not conform to realistic tones. In addition, film and television dramas that use dialects are avoided in the selection process.

To eliminate irrelevant frames from a single clip, the target clip was intercepted at level of individual video frames using Adobe Premiere Pro1 after obtaining the original video. Considering that human emotion will achieve the peaks within an average of 10 s, and in order to capture additional information other than that of the target person, the duration of a single clip was kept under 20 s. Aiming at the problems of incoherent content and unlabelled content (e.g. characters, scenes, etc.) caused by camera switching, using the face detection algorithm to extract the character's facial regions, and the scenes and people related to the labelled content are retained as much as possible in the interception of the clips.

The annotation task for the SVCEmotion dataset is more complex than that of mainstream datasets such as IEMOCAP and EMOTIC. In addition to the labelling of emotion categories, some textual aspects of description are necessary. A user interface is designed to facilitate the annotation process, as shown in Figure 1. The left side of the figure primarily contains a text annotation section, including the annotated person's spoken lines from the current video clip (or empty if no speech is present), factual descriptions, and situational descriptions.

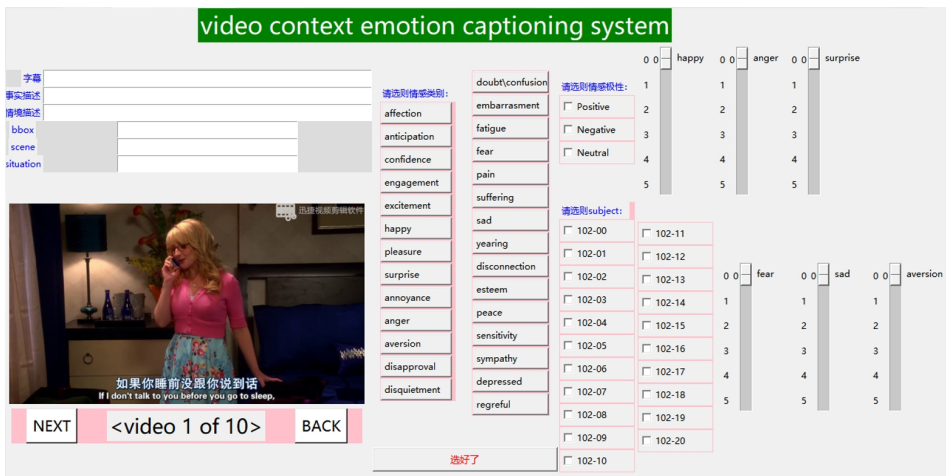


Figure 1. Display of dataset labelled content

The left side of the Figure 1 is mainly a text annotation section, including the annotated person's lines in the current video clip (or null if there are no lines), factual descriptions, and situational descriptions. Where a factual description is an objective description of what the annotated person is doing in the video, without any affective tendencies or a priori information. For example, "A woman is talking on the phone in her room" in Figure 1.

Contextual description needs to describe the events that the annotated person has experienced before the moment shown in the video, explaining the causal relationship between the previously experienced events and the current moment. The corresponding contextual description in the figure is "A woman has just arrived at the hotel where she is travelling, and her husband calls to greet her." Refining event antecedents and consequences through manual annotation is currently the best way to ensure that accurate understanding and summarisation of video content can be achieved.

"Bbox" refers to the number assigned to the target person in the video that needs to be labeled. Since there may be two or more targets in the video, this number is used to differentiate them. "Scene" refers to the scene in the video, such as home, classroom, bedroom, etc. "Situation" is the objective event that the target person is at in the video, such as talking, walking, and so on. In the annotating section of the sentiment categories, this dataset contains three types of sentiment labelled content. The annotation process involves the subdivision of the 28 specific emotion categories, with the annotator ticking the category that matches the current target person's emotion. In addition, emotional polarity options are provided, categorized as positive, negative, and neutral, with only one selection allowed. On the far right of Figure 1 are the emotional degrees of the six basic emotions. If the target person's emotion aligns with one or more of these six basic emotions, it is annotated according to the intensity of the target person's emotion, ranging from 1 to 5, with increasing intensity. If none of these six emotions are present, 0 is selected.

Except for the above-mentioned labelled content, each film and TV drama also track the annotated characters and the interpersonal relationships between all characters, and ensuring that each target character has an ID. The annotation process of this dataset requires the annotator to have a complete understanding of the collected film and TV drama. To ensure that the contextual information is understood correctly, the annotation process did not choose the conventional methods of crowdsourcing, and the same person was responsible for the collection and annotation of each film and TV drama. Totally three people were completing the collection and annotation of the SVCEmotion dataset.

Non-target characters are not deliberately avoided when capturing video for the SVCEmotion dataset. As a result, the number of characters in the footage is often two or more. In order to track the target character in each frame, a multi-person target tracking method was applied, using a combination of YOLOv5 and DeepSORT [8]. YOLOv5 is a classic target detection algorithm, and the DeepSORT is a classic multi-target tracking algorithm, which determines the position of the target in each frame of the image [9]. Using this algorithm, it is possible to identify and track all people appearing in the video. Meanwhile, an identifier will be assigned to each person's detection frame, and it will be able to obtain the coordinates, width and height of each detection frame in the image.

However, the algorithm will re-assign an identifier to the character in the shot when it encounters a shot switch. Besides, if the target character changes from a front face to a back or side face, the algorithm will not be able to detect whether

it is the same person, at which time the number of the target character will change, which is not conducive to the subsequent processing of the data, as shown in Figure 2, where the identifier of the two actors has changed after the shot switch.

Figure 2. Camera switching causes the target's number changed

Therefore, after obtaining all the detection results, manual screening is performed to standardise the target character identifier for each frame by manual modification; non-target characters were not included in the study, so that no correction was made.

Finally, the SVCEmotion dataset is derived from 16 Chinese and English films and TV dramas, consisting of a total of 1 551 videos. Each video may feature multiple target characters, with a total of 2004 annotations in the emotion category and 312 labeled individuals. The duration of the videos typically ranges from 10 to 20 seconds, and the process of facial expression changes of the target characters is recorded. Figure 3 shows the frequency (counts) of each emotion categories in the SVCEmotion dataset.

3 TRANSFER LEARNING

Deep learning models need to be trained on a large number of data samples in order to achieve a more desirable feature representation capability. However, due to the high cost of collection and labelling time of the SVCEmotion dataset, current dataset is small in size, it is insufficient to construct an effective deep-learning model. Transfer learning [10] can make use of the knowledge learnt by the pre-trained model in the source domain to assist in achieving better recognition results on the task in the target domain. The structure of transfer learning is shown in Figure 4.

3.1 VGGish-Based Audio Emotion Recognition

The VGGish network is a pre-trained model trained on AudioSet [11], a large-scale dataset of manually annotated audio events published by Google. This dataset consists of more than 2 million 10-second audio clips from YouTube, annotated with a total of 623 ontologies of audio event classes, covering a broad range of real-world

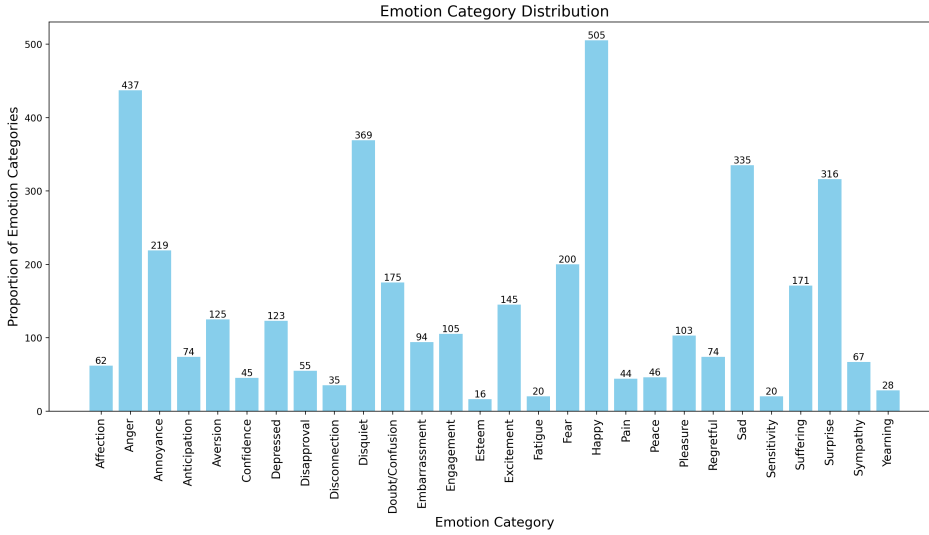


Figure 3. Frequency (counts) of each emotion categories in SVCEmotion dataset

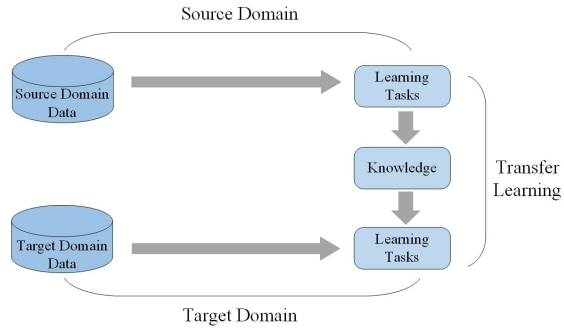


Figure 4. The structure of transfer learning

audio events and providing a more diverse and extensive collection compared to many other publicly available audio datasets. The VGGish model, by pre-training on this dataset, has learnt the vast majority of the various real-life sound categories. We use this model as a feature extraction network for audio signals in the SVCEmotion dataset to obtain high-level audio representations for emotion recognition. The specific structure of the network is shown in Figure 6, which consists of a total of four modules consisting of four convolutional and maximal pooling layers and three fully-connected layers.

The input to VGGish is the Mel-spectrogram of the audio, so first step is to convert the audio signal. The FFmpeg tool was invoked to extract the audio from

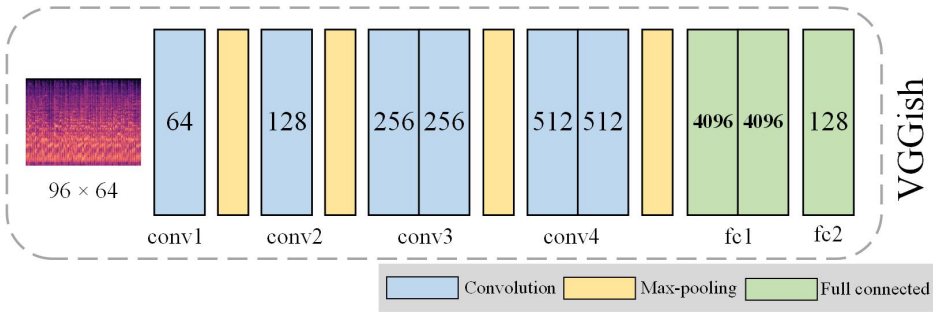


Figure 5. The structure of VGGish

the video. Then resample it to 16 kHz mono audio. Next, a Hann window with a window length of 25 ms is used to intercept the audio clip, and a short time Fourier transform is implemented on the audio clip with a frame shift of 10 ms to capture the frequency domain characteristics of the speech signal, which is converted from a time domain signal to a frequency domain signal while ensuring that the time-dependent information is not lost, and the final result is a spectrogram.

By using a Mel-scale filter bank, we first compute the Mel spectrogram and then apply the logarithmic transformation to enhance perceptual features. The transformation follows the standard formula:

$$\log(\text{Mel} - \text{spectrogram} + \varepsilon),$$

where ε is a small constant to avoid numerical instability. Each frame is 10 ms long and contains 64 Mel bands. Then the Mel spectrograms of each frame were combined without overlap in groups of 0.96 s. Thus, the size of each group of input Mel spectrograms is 96×64. The transformed Mel sound spectrograms are used as inputs to VGGish, and each set of inputs is subjected to feature extraction to obtain a 128-dimensional feature representation. Since the duration of each sample is greater than 0.96 s, multiple feature representations can be obtained from each sample after feature extraction. In the training process, a set of feature representations of the audio samples are randomly selected each time and fed into the two fully connected layers to obtain a 28-dimensional classification representation. In the validation process, the mean value of all the feature representations of each sample is taken as the feature representation of the video level and fed into the classifier to obtain the recognition results.

3.2 BERT-Based Text Emotion Recognition

The BERT model is trained by completing two unsupervised pre-training tasks using two large corpora, Books Corpus and English Wikipedia, in order to obtain a strong semantic representation of the text. The first task is to train the language model

in a masked way, where the tokens of the input sentence are randomly masked in a certain proportion, then the masked tokens are predicted to obtain a deep bi-directional linguistic feature representation.

The second task is to determine whether the two input texts are consecutive or not. Then to train the bi-directional language model's ability to learn about associations between longer text sequences. The advantage of this model is that the model architecture is unified when dealing with different downstream tasks.

In this paper, we choose the BERTbase model for text-based emotion recognition. The network has 12 layers of Encoder, the size of the hidden layer is 768, and the number of multi-head self-attention modules in each layer is 12, and the overall size is larger than that of the Transformer model, and the specific structure is shown in Figure 5.

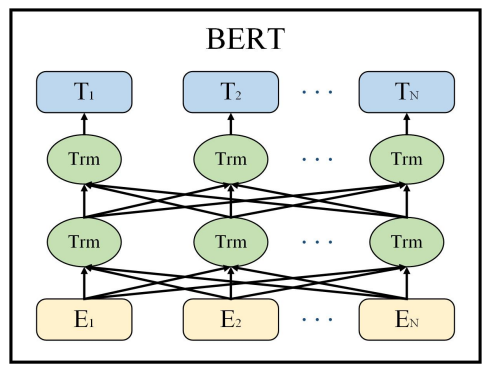


Figure 6. The structure of BERTbase. T_i represents the input token embeddings, T_{rm} denotes the transformer layers, and E_i refers to the final encoded representations of the input tokens.

The input of this model consists of three parts, firstly, the input sentences are divided into words (tokenization), and each word is converted into the corresponding 768-dimensional word embedding ID, that is the token embedding for each word. Then two special tokens [CLS] and [SEP] are inserted at the beginning and end of each input sequence, where [CLS] is the token used for the classification task, and the output state corresponding to its position integrates the semantic information of the whole sentence, and the [SEP] token serves to differentiate the input sentence pairs. Since the Transformer's self-attention mechanism will ignore the position of a word in a sentence, position embedding is taken for each word with the same dimensions as the word vector. Finally, segment embedding is added to each token to distinguish between two sentences in a sentence pair; if only one sentence is input, then all the tokens are embedded with the [SEP] token. If only one sentence is input, the segment embedding value is the same for all tokens. The three-part embeddings of each token are summed up by position, that will be the input vector of BERT. In order to unify the length of the input sequence, sentences with shorter length

will fill the remaining token positions with zeros, and the maximum length of the sequence is set to 128 in this paper.

For each input sequence, the output dimension after feature extraction is 128×768. The output features corresponding to the classification labels [CLS] are mapped to the output space through a fully connected layer to obtain the classification result for each text. In this section, feature extraction and sentiment classification are performed on factual descriptions, contextual descriptions, and the splicing of the two, respectively, and the classification results obtained are denoted as P_t, P_c and P_{tc}.

3.3 Loss Function

Since sentiment recognition in audio modality, text modality and visual modality are all multi-label classification problems, the loss function is chosen to be BCEWithLogitsLoss [12], which combines the operations of the Sigmoid activation function and BCELoss in such a way that there is no need to add the operation of the Sigmoid function at the output layer of the model.

It is made to perform more stably numerically by using the log-sum-exp trick [13]. The expression of this loss function for multi-label classification, considering all sentiment categories in the dataset, is as follows:

$$l_c(x, y) = \frac{1}{N} \sum_{n=1}^{N} l_{n,c}, \tag{1}$$

$$l_{n,c} = -w_{n,c}[p_c y_{n,c} \cdot \log \sigma(x_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - \sigma(x_{n,c}))], \tag{2}$$

where $x_{n,c}$ denotes the predicted value of the n^{th} sample in the c sentiment category, $y_{n,c}$ denotes the true labels of the n^{th} sample in the c sentiment category, N denotes the number of batch size, and σ denotes the Sigmoid activation function. $w_{n,c}$ is the customised weight corresponding to the loss of different categories in each batch, which needs to be guaranteed to have a dimension of $N \times C$, and p_c is the weight corresponding to the positive samples in each category, which is controlled by the ratio of the negative samples to the positive samples. $w_{n,c}$ and p_c are both optional parameters used to alleviate the problems of imbalance in the categories of the training samples and the distribution of positive and negative samples within the categories, respectively.

4 RESULT AND DISCUSSION

All experiments in this paper use the SVCEmotion dataset, using data from both audio and text modalities.

The training set contains 1218 videos with a total of 1589 annotations, and the validation set contains 333 videos with a total of 415 annotations.

The audio modal uses the pre-training parameters of VGGish as the initial parameters for training, the batch size is set to 6, a total of 300 epochs are trained, and the initial learning rate is set to 0.0001.

The optimiser uses an Adaptive moment estimation optimiser (Adam) with the hyper parameter momentum set to 0.9. Adam has the advantage of being easy to implement, computationally efficient and capable of dynamically adjusting the learning rate.

The weight decay rate was set to 1e−5, and the learning rate was adjusted downward by one-tenth at the 150^{th} epoch of training, and the learning rate was decreased by another one-tenth at the 230^{th} epoch of training.

The factual description of the textual modality, the contextual description, and the combination of the two use the BERTbase pre-trained model for training, the batch size is set to 128, 100 epochs are trained, the initial learning rate is set to 2e−5, and the optimiser is also chosen to be Adam, with a weight decay rate of 0.01.

Table 2 shows the parameter settings for the two modes.

Parametric	Audio	Text
Batch Size	6	128
Epoch	300	100
Learning Rate	0.0001	0.00002
Weight Decay	0.00001	0.01
Momentum	0.9	0.9

Table 2. Parameter settings

All experiments in this study were conducted using the SVCEmotion dataset. In this study, both audio and text modalities were utilized. Considering the relatively small scale of the dataset, it was divided into a training set and a validation set. The dataset contains clips from 16 movies and TV series, with 11 of them assigned to the training set and the remaining to the validation set. The training set consists of 1 218 videos with a total of 1 589 annotations, while the validation set contains 333 videos with 415 annotations. Due to the differences in the distribution of emotion categories across different genres of movies and TV shows, efforts were made to ensure that the emotion category distributions in the training and validation sets were as consistent as possible. The specific distributions of emotion categories are illustrated in Figure 7.

4.1 Audio Modal Experiment Results and Analysis

To address the issue of the small dataset size and its inadequacy for effectively training deep neural networks, the pre-trained VGGish model was selected for feature extraction and emotion recognition on audio. VGGish, pre-trained on the large-scale AudioSet dataset, has learned a wide range of real-world audio events, thus offering better feature extraction and providing a strong starting point for our dataset.

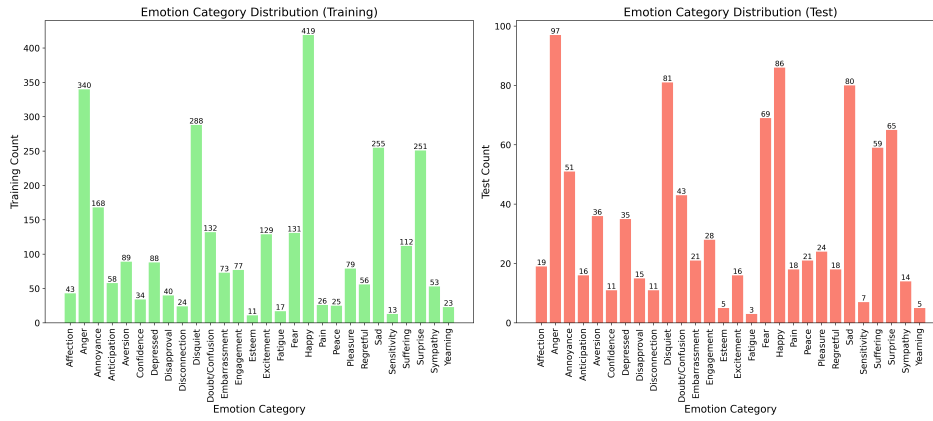


Figure 7. Distribution of emotion categories in the training and validation sets

Table 3 presents the results of audio-modal emotion recognition under two different learning rates (0.001 and 0.0001). AP refers to Average Precision, which measures the model's ability to correctly identify positive samples, and RA refers to Recognition Accuracy, which indicates the proportion of correctly classified samples. The values in the table represent percentages (%), showing the model's performance for each emotion category under different learning rates.

Average Precision (AP): AP is computed as the area under the Precision-Recall (PR) curve for each emotion category. It is obtained using the formula:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n,$$

where P_n and R_n represent the precision and recall at the n^{th} threshold, respectively.

Recognition Accuracy (RA): RA is defined as the proportion of correctly classified samples across all test instances:

$$RA = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives, respectively.

As shown in Table 3, the recognition results were comparable, with a slight advantage (about 0.5) at a learning rate of 0.0001, and most evaluation metrics for various emotion categories outperformed those with a 0.001 learning rate. This effectiveness is attributed to VGGish's comprehensive training on AudioSet's 623 audio event categories, which encompass most scenarios in our dataset, allowing efficient

Emotion Label	Learning Rate = 0.001		Learning Rate = 0.0001	
	AP %	RA %	AP %	RA %
Affection	8.21	55.58	10.65	58.65
Anger	50.08	65.71	51.90	66.70
Annoyance	22.40	54.12	27.95	59.29
Anticipation	**5.69**	**47.50**	4.88	39.55
Aversion	15.78	66.19	18.77	68.45
Confidence	**9.25**	**69.66**	8.99	65.84
Depressed	**30.58**	**76.15**	27.55	75.60
Disapproval	16.34	71.51	25.45	72.70
Disconnection	7.29	**64.92**	5.83	64.38
Disquiet	28.34	**59.39**	30.11	57.49
Doubt/Confusion	20.57	63.11	24.50	66.60
Embarrassment	**6.63**	**48.85**	6.36	47.91
Engagement	7.75	45.14	8.20	48.18
Esteem	2.91	67.59	3.23	69.09
Excitement	**19.21**	**63.68**	17.83	63.62
Fatigue	2.63	75.70	2.60	75.72
Fear	**45.98**	73.82	45.19	74.58
Happy	**35.4**	60.68	32.93	59.96
Pain	26.68	76.78	32.98	78.74
Peace	14.10	66.47	14.59	68.71
Pleasure	**23.14**	**75.35**	17.10	68.95
Regretful	8.49	56.52	8.71	56.78
Sadness	**47.61**	75.52	47.02	76.14
Sensitivity	**1.75**	**36.91**	1.62	30.01
Suffering	**29.94**	71.14	28.67	72.06
Surprise	23.07	51.50	28.30	55.83
Sympathy	8.68	62.65	8.69	67.72
Yearning	**5.04**	73.42	5.03	73.66
AVG	18.70	63.41	19.49	63.67
Mean	41.06		**41.58**	

Table 3. Audio modal sentiment recognition results

extraction of emotional features after fine-tuning with SVCEmotion. Additionally, Table 4 presents a comparative evaluation of VGGish against other methods on our dataset. The evaluation metrics include mean Average Precision (mAP), mean Recognition Accuracy (mRA), and M, which represents the mean score combining these metrics for overall performance assessment. Higher values indicate better performance. The results demonstrate that VGGish achieves the highest mAP, mRA, and Mean scores, highlighting its effectiveness in feature extraction for emotion recognition.

Network Model	mAP	mRA	M
3D-ACRNN	18.95	62.79	40.87
Light-SERNet	19.06	63.11	41.09
VGGish	19.49	63.67	41.58

Table 4. Experimental comparison of VGGish with other methods on SVCEmotion

4.2 Text Modal Experiment Results and Analysis

Two parts of textual modal data, factual and situational descriptions from the SVCEmotion dataset were used to extract and recognise the emotion feature landscape features of the two textual descriptions respectively using the BERTbase pre-trained model. In addition, the two textual descriptions from each sample were also used for emotion recognition using this pre-trained model after a direct simple collocation, and the results of these three emotion recognitions were compared.

BERTbase obtains powerful semantic representation capabilities by performing two pre-training tasks on a large corpus. The pre-training task using randomly masked input tokens equips the model with deep bi-directional representations, enabling richer semantic information to be obtained jointly with left and right contextual contexts. Thus, the model is able to effectively learn the associations between text descriptions and emotions and achieve emotion classification. The parameters and details of the three experiments of the text modal were conducted under the same conditions, and Table 5 demonstrates the specific results of the experiments.

Firstly, comparing the factual description and contextual description, both textual descriptions achieve the emotion classification function through BERTbase for feature extraction. The overall effect of emotion recognition of the contextual context description is better than that of the factual description, with a higher evaluation index of 7.77. This is due to the fact that the experience of the target person is described in the contextual description, and the current possible emotional state can be inferred based on the experienced events.

For example, if a girl is being ridiculed by her boyfriend's family, it can be inferred that her mood may be "sad", "puzzled", or "tormented" when she converses with him. Factual descriptions, on the other hand, are only objective descriptions of the events that the target character is currently in or doing, and the text is basically free of emotion, so it is not easy to relate to affective states, for example, "the girl is having a conversation with a boy".

In the emotion categories of "sensitive", the recognition effect of factual description is slightly higher than that of situational recognition, probably because these emotion categories are mainly expressed through the behaviour of the character, and the events done by the character are described in the objective description, while the situational description cannot highlight what the target character is doing at present, so the recognition effect is poorer.

For example, "A woman is sitting at her desk" may presuppose a release from something, but it is easier to infer "calm" from a factual description.

Emotional Labels	Factual Description		Situation Description		Factual + Situation Description	
	AP	RA	AP	RA	AP	RA
Affection	10.6	72.65	23.15	68.49	26.29	87.47
Anger	34.36	64.92	**56.67**	81.99	55.64	82.15
Annoyance	18.4	56.74	16.75	63.55	24.85	71.06
Anticipation	4	45.15	9.61	71.9	17.89	81.56
Aversion	9.8	55.12	**24.18**	**78.5**	23.44	70.62
Confidence	7.82	78.45	5.75	69.24	32.16	87.08
Depressed	13.11	57.23	20.33	**74.59**	28.53	72.61
Disapproval	11.77	75.54	30.41	87.25	32.97	90.55
Disconnection	7.23	73.09	5.25	58.03	12.07	85.1
Disquiet	25.67	58.66	37.34	74.62	39.03	77.21
Doubt/Confusion	13.47	61.01	21.32	66.73	26.59	73.29
Embarrassment	4.91	43.2	8.76	**69.4**	12.65	65.24
Engagement	11.17	65.28	19.57	76.14	26.66	76.14
Esteem	5.16	66.24	6.63	60.15	16.65	81.32
Excitement	5.5	54.82	**10.94**	**76.58**	9.03	60.89
Fatigue	1.02	44.5	**2.71**	**60.76**	1.72	48.87
Fear	47.32	79.93	**48.76**	**82.87**	47.27	81.37
Happy	30.4	62.06	72.37	**88.38**	75.34	86.18
Pain	26.72	80.73	31.34	73.83	35.15	88.18
Peace	**7.6**	57.34	4.87	42.63	7.56	60.66
Pleasure	11.03	68.27	35.23	**84.41**	33.96	81.47
Regretful	4.83	51.97	6.46	58.84	7.87	59.71
Sadness	33.92	60.86	49.92	80.16	53.97	83.54
Sensitivity	**4.37**	**70.83**	1.69	43.73	3.97	70.83
Suffering	31.8	66.92	36.52	79.55	45.1	80.58
Surprise	17.98	55.37	30.98	69.16	37.3	76.78
Sympathy	4.27	55.76	9.44	60.03	10.12	70.64
Yearning	2.2	59.56	1.57	53.85	2.68	61.02
AVG	14.52	62.22	22.44	69.83	**26.66**	75.43
Mean	38.37		46.14		**51.01**	

Table 5. Text modal sentiment recognition results

In order to explore the role of situational context information as an aid to general factual descriptions, two textual descriptions were spliced as input, i.e., "Factual + Situational Description". The recognition effect was significantly improved, by 12.64 compared to the factual description.

The analysis suggests that the information described in the two texts is complementary. By integrating current events with past experiences, the model can more effectively infer the target person's emotions. Emotion categories with previously low recognition rates, such as "boredom", "embarrassment", and "confidence", have all shown varying degrees of improvement. Notably, the RA indicator for "disap-

proval" reached 90.55. However, certain emotion categories experienced a decline in recognition performance. This deterioration may stem from the fact that factual and situational descriptions contribute differently to emotion categorization depending on the context. When combined, these descriptions may weaken the intensity of emotional expression. Therefore, how to more effectively integrate these two types of descriptions remains an important area for future research.

The experiment results indicate that text remains the most informative modality for emotion recognition, while audio serves as a useful but weaker complementary feature. Although Audio + Text fusion improves over Audio alone, it does not outperform Text alone, emphasizing the need for more sophisticated fusion techniques to better integrate multi-modal information.

4.3 Audio + Text Experiment Results and Analysis

To evaluate the effectiveness of multi-modal fusion, we conducted experiments using Audio, Text, and Audio + Text modalities. The text modality includes factual descriptions and situational context descriptions from the SVCEmotion dataset, while the audio modality utilizes extracted features from VGGish.

For the Audio + Text fusion, we employed a decision-level fusion strategy. In this approach, independent classifiers were trained for audio and text separately, and their prediction scores were combined at the decision level using a weighted averaging strategy. This method allows each modality to contribute independently to the final classification decision.

Modalities	mAP	mRA	Mean
Audio	19.49	63.37	41.58
Text	26.66	75.43	51.01
Audio + Text	22.18	69.81	46.00

Table 6. Fusion result

Table 6 presents the results of emotion recognition using Audio, Text, and Audio + Text fusion. From the results, Text modality achieves the highest performance, with mAP = 26.66 and mRA = 75.43, indicating that textual descriptions provide rich emotional cues and strong semantic representations. Audio-only performance is lower (mAP = 19.49, mRA = 63.37), suggesting that audio signals alone may not be sufficient to capture nuanced emotions, especially when background noise or speech variations are present. Audio + Text fusion improves upon the Audio-only results, showing that combining text and audio contributes to a more comprehensive emotion representation (mAP = 22.18, mRA = 69.81). However, the fusion does not surpass the standalone text modality, indicating that text remains the dominant factor in emotion recognition.

5 CONCLUSION

In this study, a context-based video sentiment dataset was constructed. Employing a multi-person target tracking algorithm combining YOLOv5 and DeepSORT, to localize the target subjects in the video. Each subject was labelled with 28 multi-label emotion categories, video timestamps, factual descriptions, and contextual descriptions.

Aiming at the small size of the dataset constructed in this paper, a transfer learning-based approach is used for sentiment recognition of information in audio and text modalities. The pre-training model VGGish is used for feature extraction and sentiment recognition of the audio signals in the dataset of this paper. The BERTbase pre-trained model is used for feature extraction and sentiment recognition for factual descriptions, textual descriptions, and textual descriptions spliced between the two, respectively. The experimental results demonstrate that the pre-trained models used for both modalities can learn sentiment features effectively on the dataset of this paper. By comparing the experimental results of the textual modality, it is verified that the contextual descriptions under the wide time category can provide cues for sentiment analysis to the sentiment recognition model.

Through experiments on audio and text description data from the SVCEmotion dataset, it is demonstrated that both VGGish and BERTbase pre-trained models achieve good results on the dataset used in this paper, and that the parameters learnt by the models in the pre-training process can effectively improve their performance on the target task. By comparing VGGish with other audio emotion recognition methods, it is proven that VGGish is more suitable for audio emotion feature extraction on the dataset constructed in this paper. The comparison experiments with textual descriptions demonstrate that the contextual descriptions introduced in the SVCEmotion dataset for the emotion recognition task under the wide time category can provide clues for emotion recognition, and the combination with factual descriptions can significantly improve the emotion recognition effect.

Acknowledgments

This paper was supported by a grant from the 2023 Zhuhai Basic and Applied Basic Research Projects (No. 2320004002492).

REFERENCES

[1] D'MELLO, S. K.—KORY, J.: A Review and Meta-Analysis of Multimodal Affect Detection Systems. ACM Computing Surveys (CSUR), Vol. 47, 2015, No. 3, Art. No. 43, doi: 10.1145/2682899.

[2] TENSORFLOW AUTHORS: VGGish: A VGG-Like Audio Classification Model. GitHub Repository, 2017, https://github.com/tensorflow/models/tree/master/research/audioset.

[3] Devlin, J.—Chang, M. W.—Lee, K.—Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019). ACL, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[4] Clavel, C.—Vasilescu, I.—Devillers, L.—Richard, G.—Ehrette, T.—Sedogbo, C.: The SAFE Corpus: Illustrating Extreme Emotions in Dynamic Situations. LREC Workshop on Corpora for Research on Emotion and Affect, 2006, pp. 76–79, http://lrec.elra.info/proceedings/lrec2006/workshops/W09/Emotion-proceeding.pdf#page=85.

[5] Ekman, P.—Friesen, W. V.: Constants Across Cultures in the Face and Emotion. Journal of Personality and Social Psychology, Vol. 17, 1971, No. 2, pp. 124–129, doi: 10.1037/h0030377.

[6] Busso, C.—Bulut, M.—Lee, C. C.—Kazemzadeh, A.—Mower, E.—Kim, S.—Chang, J. N.—Lee, S.—Narayanan, S. S.: IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. Language Resources and Evaluation, Vol. 42, 2008, No. 4, pp. 335–359, doi: 10.1007/s10579-008-9076-6.

[7] Kosti, R.—Alvarez, J. M.—Recasens, A.—Lapedriza, A.: Emotion Recognition in Context. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1960–1968, doi: 10.1109/CVPR.2017.212.

[8] Broström, M.: Real-Time Multi-Object Tracker Using YOLOv5 and Deep Sort. 2020, https://github.com/mikel-brostrom/Yolov5_DeepSort_Pytorch.

[9] Pujara, A.—Bhamare, M.: DeepSORT: Real Time & Multi-Object Detection and Tracking with YOLO and TensorFlow. 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), IEEE, 2022, pp. 456–460, doi: 10.1109/ICAISS55157.2022.10011018.

[10] Pan, S. J.—Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, Vol. 22, 2009, No. 10, pp. 1345–1359, doi: 10.1109/TKDE.2009.191.

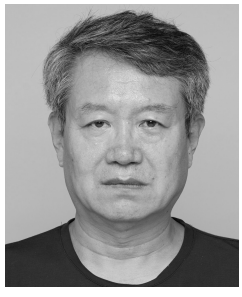
[11] Gemmeke, J. F.—Ellis, D. P. W.—Freedman, D.—Jansen, A.—Lawrence, W.—Moore, R. C.—Plakal, M.—Ritter, M.: Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.

[12] PyTorch Authors: torch.nn.BCEWithLogitsLoss. PyTorch Documentation. 2021, https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html.

[13] Boyd, S. P.—Vandenberghe, L.: Convex Optimization. Cambridge University Press, 2004, doi: 10.1017/CBO9780511804441.

Guiping Lu is Professor and Master's supervisor with a Master's degree. Her main research interests include artificial intelligence theory and applications.

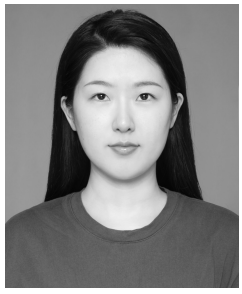
Honghua Liu is a postgraduate student and a member of the China Communications Society. He obtained his Bachelor's degree in automation from the Beijing Institute of Technology, Zhuhai, and his Master's degree in artificial intelligence and digital media from the Beijing Normal – Hong Kong Baptist University United International College. His primary research interests lie in artificial intelligence and multimodal emotion recognition.



Kejun Wang holds his Ph.D. and completed postdoctoral research at the Harbin Engineering University. He is a member of the China Democratic League and currently serves as a council member of the Chinese Association for Artificial Intelligence. His research focuses on biometric recognition and deep learning. He has published over 450 academic papers, received several ministerial-level scientific and technological awards, and holds 48 authorized invention patents. He has been listed among the world's top 2 % of scientists.

Weidong Hu is Professor and a doctoral supervisor at the School of Integrated Circuits and Electronics, Beijing Institute of Technology. He is a recipient of the National Leading Talent award and serves as the Deputy Director of the Beijing Key Laboratory of Millimeter Wave and Terahertz Technology. His research focuses on terahertz space detection and remote sensing technology. He has led major projects under the National Natural Science Foundation of China and the National Civil Aerospace Terahertz Imaging Program, contributing to the development of China's Fengyun meteorological satellites. He has received multiple honors, including the Second Prize of the Beijing Science and Technology Progress Award. As the head delegate for terahertz topics in the Chinese delegation to the International Telecommunication Union (ITU), he has made significant contributions to safeguarding China's spectrum rights.

Wenliang Peng is Associate Professor and Master's supervisor at the Marine Science and Technology Domain, Beijing Institute of Technology, Zhuhai. He also serves as the faculty advisor for the "Yiheng Team", an award-winning student organization recognized as an Excellent Association of Guangdong Province. He was a visiting scholar in the Mechatronics program at Wuhan University of Technology. His research interests include fault diagnosis, machine vision, and cooperative control.



Tao Yang holds her Master's degree in electronic information from the Harbin Engineering University. Her research focuses on multimodal emotion recognition. She is currently working at BYD Auto Industry Co., Ltd., where she is engaged in commercial operations.

Shan Lu graduated from BUAA with a Bachelor's degree in 2002 and a Master's degree in 2005. He is currently employed by BMW Brilliance Automotive Ltd., where he focuses on the development and testing of intelligent driving technologies for the NEV powertrain system and electric drive machine.