Hindawi
Security and Communication Networks
Volume 2021, Article ID 7650483, 12 pages
https://doi.org/10.1155/2021/7650483

Research Article

A Deep Multimodal Model for Predicting Affective Responses Evoked by Movies Based on Shot Segmentation

Chunxiao Wang,[1,2,3] **Jingjing Zhang**,[1,2,3] **Wei Jiang**,[1,2,3] **and Shuang Wang**[1,2,3]

[1] *State Key Laboratory of Media Convergence of Communication, Communication University of China, Beijing 100024, China*
[2] *Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Communication University of China, Beijing 100024, China*
[3] *Beijing Key Laboratory of Modern Entertainment Technology, Communication University of China, Beijing 100024, China*

Correspondence should be addressed to Jingjing Zhang; zjj_cuc@cuc.edu.cn

Received 2 July 2021; Revised 27 August 2021; Accepted 6 September 2021; Published 28 September 2021

Academic Editor: Zhenhua Tan

Copyright © 2021 Chunxiao Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Predicting the emotions evoked in a viewer watching movies is an important research element in affective video content analysis over a wide range of applications. Generally, the emotion of the audience is evoked by the combined effect of the audio-visual messages of the movies. Current research has mainly used rough middle- and high-level audio and visual features to predict experienced emotions, but combining semantic information to refine features to improve emotion prediction results is still not well studied. Therefore, on the premise of considering the time structure and semantic units of a movie, this paper proposes a shot-based audio-visual feature representation method and a long short-term memory (LSTM) model incorporating a temporal attention mechanism for experienced emotion prediction. First, the shot-based audio-visual feature representation defines a method for extracting and combining audio and visual features of each shot clip, and the advanced pretraining models in the related audio-visual tasks are used to extract the audio and visual features with different semantic levels. Then, four components are included in the prediction model: a nonlinear multimodal feature fusion layer, a temporal feature capture layer, a temporal attention layer, and a sentiment prediction layer. This paper focuses on experienced emotion prediction and evaluates the proposed method on the extended COGNIMUSE dataset. The method performs significantly better than the state-of-the-art while significantly reducing the number of calculations, with increases in the Pearson correlation coefficient (PCC) from 0.46 to 0.62 for arousal and from 0.18 to 0.34 for valence in experienced emotion.

1. Introduction

When watching movies, audiences experience a range of emotions over time based on the visual and auditory information they receive. This phenomenon has been a concern of and has been studied by psychologists [1]. As people always evaluate, select, edit, and split movies based on their affective characteristics, recognizing the continuous dynamic emotion evoked by movies can be used to build better multimedia intelligent applications, such as computational affective video-in-video advertising [2] and personalized multimedia content [3], and to create automatic summaries and adaptive playback speed adjustment for long videos, etc. Movies have always been one of the main objects of video sentiment analysis. Unlike short videos on social media, movies are much longer and can induce a rich emotional response from an audience. Additionally, the rich emotional content in movies is inherently multimodal. The complex interplay between audio and video modalities determines the perceived emotion. Therefore, both the complexity of the film data and the dynamic interactivity of the emotional content of the movie make quantifying and automatically predicting the emotions audiences experience a challenging problem.

There are three different "types" of movie emotion. The intended emotion describes the emotional response the movie tries to evoke in audiences, the experienced emotion describes the actual emotions felt by the viewer while watching the movie, and the expected emotion is the experienced emotion expected over a population [4]. Researchers have combined existing models of affective psychology to measure emotional responses, such as dimensional and categorical approaches. The dimensional method has been used in most predictive studies [5–10] because the dimensional method constituted by arousal and valence dimensions can effectively represent the emotions elicited by pictures, videos, sounds, etc. [11]. In particular, Hanjalic and Xu [12] use the arousal and valence dimensions to measure the intensity and type of feeling or emotion that a user experiences while watching a video. Malandrakis et al. [4] proposed a database with continuous valence-arousal scale annotation of intended and experienced emotions over a continuous time, in which valence and arousal are annotated in the range of [−1, 1] by several subjects. The closer the valence is to 1, the more pleasant the emotions that the audience feels, and the closer it is to −1, the more negative the emotions that the audience feels. The closer the arousal is to 1, the more active the audience is, and the closer it is to −1, and the more passive the audience is. When both are closer to 0, the audience feels more neutral. Based on this modeling method, we can measure the emotions that audiences experience as a range of emotions over time by continuous valence-arousal scale annotation.

For predicting the intended or experienced emotion on a continuous valence-arousal scale, Malandrakis et al. [4] proposed a supervised learning method to model the continuous affective response by independently using hidden Markov models in each dimension. Goyal et al. [6] proposed a mixture of experts- (MoE-) based fusion model that dynamically combines information from audio and video modalities for predicting the dynamic emotion evoked in movies. Sivaprasad et al. [7] presented a continuous emotion prediction model for movies based on long short-term memory (LSTM) [13] that models contextual information while using handcrafted audio-video features as input. Joshi et al. [8] proposed a method to model the interdependence of arousal and valence using custom joint loss terms to simultaneously train different LSTM models for arousal and valence prediction. Thao et al. [9] presented a multimodal approach that uses pretrained models to extract visual and audio features to predict the evoked/experienced emotions of videos. Thao et al. [10] presented AttendAffectNet, a multimodal approach based on the self-attention mechanism that can find unique and cross-correspondence contributions of features extracted from multiple modalities. However, these methods usually do not take into account the temporal structure or semantic units of the film, ignore the changes in the audio-visual information in the subclip, and have high computational complexity in the feature extraction process.

Because experienced emotion prediction is more complicated than intended emotion prediction, most researchers have focused on intended emotion prediction. In 2019, Thao et al. [9] used the same features and models to predict intended or experienced emotion, and this was chosen as the baseline of this paper. For predicting the emotion on a continuous valence-arousal scale over time, some methods [6, 7, 9, 10] need to extract image content and optical flow information from each frame of movies, which has high computational complexity. Goyal et al. [6] proposed splitting all movies into nonoverlapping 5-second samples after a frequency response analysis of the intended and experienced emotion labels to find a suitable unit for affective video content analysis. These methods use 5 s subclip-level averaging features without considering the temporal and semantic structure of the movies. There are two temporal structure levels of movies, shots and scenes. As the shot is the minimal visual unit of a movie [14], a multimodal prediction model based on video shot segmentation is proposed in this paper.

This paper focuses on experienced emotion prediction and evaluates the proposed method on the extended COGNIMUSE dataset [4, 15]. First, the movies were divided into short clips by shot boundary detection. Then, the audio and visual features of each shot clip were extracted and combined with shot-based audio-visual feature representation, as we define in Section 2. Finally, the shot-level audio-visual features were fed into our multimodal deep model to predict the evoked emotion. This method obtained a significantly better result than the state-of-the-art while significantly reducing the number of calculations, with increases in Pearson correlation from 0.46 to 0.62 for arousal in experienced emotion and from 0.18 to 0.34 for valence in experienced emotion. Interestingly, we found that the feature combination method based on shot fragments can significantly improve the performance of the method in [9]. The experimental results show that considering the temporal structure and semantic units of movies is of great significance to predicting experiential emotions.

2. Shot Audio-Visual Feature Representation

To consider a movie's temporal structure and semantic unit, the variation in audio and visual information in subclips, and to reduce the high computational complexity of the feature extraction process, a multimodal shot audio-visual feature representation was proposed as follows.

2.1. Video Subset Segmentation Based on Shot Boundary Detection. A shot is a sequence of frames recorded by the same camera and is the minimal visual unit of the movie [16]. The semantic information in one shot clip does not change much, but there are apparent changes in one 5 s clip, as shown in Figure 1. The semantic variation in the 5-s clip increases the difficulty of model learning, so we believe that the segmentation of the video shot subset can obtain higher accuracy in experienced emotion prediction.

Sidiropoulos et al. [17] jointly exploited low-level and high-level features automatically extracted from visual and auditory channels, possessing better shot boundary segmentation, so this method was used to obtain shot subset segmentation of movies in this paper. The average value of the arousal/valence labels of all frames in each shot was used as the emotion label for each shot subclip, similar to the emotion label for the 5 s clips in previous studies [6–10], giving us approximately 5902 samples.

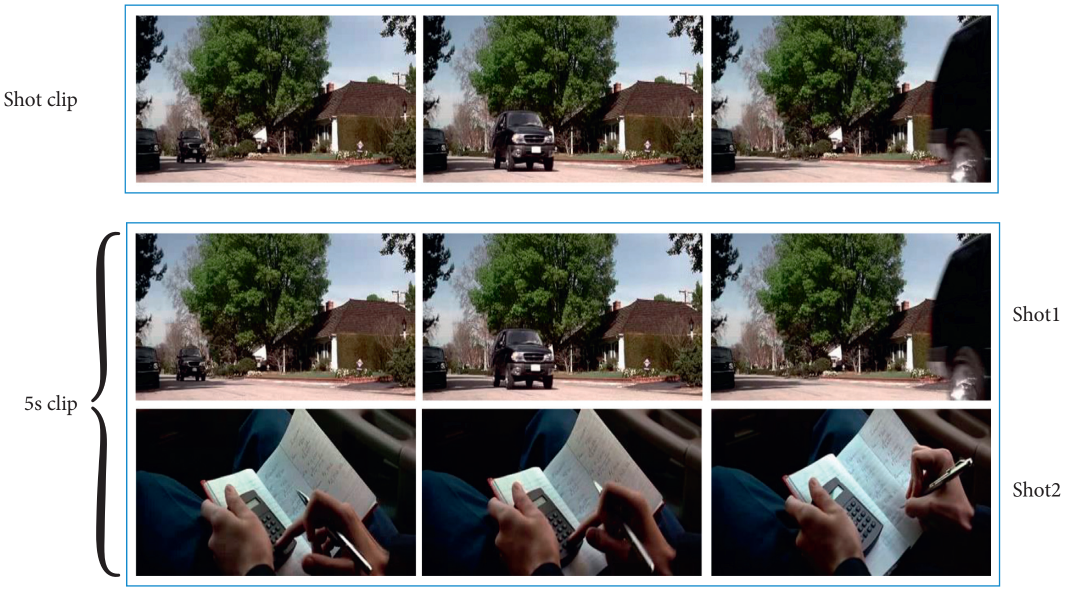

Figure 1: 5 s clip versus shot clip. There is little change in the semantic information in the shot clip compared to the 5 s clip.

2.2. Multimodal Emotion Features Extraction. As is known, the combined effect of the audio-visual messages of movies evokes emotional responses in the audience, and we hypothesize that the comprehensive effect of audio-visual information can be approximated by the interaction of semantic units, so the extracted features need to have the ability to describe the interaction of each semantic unit in the movies. In previous studies [6–10], a global average feature extracted from each 5 s clip was widely used, ignoring the change in the audio-visual information of each subclip. Therefore, to capture the interaction of each semantic unit from the audio and visual information of each shot clip, a targeted method for extracting and combining the audio and visual information features of each shot clip was designed as follows.

First, three keyframes of equal time intervals and one audio file in .wav format were obtained from each shot subset, as shown in Figure 2. These three keyframes corresponded to the beginning, development, and end of the visual information of each shot clip.

To consider semantic units and reduce computational complexity, this method needed to distinguish high-level semantic elements and to replace the extraction of the optical flow information. Thus, inspired by [16], four aspects were used to represent the shot: action, face, person, and place. Specifically, this paper utilizes (1) a partial temporal action detection model based on Fast-RCNN NonLocal-I3D-50 [14] pretrained on the AVA dataset [18] to obtain the action features, (2) a multitask cascaded convolutional network (MTCNN) [19] to detect the faces in each keyframe and InceptionResnetV1 [20] to extract features of the face, (3) a cascade region-based convolutional neural network (R-CNN) [21] that was trained with the B-box annotations in MovieNet [14] based on the detection codebase MMDetection [22] to detect the people in each keyframe and a ResNet50 [23] trained with the cast annotations in MovieNet [14] to extract the person features, and (4) ResNet50 [23] pretrained on the Places dataset [24] on keyframe images to obtain place features. Finally, these aspects were combined into a visual feature representation of each shot clip suitable for learning the temporal variation in the visual information in shot subclips.

As we know, the movie's audio may be speech, music, or sound effects, etc. But since the duration of a shot clip is too short, the audio of each shot subset does not have complete semantic information. To effectively describe the characteristics of various types of sounds, audio features were extracted using the OpenSMILE toolkit [25] and a pretrained VGGish model [26], as in AttendAffectNet [10]. For OpenSMILE feature extraction, the configuration file “emobase2010” in the INTERSPEECH 2010 paralinguistics challenge [27] was used to extract 1,582 features from each audio of the shot subset with default parameters. The extracted feature set included low-level descriptors, including the jitter, loudness, pitch, mel frequency cepstral coefficients (MFCCs), mel filter bank, line spectral pairs with their delta coefficients, functionals, duration in seconds, and the number of pitch onsets [28]. For VGGish feature extraction, the pretrained VGGish network on the AudioSet dataset [29] was used. For each 0.96-second audio segment, 128-d audio features were obtained, and the extracted features and overall parts were calculated to obtain the elementwise averaging to finally obtain a 128-feature vector for each movie shot excerpt to describe the mid-level and high-level characteristics of audio. Finally, we obtained the keyframe-level visual and shot-level acoustic features.

2.3. Shot Segment Level Features. In this paper, the keyframe-level visual features and the shot-level audio features are combined into shot-level features that are suitable for describing the interaction between semantic units in the audio-

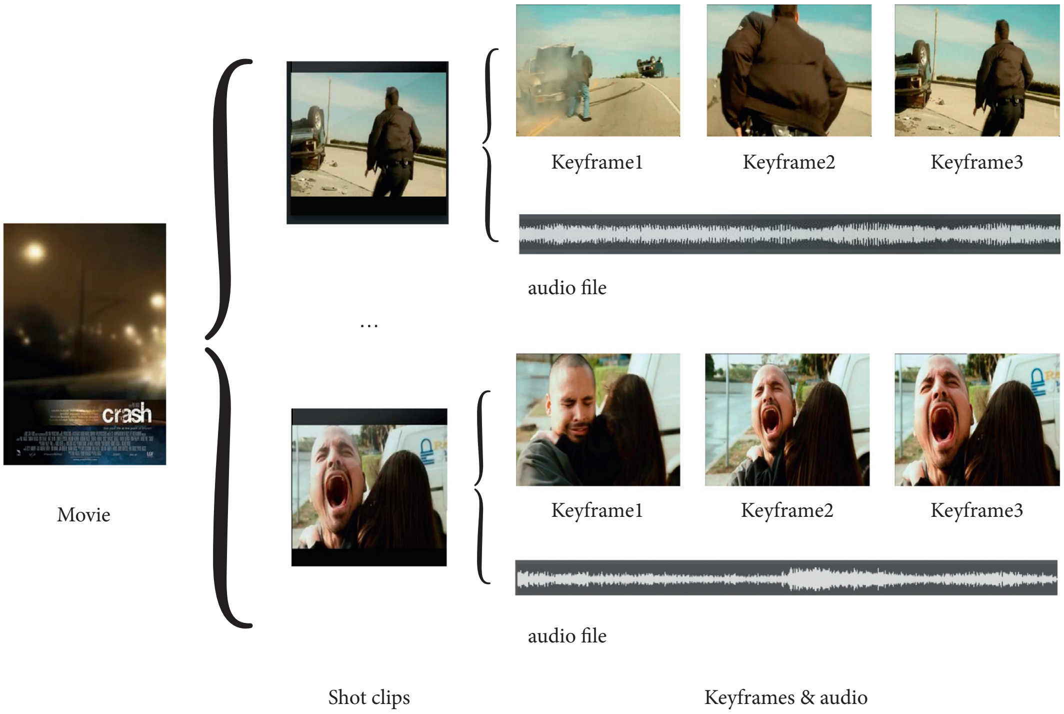

Figure 2: Preprocessing operation of the shot fragment data.

visual information of each shot segment. The process of feature processing is shown in Figure 3.

For visual features, the keyframes of the shot S_i, $i \in [1, N]$ were denoted as K_{ij}, $j \in [1, 3]$. For all four types of features of each keyframe, the action, face, person, and place features were denoted in turn as $\{Ac_{Kij}, Fa_{Kij}, Pe_{Kij}, Pl_{Kij}\}$, $i \in [1, N]$, $j \in [1, 3]$. To extract action features, person detection of each keyframe using a cascade R-CNN [21] was performed first, next a spatial-temporal action detection model was used, and then a 2048-dimensional action feature vector of one person for each keyframe was obtained, denoted as Ac_{Kij}. Then, the elementwise averaging of Ac_{Kij}, $j \in [1, 3]$ was calculated to obtain the shot-level action feature Ac_{Ki}. When there were no people in the keyframe, a 2048 zero vector was set as its action features. There could be multiple faces or people in each keyframe, so after the number of faces and people in all keyframes was counted, the number of faces and people in each keyframe with the calculated feature was set to 3. Therefore, to extract the facial feature, facial detection was performed first, and 512-dimensional feature vectors for each of the three faces in each keyframe were extracted and then concatenated to obtain a 1582-d facial feature vector Fa_{Kij}. If the number of faces in the keyframe was less than 3, the feature of each undetected face was set to a 512-d zero vector. The process of extracting the person features was similar to that of the facial features, but each person had a feature vector of 256, so a 768-d person feature vector Pe_{Kij} was obtained. For the place feature, a 2048-d vector Pl_{Kij} was extracted for each keyframe. Finally, all four features were concatenated as a visual feature representation for each keyframe. To simplify the operation, the features of the three keyframes were averaged as a shot visual feature $\{Ac_i, Fa_i, Pe_i, Pl_i\}$. For the auditory features, the 1582-dimensional features extracted by the OpenSMILE tool and the 128-dimensional features extracted by pretrained VGGish were directly concatenated and combined as the acoustic features of the shot. Then, the shot-level visual feature was concatenated with the shot-level auditory feature, and the feature was normalized as a shot-level audio-visual feature $\{Ac_i, Fa_i, Pe_i, Pl_i, Vgg_i, Op_i\}$ of shot i.

3. Multimodal Model for Emotion Prediction

The emotion evoked in an audience is related to the audio-visual information received at the current moment and the emotional state reached previously. For predicting the experienced emotion of viewers, the definition of the problem needs to be clarified. Given n shot clips, i.e., $\{s_1, s_2, \ldots, s_n\}$, and $(n-1)$ labels, denoted $\{y_1, y_2, \ldots, y_{n-1}\}$, the $n-$ th label y_n of the $n-$ th shot clip s_n needed to be predicted. A feature set $x_i = \{Ac_i, Fa_i, Pe_i, Pl_i, Vgg_i, Op_i\}$ was used to represent shot s_i. Therefore, this question translated to a nonlinear autoregressive exogenous problem. Given the n-d driving series, i.e., $\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{f_d \times n}$, where f_d is the dimension of the shot-level features of shot s_i, and the previous values of the target series $\{y_1, y_2, \ldots, y_{n-1}\}$ with $y_i \in [-1, 1]$, nonlinear mapping was needed to learn to predict the current emotion value y_n:

$$\hat{y}_n = F(y_1, \ldots, y_{n-1}, x_1, \ldots, x_n), \quad (1)$$

where $F(\cdot)$ is a nonlinear mapping function that needs to be learned.

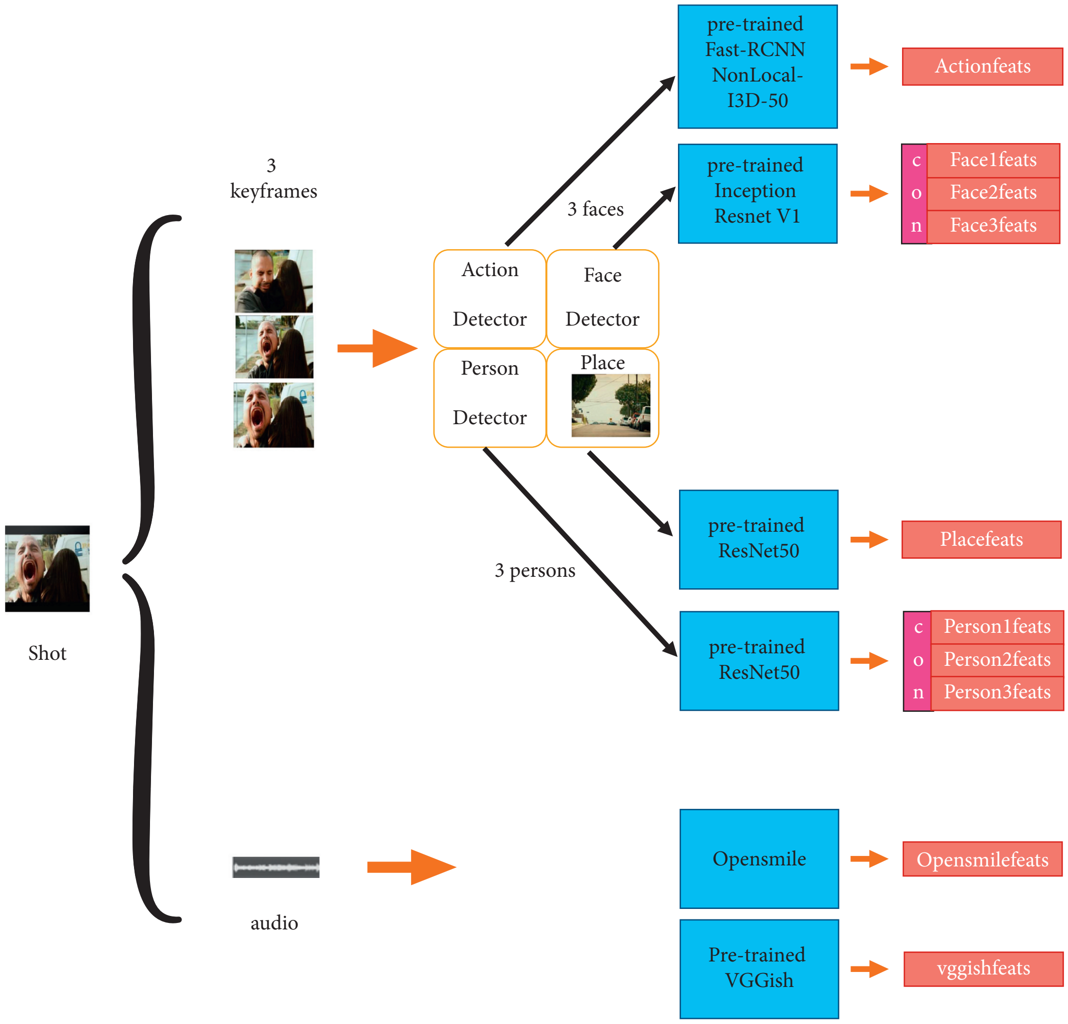

Figure 3: Extraction of vision and audio features of a shot.

To solve this problem, an LSTM model incorporating a temporal attention mechanism is proposed in this paper. Specifically, four components are included: a nonlinear multimodal feature fusion and dimensionality reduction layer, a temporal feature capture layer, a temporal attention layer, and a sentiment prediction layer, as shown in Figure 4.

3.1. Feature Fusion and Dimensionality Reduction. Inspired by the basic structure of the encoder-decoder [30], an encoder $f_1(\cdot)$ was proposed to capture the nonlinear dimensionality reduction in each feature sequence in the feature set $x_i = \{Ac_i, Fa_i, Pe_i, Pl_i, Vgg_i, Op_i\}$ and to then obtain the dimensionality reduction data $h_{1i} = \{f_1(Ac_i), f_1(Fa_i), f_1(Pe_i), f_1(Pl_i), f_1(Vgg_i), f_1(Op_i)\}$. $f_1(\cdot)$ is a nonlinear activation function that could be LSTM or something else. An LSTM unit was used as $f_1(\cdot)$ in this paper, and the update of the LSTM unit can be summarized as follows:

$$\begin{aligned}
i_{nf} &= \sigma\left(W_i\left[x_{1f}, \ldots, x_{nf}\right] + b_i\right)\\
f_{nf} &= \sigma\left(W_f\left[x_{1f}, \ldots, x_{nf}\right] + b_f\right),\\
g_{nf} &= \tanh\left(W_g\left[x_{1f}, \ldots, x_{nf}\right] + b_g\right),\\
o_{nf} &= \sigma\left(W_o\left[x_{1f}, \ldots, x_{nf}\right] + b_o\right),\\
c_{nf} &= f_{nf} \odot c_{(n-1)f} + i_{nf} \odot g_{nf},\\
h_{1nf} &= o_{nf} \odot \tanh\left(c_{nf}\right)
\end{aligned} \tag{2}$$

where W_i, W_f, W_g, and W_o and b_i, b_f, b_g, and b_o are the parameters to learn and σ and $\odot$ are a logistic sigmoid function and elementwise multiplication, respectively. $x_{nf} \in \{Ac_i, Fa_i, Pe_i, Pl_i, Vgg_i, Op_i\}$, so LSTM could be used to capture long-term dependencies of feature series to describe the interaction between high-level semantic features. Then, all dimensionality reduction features were concatenated as h_{icon}, so we obtain $h_{\text{icon}} = \{f_1(Ac_i) \oplus f_1(Fa_i) \oplus f_1(Pe_i) \oplus f_1(Pl_i) \oplus f_1(Vgg_i) \oplus f_1(Op_i)\}$, where $\oplus$ is the concatenation operators. Batch normalization [31] was used

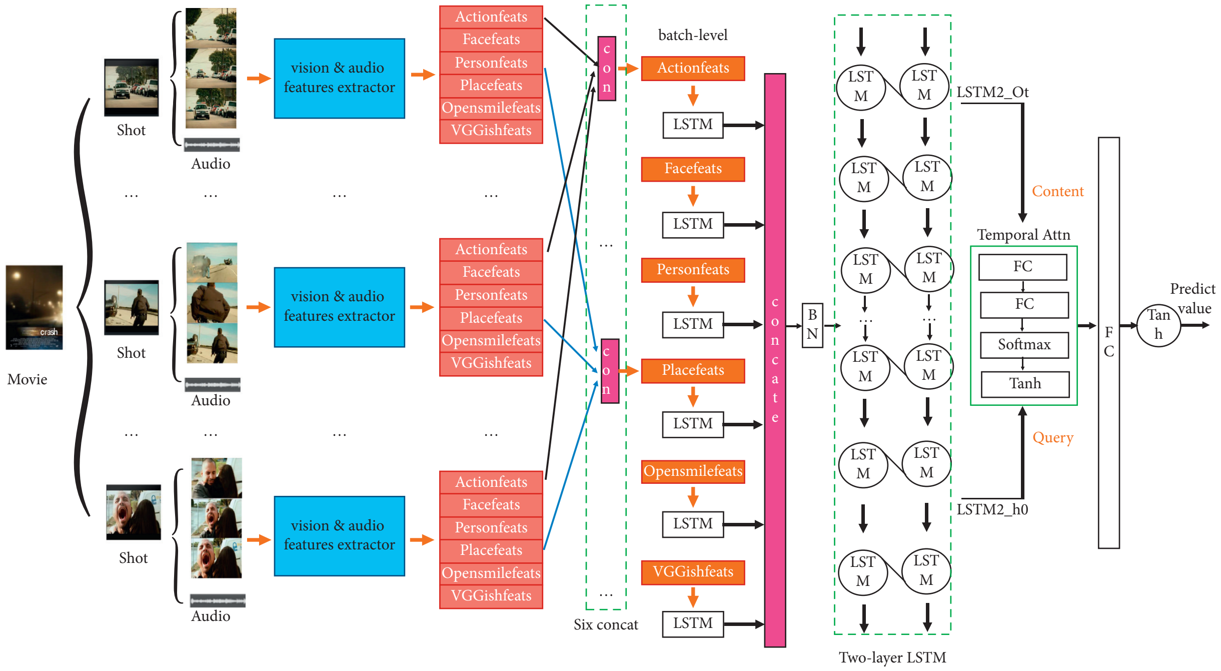

Figure 4: A model for arousal and valence prediction separately.

to normalize the feature data by recentering and rescaling, as shown in

$$h_2 = \frac{\sum_0^{bs-1} h_{\text{icon}} - E\left[\sum_0^{bs-1} h_{\text{icon}}\right]}{\sqrt{h_{\text{icon}} + \varepsilon}}. \qquad (3)$$

3.2. Temporal Feature Capture. To simulate the effects of current and previous audio-visual information on audiences, the features h_2 of shots were fed to two layers of LSTM to incorporate the time dependencies of the shot features, each with a hidden size of m units, as in Figure 4. In addition, in this paper, m was set to 30. Therefore, after the two LSTM layers, we could obtain LSTM2_Ot and LSTM2_$h0$ as follows:

$$\begin{aligned} \text{LSTM2}_{\_Ot} &= \sigma\left(W_{io}\delta^{\text{LSTM1}} h_t^{\text{LSTM1}} + b_{io} + W_{hg}h_{t-1} + b_{ho}\right), \\ \text{LSTM2}_{\_h0} &= \text{LSTM2}_{\_Ot} \odot \tanh\left(\text{LSTM2}_{\_Ct}\right) \end{aligned} \qquad (4)$$

where the input of the second layer of LSTM is the hidden state h_t^{LSTM1} of the first layer multiplied by the dropout δ^{LSTM1} and δ^{LSTM1} is a Bernoulli random variable. W_{io}, b_{io}, W_{hg}, and b_{ho} are parameters that need to be learned. LSTM2_Ot and LSTM2_$h0$ are the output and hidden states of the second LSTM layer, respectively.

3.3. Temporal Attention. To adaptively learn the influence weights of different temporal features on this paper's task, we introduce the temporal attention mechanism. We assign LSTM2_$h0$ to Query and substitute LSTM2_Ot into Content, as shown in Figure 4. Therefore,

$$S_t^k = W_2^T \tanh\left(W_1\left[\text{LSTM2}_{\_h0}; \text{LSTM2}_{\_Ot}\right]\right), \qquad (5)$$

$$\alpha_t^k = \frac{\exp\left(S_t^k\right)}{\sum_{i=1}^n S_t^i}. \qquad (6)$$

The parameters that need to be learned are W_2^T and W_1, where α_t^k is the attention weight measuring the importance of the k-th temporal feature at time t. Equation (6) is a softmax function that ensures that all the attention weights sum to 1.

Then, the temporal feature LSTM2_Ot changes to

$$\overline{\text{LSTM2}_O t} = \alpha_t^k \text{LSTM2}_{\_Ot}. \qquad (7)$$

3.4. Sentiment Value Prediction. Finally, we need to obtain the predicted value of each shot clip. Because the range of values of the valence and arousal is $[-1, 1]$, we put $\overline{\text{LSTM2}_O t}$ passing through a fully connected layer with one unit output first and then through a tanh layer, as shown in Figure 4. The calculation process is as

$$\text{Predict}_{\_\text{value}} = \tanh\left(W_3 \overline{\text{LSTM2}_O t}\right). \qquad (8)$$

Additionally, W_3 is a learnable parameter. For simplicity, the bias term was omitted.

3.5. Loss Function. Two loss functions, Loss 1 (as in equation (9)) and Loss 2 (as in equation (10)), were chosen and compared in our experiments.

For the loss function, Loss1 was defined as the mean squared error (MSE) of prediction:

$$\mathrm{Loss1} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Value}_{\mathrm{pred},i} - \mathrm{Value}_{\mathrm{ground},i}\right), \tag{9}$$

where n is the total number of shot clips of the validation set, $\mathrm{Value}_{\mathrm{pred},i}$ is the predicted value for the i − th shot clip, and $\mathrm{Value}_{\mathrm{ground},i}$ is the ground truth of the i − th shot clip. Loss2 was defined as

$$\mathrm{Loss2} = \mathrm{loss1} + \left(1 - p\left(\mathrm{Value}_{\mathrm{pred},i}, \mathrm{Value}_{\mathrm{ground},i}\right)\right), \tag{10}$$

where p is the Pearson correlation coefficient (PCC), computed from the predicted arousal/valence values and the ground truth.

4. Experiments

4.1. Dataset. As in existing research [4, 6–10], the extended COGNIMUSE dataset [4, 15], which consists of twelve half-hour additional Hollywood movie clips, was used. This dataset has the intended and experienced emotion labels at the frame level. Emotion is represented by continuous arousal and valence values in the range [−1, 1]. This paper focused mainly on experienced emotion, which is equivalent to evoked emotion and was described in terms of valence and arousal values computed as the average of twelve annotations. For comparison with previous work, the results for intended emotions were reported, representing the intention of the filmmakers, and were annotated in terms of valence and arousal values, computed as the average of three annotations done by the same expert at the frame level. In both cases, the emotion values (valence and arousal), which ranged between −1 and 1, were quantized into shot emotion labels as defined in Section 2.1. All movies in the extended COGNIMUSE dataset were used, including two animated movies, namely, “Ratatouille” and “Finding Nemo.”

4.2. Evaluation Metrics. To evaluate our proposed method, leave-one-out cross-validation was used, and the MSE and PCC between the predicted values and the ground truth for arousal/valence were chosen as our evaluation metrics, as in [9, 10]. For leave-one-out cross-validation, we selected each movie in turn as the validation set and the other movies as the training set, and the averages of all training results (MSE and PCC) were used as the overall results. For the evaluation metrics, the closer the MSE was to 0 and the closer the PCC was to 1, the better the prediction.

4.3. Shot Clip versus 5 s Clip. First, a preexperiment to briefly verify and compare the effectiveness of the two segmentation methods was conducted. To simplify the experiment, this part used the same features as Thao et al. [9]. We implemented a regression deformation based on the sequence memory model [9] to predict arousal and valence. In this model, Adam optimization [32] training was used; the learning rate of arousal and valence prediction was 0.0005; the momentum was set to 0.00002, weight attenuation was not performed, and two losses (Loss 1 and Loss 2) were used separately. The fully connected layer for feature reduction had ten cells. For LSTM, a fixed sequence length equal to 5 and 64 hidden units were used.

As shown in Table 1, under the same model and hyperparameter settings, the PCC of the shot-based prediction of the experienced arousal is 0.13 larger than that of the 5 s-segment-based, and the MSE is approximately 0.01 litter than that of the 5 s-segment-based. However, the MSE of the experienced valence prediction based on the shot segment is 0.01 larger than that of the prediction based on the 5 s segment, while the PCC of the experienced valence is 0.006 larger than that of the 5 s segment. Therefore, shot-based segmentation is more favorable to the experienced arousal prediction task than 5 s segmentation. For the experienced valence prediction task, shot-based segmentation can obtain a better PCC even with a slightly larger MSE.

4.4. Implementation Details. For separate arousal and valence prediction, the model was trained using Adam optimization with learning rates of 0.01 for arousal and 0.05 for valence. Both momenta were set to 0.005, without weight decay, and two losses (Loss 1 and Loss 2) were separately applied. The models were trained for 500 epochs, each batch size was 128, and the early stopping patience was 70 epochs. For the LSTM, the fixed sequence length was set to 5, and both had 30 hidden units. All models were implemented in Python 3.6 with PyTorch 1.4 and were run on an NVIDIA GTX 2080ti.

5. Results and Analysis

5.1. Comparison with State-of-the-Art Results. As experienced emotion prediction is more difficult than intended emotion prediction, most researchers have focused on intended emotion prediction. In 2019, Thao et al. [9] used the same features and models to predict intended or experienced emotion, respectively, and this was chosen as the baseline in this paper. Therefore, the model was trained and validated on the experienced and intended emotion annotations in the COGNIMUSE dataset for effective comparison. The results are summarized in Table 2 for experienced emotion prediction and in Table 3 for intended emotion prediction.

As shown in Table 2, the performance of our prediction method is significantly better than that of Thao et al. [9]. The best results are obtained for each forecasting task when arousal or valence is predicted with Loss1 or Loss2, respectively. In particular, the MSE of the arousal prediction task decreases from 0.04 to 0.027, and the PCC increases from 0.46 to 0.62; the MSE of the valence prediction task changes from 0.06 to 0.063, and the PCC improves from 0.18 to 0.34. As shown in Figures 5 and 6, the curve of the predicted values always exhibits a sudden large change over time. When the forecasting model does not have sufficient fitting power, the overall curve consisting of all the forecast values flattens out, resulting in a small MSE and a small PCC, which does not fit the sharp fluctuations in the curve. Therefore, we prefer methods that give higher PCC values to prioritize the model itself and to ensure better fitting

Table 1: Comparison of the performance of the two segmentation methods.

Features	Experienced arousal		Experienced valence	
	MSE	PCC	MSE	PCC
5 s features4 (loss2)	0.0476	0.4543	**0.0707**	0.2145
Shot features4 (loss2)	**0.0377**	**0.5822**	0.0820	**0.2203**

The best result for each indicator is marked in bold.

Table 2: Comparison of state-of-the-art results for experienced emotion prediction.

Method	Arousal		Valence	
	MSE	PCC	MSE	PCC
Thao et al. [9]	0.04	0.46	0.06	0.18
Ours (loss1)	**0.0275**	**0.6187**	**0.0490**	0.2828
Ours (loss2)	0.0403	0.5569	0.0632	**0.3443**

Table 3: Comparison of state-of-the-art results for intended emotion prediction.

Models	Arousal		Valence	
	MSE	PCC	MSE	PCC
Malandrakis et al. [4]	0.17	0.54	0.24	0.23
Goyal et al. [6]	—	0.62 ± 0.16	—	0.29 ± 0.16
Sivaprasad et al. [7]	0.08 ± 0.04	**0.84 ± 0.06**	0.21 ± 0.06	0.50 ± 0.14
Thao et al. [9]	0.13	0.62	0.19	0.25
Thao et al. [10]	0.124	0.630	0.178	**0.572**
Ours (loss1)	**0.1022**	**0.6748**	**0.1654**	0.3167
Ours (loss2)	0.1141	0.6582	0.1704	**0.4025**

performance. Therefore, Loss1 is more suitable for arousal prediction tasks, while is better for valence prediction.

As shown in Table 3, our model obtains results comparable to the state-of-the-art results for both arousal and valence in intended emotion prediction. For arousal-intended emotion prediction, Sivaprasad et al. [7] obtained the best results, with a PCC equal to 0.84 and an MSE of 0.08. However, their data excludes the two animated movies from the COGNIMUSE dataset just as in [6]. The PCC and MSE of our method are 0.67 and 0.10, respectively, which are the best results without handcrafted features. For valence-intended emotion prediction, Thao et al. [10] obtained the best results, with a PCC of 0.57 and an MSE of 0.17. The results of our method are slightly worse, with a PCC of 0.40 and an MSE of 0.17. This may be because our model is not as complex as theirs, but despite this, our arousal prediction performance surpasses theirs, indicating that our feature combination is more effective. Compared to the model of equal-level complexity reported by Thao et al. [9], a significant improvement is reported for the intended emotion prediction. Specifically, both arousal and valence prediction have smaller MSE and larger PCC.

The arousal and valence dimensions of the experienced emotion of two movies, "American Beauty" and "A Beautiful Mind," are visualized in Figures 5 and 6, respectively.

As shown in Figures 5 and 6, the predicted result of arousal is much better than the valence value, which means that the nonlinear mapping function from audio-visual cues to valence is more challenging to learn. There are clear opposite trends for the predicted arousal values and ground truth in some periods. This phenomenon may mean that there are different ways in which the same features interact with each other in various movies and that these ways of interaction are affected by the specific values of the features.

5.2. Ablation Experiments of Features. To verify the contribution of each visual or auditory feature to the overall effect, feature ablation experiments were conducted. This section was validated using the same model with the same hyperparameters as that in Section 4.4.

As shown in Table 4, we subtract one feature each time we validate, "Action features," "Face features," "Person features," "Place features," "VGGish features," or "OpenSMILE features," and use only the visual and auditory features for prediction. The experimental results show that the contribution of each visual or auditory feature to the overall arousal prediction results is generally comparable. Although the performance of prediction by visual features alone is still worse than that by only acoustic features, they are comparable, indicating that our proposed visual features have a better ability to describe arousal information. For the valence prediction task, the performance of prediction by auditory features alone is much higher than that of visual features, indicating that the proposed audio feature set has a better ability to describe valence attributes than the visual feature set. Without the features of the person, the multimodel obtains the best PCC, which is 0.37, and a slightly worse MSE than all features for valence prediction. This may be because a variety of different emotions can be evoked for

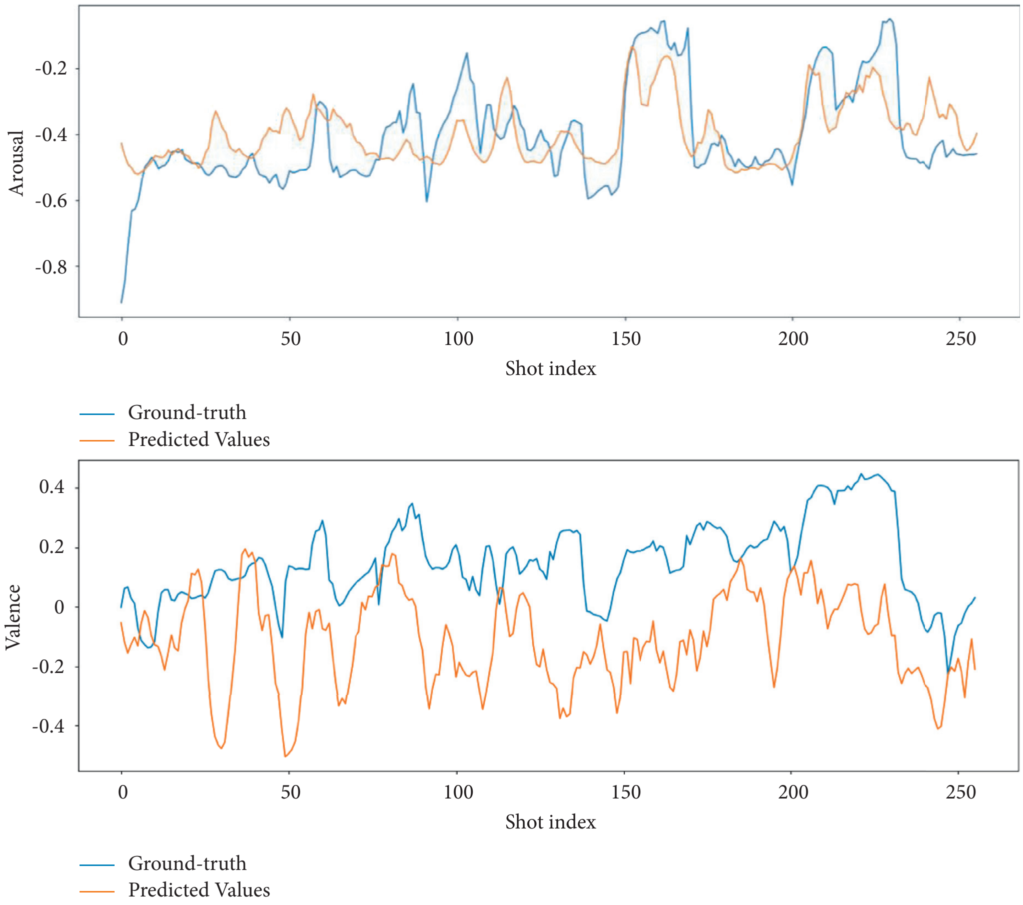

Figure 5: Visualization of the arousal and valence dimensions of the experienced emotion of the movie named "American Beauty."

the videos with the same human pose and the current model is not complex enough to differentiate it.

5.3. Ablation Experiments of LSTM-One-Layer Encoder. This section aims to validate the ability to capture the nonlinear dimensionality reduction in each feature sequence in the feature set using LSTM and its contribution to the overall arousal and valence predictions. A comparative model was designed for feature downscaling using a fully connected layer and for predicting arousal and value using Loss1 or Loss2. As shown in Table 5, the use of LSTM to capture the nonlinear dimension reduction in each feature sequence in the feature set has a better effect on the overall prediction performance. In particular, the MSE of the arousal prediction task decreases from 0.0288 to 0.0275, and the PCC increases from 0.5826 to 0.6187; the MSE of the valence prediction task decreases from 0.0751 to 0.0632, and the PCC improves from 0.3276 to 0.3443. Therefore, our LSTM-one-layer encoder helps to improve the results of all tasks.

5.4. Ablation Experiments of the Time Attention Mechanism. In this part, we validated the performance of the time attention mechanism. As shown in Table 6, the prediction performance of the time attention mechanism is significantly improved. Specifically, the PCC values for both arousal and valence emotion prediction improve from 0.574 and 0.296 to 0.618 and 0.34, respectively. The MSE decreases from 0.034 and 0.071 to 0.027 and 0.063, respectively, for arousal and valence. This proves the effectiveness of the time attention mechanism.

5.5. Computational Complexity Comparison. In the feature extraction process, the computational complexity of visual features is much higher than that of audio features, and the extraction process of auditory features in existing studies is similar, with computational complexity on the same order of magnitude. Therefore, we focus on the computational complexity of the visual feature extraction process for comparison. As shown in Table 7, existing studies need to extract every frame's features in the video, while the number of frames to be processed is reduced to 3.21% of the original number of frames in this paper. Optical flow information extraction is replaced by action feature extraction, which has low computational complexity and consumes less time. Therefore, this method significantly reduces the number of calculations and time.

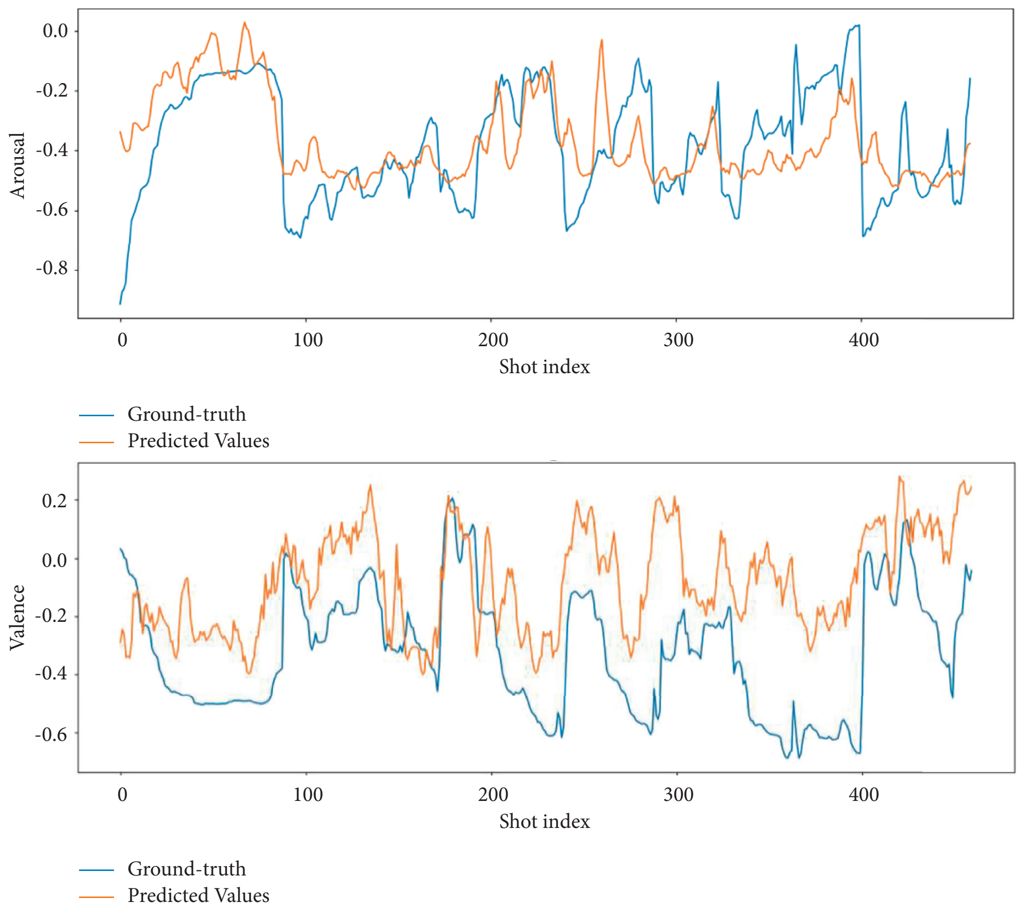

Figure 6: Visualization of the arousal and valence dimensions of the experienced emotion of the movie named "A Beautiful Mind."

Table 4: Comparison of state-of-the-art results for experienced emotion prediction.

Features	Arousal (loss1)		Valence (loss2)	
	MSE	PCC	MSE	PCC
All features	**0.0275**	**0.6187**	**0.0632**	0.3443
−Action features	0.0291	0.6038	0.0673	0.3259
−Face features	0.0277	0.6136	0.0637	0.3667
−Person features	0.0280	0.6181	0.0653	**0.3726**
−Place features	0.0280	0.5981	0.0663	0.3315
−VGGish features	0.0290	0.5952	0.0669	0.3444
−OpenSMILE features	0.0295	0.6003	0.0666	0.3345
All_visual_features	0.0316	0.4931	0.0751	0.2694
All_audio_features	0.0297	0.6141	0.0726	0.3356

"−" indicates without the feature.

Table 5: With or without capture changes in audio and visual feature sequences using LSTM.

Model (with Features6)	Experienced arousal (loss1)		Experienced valence (loss2)	
	MSE	PCC	MSE	PCC
Ours without LSTM	0.0288	0.5826	0.0751	0.3276
Ours	**0.0275**	**0.6187**	**0.0632**	**0.3443**

Table 6: With or without time attention mechanism.

Model (with Features6)	Experienced arousal (loss1)		Experienced valence (loss2)	
	MSE	PCC	MSE	PCC
Ours without attention	0.0342	0.5746	0.0718	0.2964
Ours	**0.0275**	**0.6187**	0.0632	**0.3443**

Table 7: Computational complexity comparison.

Models	Number of frames	Optical flow
Goyal et al. [6]	551112	✓
Sivaprasad et al. [7]	551112	✓
Thao et al. [9]	551112	✓
Thao et al. [10]	551112	✓
Ours	**17706**	×

6. Conclusion

In this paper, a multimodal prediction model based on video shot segmentation for predicting affective responses evoked by movies is presented. Unlike many existing studies, this paper introduces the shot clip as the minimum emotion prediction of the video unit and avoids the optical flow calculation in feature extraction. This method enables our model to focus on analyzing each semantic unit's audio and visual information in the interaction process. Therefore, the emotion prediction task performance was significantly improved, with an increase from 0.46 to 0.62 for arousal and from 0.18 to 0.34 for valence in experienced emotion.

The movie was divided into short clips by shot boundary detection first. Then, three keyframes were extracted from each shot clip. Four types of features—action, face, person, and place features—for each keyframe of each shot clip were extracted. For each shot clip's audio, we used the OpenSMILE tool and a pretrained VGGish model to extract audio features. Then, they were combined as shot audio-visual feature representations. Finally, the shot audio-visual features were fed into our deep multimodal model based on LSTM incorporating temporal attention to predict emotion.

The effects of arousal and valence predicted separately were compared by using our model with two types of loss functions. Loss1 is more suitable for arousal prediction tasks, while Loss2 is better for valence prediction. In the future, we will work on designing scene-level video feature calculation methods and a better model for mapping the complex changes in visual and audio information at different temporal levels and segment levels of the movie to the experienced emotion. Interestingly, the feature combination method based on shot fragments can significantly improve the performance of the method in [9]. The experimental results of this paper reveal the importance of considering the time structure and semantic unit of movies for experienced emotion prediction.

Data Availability

The official website of the COGNIMUSE dataset is http://cognimuse.cs.ntua.gr/database. The complete data are available from this website and by contacting the creators of this dataset.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (CUC19ZD005 and CUC200D050).

References

[1] L. Fernández-Aguilar, B. Navarro-Bravo, J. Ricarte, L. Ros, and J. M. Latorre, "How effective are films in inducing positive and negative emotional states? A meta-analysis," *PloS one*, vol. 14, no. 11, Article ID e0225040, 2019.

[2] K. Yadati, H. Katti, and M. Kankanhalli, "CAVVA: computational affective video-in-video advertising," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 15–23, 2013.

[3] D. Aditya, R. G. Manvitha, M. Samyak, and B. S. Shamitha, "International conference on computing system and its applications emotion based video player," *Global Transitions Proceedings*, vol. 2, no. 1, 2021.

[4] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proceeding of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 22-27 May 2011.

[5] Y. Baveye, E. Dellandrea, and C. L. Chamaret, "LIRIS-ACCEDE: a video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[6] A. Goyal, N. Kumar, T. Guha, and S. S. Narayanan, "A multimodal mixture-of-experts model for dynamic emotion prediction in movies," in *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 20-25 March 2016.

[7] S. Sivaprasad, T. Joshi, R. Agrawal, and N. Pedanekar, "Multimodal continuous prediction of emotions in movies using long short-term memory networks," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, Yokohama Japan, 05 June 2018.

[8] T. Joshi, S. Sivaprasad, and N. Pedanekar, "Partners in crime: utilizing arousal-valence relationship for continuous prediction of valence in movies," in *Proceedings of the 2nd Workshop on Affective Content Analysis (AffCon 2019) co-located with Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019)*, Honolulu, USA, January 2019.

[9] H. T. P. Thao, D. Herremans, and G. Roig, "Multimodal deep models for predicting affective responses evoked by movies," in *Proceedings of ICCV Workshops*, 27-28 Oct. 2019.

[10] H. T. P. Thao, B. T. Balamurali, D. Herremans, and G. Roig, "AttendAffectNet: self-attention based networks for predicting affective responses from movies," in *25th International Conference on Pattern Recognition (ICPR)*, 10-15 Jan. 2021.

[11] R. Dietz and A. Lang, "Affective agents: effects of agent affect on arousal, attention, liking and learning," in *Proceedings of the Third International Cognitive Technology Conference*, San Francisco, August 1999.

[12] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: a holistic dataset for movie understanding," 2020. arXiv preprint.

[15] A. Zlatintsi, P. Koutras, G. Evangelopoulos et al., "COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–24, 2017.

[16] A. Rao, L. Xu, Y. Xiong et al., "A local-to-global approach to multimodal movie scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13-19 June 2020.

[17] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "Temporal video segmentation to scenes using high-level audiovisual features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1163–1177, 2011.

[18] C. Gu, C. Sun, D. A. Ross et al., "Ava: a video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18-23 June 2018.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, California, USA, February 2017.

[21] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 18-23 June 2018.

[22] K. Chen, "MMDetection: open mmlab detection toolbox and benchmark," arXiv preprint, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 27-30 June 2016.

[24] B. Zhou, A. Lapedriza, and A. A. Khosla, "Places: a 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1452–1464, 2017.

[25] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Springer, Berlin/Heidelberg, Germany, 2015.

[26] S. Hershey, S. Chaudhuri, D. P. W. Ellis et al., "CNN architectures for large-scale audio classification," in *Proceedings of 2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 5-9 March 2017.

[27] B. Schuller, S. Steidl, A. Batliner et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[28] F. Eyben, F. Weninger, M. Wöllmer et al., *Open-source media interpretation by large feature-space extraction*, TU Munchen, Munich, Germany, 2016.

[29] J. F. Gemmeke, D. P. W. Ellis, D. Freedman et al., "Audio set: an ontology and human-labeled dataset for audio events," in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5-9 March 2017.

[30] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*arXiv preprint, Doha, Qatar, October 2014.

[31] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning PMLR*, pp. 448–456, Lille, France, July 2015.

[32] D. P. Kingma and B. Jimmy, "Adam: a method for stochastic optimization," arXiv preprint, 2014.