Hindawi
IET Software
Volume 2023, Article ID 5566781, 10 pages
https://doi.org/10.1049/2023/5566781

Research Article

Analysis of Emotional Deconstruction and the Role of Emotional Value for Learners in Animation Works Based on Digital Multimedia Technology

Shilei Liang

Sch Ary and Design, Mudanjiang Normal University, Mudanjiang 157000, Heilongjiang, China

Correspondence should be addressed to Shilei Liang; 0509019@mdjnu.edu.cn

Received 18 August 2023; Revised 24 October 2023; Accepted 1 November 2023; Published 22 November 2023

Academic Editor: Sangbing Tsai

With the rapid development of artificial intelligence and digital media technology, modern animation technology has greatly improved the creative efficiency of creators through computer-generated graphics, electronic manual painting, and other means, and its number has also experienced explosive growth. The intelligent completion of emotional expression identification within animation works holds immense significance for both animation production learners and the creation of intelligent animation works. Consequently, emotion recognition has emerged as a focal point of research attention. This paper focuses on the analysis of emotional states in animation works. First, by analyzing the characteristics of emotional expression in animation, the model data foundation for using sound and video information is determined. Subsequently, we perform individual feature extraction for these two types of information using gated recurrent unit (GRU). Finally, we employ a multiattention mechanism to fuse the multimodal information derived from audio and video sources. The experimental outcomes demonstrate that the proposed method framework attains a recognition accuracy exceeding 90% for the three distinct emotional categories. Remarkably, the recognition rate for negative emotions reaches an impressive 94.7%, significantly surpassing the performance of single-modal approaches and other feature fusion methods. This research presents invaluable insights for the training of multimedia animation production professionals, empowering them to better grasp the nuances of emotion transfer within animation and, thereby, realize productions of elevated quality, which will greatly improve the market operational efficiency of animation industry.

1. Introduction

With the relentless advancement of science and technology, accompanied by the rapid development of digital media technology, multimedia paintings have acquired a prominent and indispensable role in contemporary society. The advent of digital media has granted animation the liberty to transcend traditional constraints, bestowing upon it a more intricate and diverse visual panorama. As an imaginative and creative form of artistic expression, animation has seamlessly ingrained itself into people's daily lives, emerging as an indispensable medium of entertainment [1]. Nevertheless, as animation evolves, the emotional facets embedded within its works have garnered increasing attention. The transmission and manifestation of emotions in animation have become a captivating area of inquiry, impelling researchers to delve into the profound dissection of emotions within animation and their profound impact on the audience. Animation, being a visual medium, possesses a distinctive means of expression, skillfully relaying an abundant array of emotional content to its viewers through elements such as colors, shapes, music, and dynamic imagery. When we immerse ourselves in animation, we often experience emotional resonance engendered by the emotions interwoven within, including but not limited to joy, sadness, and surprise [2]. In the continuous development of computer technology today, many animation works are generated through computers. CGI technology can help animator complete realistic 3D images and improve creative efficiency. At the same time, the digitization of hand-drawn animation and the widespread application of VR, AR, and other technologies have also given current animation works more forms of

expression. These animations also convey more emotions through a wider range of media dissemination methods, meeting people's emotional needs through high-precision and diverse technologies.

Sentiment analysis has progressively emerged as a consequential research domain in contemporary society. The widespread adoption of the Internet and social media has led to an extensive collection and dissemination of user-generated sentiment data, encompassing a wealth of information about individuals' emotional states. Leveraging sentiment analysis, we can gain a deeper comprehension of how human emotions are expressed and communicated, thereby obtaining insights into emotional tendencies within social groups. This knowledge finds applications in diverse fields such as market research, mental health assistance, and beyond [3]. Particularly, in the realm of education, comprehending learners' emotional states during the learning process proves invaluable, facilitating the optimization of teaching strategies and enhancing learners' overall learning experience and outcomes. However, when it comes to vast digital multimedia works, such as animation, manual analysis of emotions within them is a laborious and time-consuming endeavor. Thankfully, the progress in artificial intelligence technology allows us to expeditiously and intelligently analyze emotions in animation works by harnessing machine learning and deep learning techniques [4]. Through the construction of emotion recognition and emotion intensity analysis models, we can automatically extract emotion-related information from animation works, empowering us to thoroughly decipher the expression of emotions within the animation [5]. Moreover, by combining the outcomes of sentiment analysis with learner emotion data, we can further explore the influence of emotions conveyed through animation works on learners' emotional states, thus facilitating the refinement of teaching methodologies to enhance learners' overall educational experience and effectiveness.

The intelligent analysis facilitated by artificial intelligence will enable us to better comprehend and harness the emotional components inherent in digital multimedia technology, ushering in novel opportunities and challenges for the realms of animation creation and education. Therefore, this paper intends to study the emotions conveyed by animation works, hoping to speed up the production and review process of animation works under the wave of digital media by intelligently analyzing the emotions conveyed by animation works. The specific work of this paper is as follows:

(1) In this paper, the timing feature extraction of two types of information is accomplished by using gated recurrent unit (GRU) for audio signal and video signal in emotion recognition of multimedia animation works.
(2) Based on the extracted temporal features, the multimodal information fusion is realized by using multi-head attention (MHA) method.
(3) Through the fusion of the existing dataset and the self-made dataset to complete the three types of emotion recognition of animation works, its recognition rate for positive, neutral, and negative emotions is higher than 90%.

The rest of the paper is arranged as follows: in Section 2, related works for the emotion analysis and deep learning-based affective computing are described; Section 3 established the framework for the emotion analysis using the GRU feature extraction and the MHA; in Section 4, the experiment and result analysis is given. Section 5 discusses the related works and the meaning for the investigation. The conclusion is drawn at last.

2. Related Works

2.1. Current Status of Sentiment Analysis. The term "emotion" traces its origin back to the Greek word "Pathos," which originally denoted people's heartfelt responses toward tragedy [6]. The complexity of emotion has led to a lack of consensus among scholars regarding its definition, resulting in over 150 theories on the subject [7]. Picard distinguished "emotion" from "sentiment" in her book, asserting that emotion represents a longer-lasting affective state than mood. They manifest through various means, including gestures, footsteps, voices, and facial expressions, with facial expressions being the most controllable form of emotional expression. Facial expressions serve as external manifestations of internal emotions, with different emotions being associated with distinct changes in facial muscles, especially around the eyes and mouth, although other changes may not be as conspicuous and are harder to detect. In the realm of affective computing, researchers have adopted a combination of methods, primarily multimodal integration, augmentation of additional information, and technology-infused solutions that remain attuned to the demands of the era, leading to improved emotion recognition rates. Amer et al. [8] emphasized the significance of the eyes and mouth as crucial regions for recognizing basic emotions, emphasizing the necessity of employing a multifaceted approach to facial expression recognition. Ashfahani et al. [9] proposed the novel concept of radial coding inspired by the human visual cortex to enhance emotion recognition performance, effectively integrating complex facial expressions and achieving hybrid recognition of facial expressions. Bengio [10] leveraged textual and speech emotion rhymes from linguistic data federation to achieve deep emotion recognition through cross-validation and rhyme feature extraction and classification. Taboada et al. [11] employed an emotion lexicon, combining the emotion polarity of each word with contextual transfer information to ascertain the emotion polarity of text. Bollegala et al. [12] devised an effective solution for cross-domain sentiment analysis by utilizing common sentiment expressions to enhance cross-domain sentiment classification. Pang et al. [13] employed the support vector machines (SVM) classification algorithm, combining topic text classification and sentiment analysis of movie review text, demonstrating superior classification performance of the SVM algorithm through their experiments.

2.2. Deep Learning-Based Multimodal Emotion Recognition Approach. By leveraging modality-specific information, multimodal techniques excel at accessing and analyzing more intricate and comprehensive data compared to single-modality approaches. Previous research efforts have focused on fusing diverse single-modal data to unveil complementary and cross-modal information, aiming to achieve the integration of complex information. For instance, Dong et al. [14] proposed a multimodal information generalization and correlation analysis of linear relationships between multiple modalities. Miwakeichi et al. [15] identified a set of variables from multisource datasets that conform to their proposed relationship based on multivariate variables, known as the partial least squares modeling relationship. Sun et al. [16] discovered and introduced a multimodal modal independence technique involving independent component analysis for each modality, enabling probabilistic modeling within a Bayesian framework for multimodal data analysis.

Due to previous multimodal data exhibiting certain correlations, a simple shallow feature representation of such data was insufficient to adequately capture the external relationships and internal structures among multiple modalities. Consequently, proposed multimodal data fusion methods have been characterized by several traits, including large volumes of diverse data, high processing speeds, and enhanced accuracy of multimodal data [17]. Researchers have furthered exploration of multimodal sentiment analysis by embracing deep learning architectures, enabled by the increasing maturity of hardware and software facilities. Thao et al. [18] pursued feature layer fusion, initially extracting feature representations of different modalities using three submodules. Subsequently, these features were concatenated, and modal relationships were adapted using downstream task modules to achieve sentiment classification results. Zadeh et al. [19] introduced the tensor fusion network (TFN) for modal fusion, wherein text, video, and audio feature vectors encoded by subnetworks were subjected to outer product operations to obtain bimodal intermodal correlations. Further, the outer product was applied with the remaining unimodal encodings to achieve trimodal fusion information. Tsai et al. [20] employed multiple pairs of bidirectional crossover encoders to directly focus on and fuse low-level features from each modality.

While unimodal-based sentiment analysis has shown promising results, human emotional expression fundamentally stems from the combination of modalities, where potential connections between various elements like body posture, facial expression, and speech exist. Neglecting such connections in unimodal sentiment analysis may lead to a loss of valuable information. Consequently, researchers are striving to establish intermodal correlation information through multimodal fusion, seeking more accurate and comprehensive emotional representations. For animation works, analyzing emotions solely based on video data proves challenging due to the varying expertise levels between design and production units. Therefore, it becomes essential to incorporate other emotional data to achieve a more comprehensive analysis. This paper aims to leverage existing multimodal emotion recognition and classification foundations to achieve high-precision emotion classification by utilizing both audio data and animation video data.

3. Emotion Recognition Modeling by Fusing Audio and Video Images

3.1. GRU-Based Data Feature Extraction. The essence of emotion classification task is a classification decision task based on time-series data. Therefore, this article uses the GRU method in recurrent neural networks for feature extraction. GRU can accept audio and video feature sequences, which helps to model the evolution of emotional changes in sound and images over time, thereby improving the accuracy and performance of emotional analysis and providing strong support for multimodal emotional analysis. Considering the characteristics of animation works, this paper selects audio modal data and video modal data as the input of the model, and denoising of audio data and video data is realized by manual annotation and filtering methods in the preprocessing process of the data, and the obtained, e.g., feature vectors are shown in Equations (1) and (2):

$$F_v = \{v_1, v_2, \cdots, v_n\}, \tag{1}$$

$$F_a = \{a_1, a_2, \cdots, a_n\}, \tag{2}$$

where F_v denotes the obtained video feature input and F_a denotes the audio feature input; after completing the data preprocessing, we carried out the feature processing of time-series data through the GRU module. In the field of deep learning, recurrent neural networks (RNNs) are specialized neural networks designed for processing sequential data, such as natural language, time series, and audio [21]. However, traditional RNNs encounter challenges with long-term dependencies, experiencing issues like gradient vanishing and gradient explosion, which hinder their ability to capture complex relationships between lengthy sequences. To address these problems, researchers have introduced an improved RNN model known as the long short-term memory network (GRU). GRU is a gated RNN that outperforms traditional RNNs in capturing long-term dependencies when processing sequence data. It incorporates two key mechanisms called reset gate and update gate. These gating structures empower GRU to decide when to retain historical information and when to discard previous information, allowing it to handle long-term dependencies more effectively [22]. In the GRU architecture, two control units, namely the update gate and reset gate, play a central role. They are computed as demonstrated in Equations (3) and (4):

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \tag{3}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]). \tag{4}$$

The update gate, denoted as z_t, plays a crucial role in controlling the impact of inputs from the current time step on the hidden state. It determines how much information from the current input should be incorporated into the hidden state.

The reset gate, on the other hand, is responsible for controlling the extent to which the hidden state from the previous time step should be forgotten or reset. It influences how much of the previous hidden state should be combined with the current input to calculate the new candidate hidden state.

The candidate hidden state (also known as the temporary memory or proposed update), represented as h_t', is another vital component of the GRU structure. It is responsible for combining the hidden state from the previous time step, after considering the reset gate, with the input from the current time step. This combination results in the calculation of the new candidate hidden state, which is used in the subsequent steps of the GRU computation. The calculation of the candidate hidden state is depicted in Equation (5).

$$h_t^{'} = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]). \tag{5}$$

By updating the gate resetters, we can get the latest hidden state of the whole time series, as shown in Equation (6):

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t^{'}. \tag{6}$$

Therefore, the two classes, we obtained by the GRU method, are represented by Equations (7) and (8), respectively:

$$G_v = \{G_1^v, G_2^v, \cdots, G_n^v\}, \tag{7}$$

$$G_a = \{G_1^a, G_2^a, \cdots, G_n^a\}. \tag{8}$$

The features extracted through GRU are subsequently processed to obtain independent features of multimodal information, based on which the attention mechanism is utilized to complete the sentiment classification, and the network construction based on the attention mechanism is described in detail in the next subsection.

3.2. Feature Fusion and Sentiment Analysis Modeling Based on Multiple Attention Mechanisms. Considering the need to focus on the core data video in the process of analyzing the sentiment structure of animation works, this paper adopts a MHA mechanism model to analyze the data [23]. The role of the MHA mechanism model in emotion classification based on video and audio data is mainly reflected in its ability to integrate multimodal features, improve interpretability, adaptively learn feature weights, and, thus, improve the accuracy and robustness of emotion classification. This makes it a powerful tool for handling multimodal emotional analysis tasks. Through the attention mechanism to focus on the features of the video data to amplify, so as to improve the effect of sentiment classification, the MHA mechanism used in this paper is shown in Figure 1.

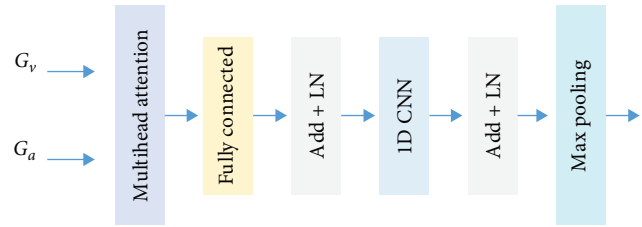

FIGURE 1: The structure of the multihead attention.

For better model analysis and subsequent comparison, this paper applies the proposed MHA method only to the emotion classification study after feature extraction is completed. When modal fusion is performed, multiple independent attention heads are then utilized to perform scaled dot product attention computation, and each head computes the similarity in a different subspace to obtain the attention distribution. Each head calculates the similarity in different subspaces to obtain the attention distribution, which is used as video information to extract the common features among modalities. For a single attention head, it is calculated as follows:

$$\text{Head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{9}$$

$$\text{Attention}\,(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{in}}}\right)V, \tag{10}$$

where Q, K, and V stand for Q(query), K(key), and V(value), respectively, and the three inputs are obtained from the features G_a and G_v weighted by weights W_i^Q, W_i^K, W_i^V. The results of combining these subspaces after performing the attention head computation are as follows:

$$\text{Multihead}\,(Q, K, V) = \text{Concat}\,(\text{Head}_1, \ldots, \text{Head}_n)W^O. \tag{11}$$

After several feature extraction and combination using the structure, as shown in Figure 1, we get the features for final classification, as shown in Equation (12):

$$F_{\text{fusion}} = \text{Concat}\left(F_{ta}^{'}, F_{tv}^{'}\right). \tag{12}$$

In order for the model to better discriminate the speaker's emotion, a speaker gender recognition auxiliary task is added in the classification stage, taking into account the characteristics of the animated characters. The resulting loss function is shown in Equation (13):

$$\text{Loss} = \frac{1}{N}\sum_{n=1}^{N}\left(\mu L(y_e^n, \widehat{y}_e^n) + (1-\mu)L\left(y_g^n, \widehat{y}_g^n\right)\right), \tag{13}$$

where μ is a hyperparameter that represents the weight value of the loss function for the emotion recognition task, which is

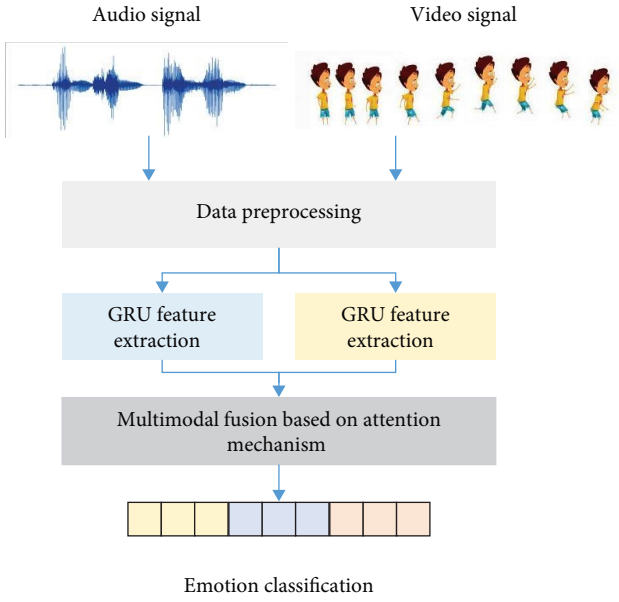

Figure 2: The framework for the emotion classification using the video and audio information.

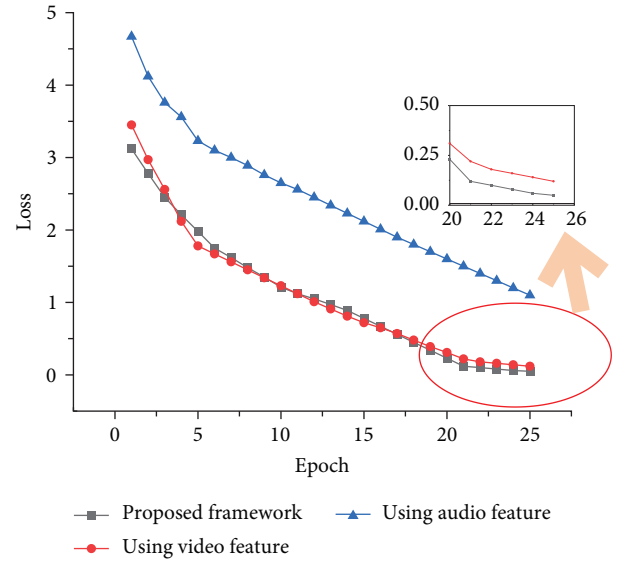

Figure 3: The training process and the loss for the framework using different features.

used to regulate the training of the two tasks and y denotes the prediction value. After the completion of feature extraction, network construction and model loss function definition can be realized in the animation works of sentiment analysis, this paper describes the sentiment structure analysis framework, as shown in Figure 2.

4. Experiment Result and Analysis

Once the model construction was completed, we proceeded to select the MOSI dataset along with some animation clips from the MOSEI dataset for the model training process. Additionally, videos and animation clips were collected from various online video websites to form the model data [24]. The MOSI dataset contains over 2,000 video clips from YouTube. These fragments are divided into three categories: positive, negative, and neutral emotions, with data types mainly including text transcription, audio, and video. The MOSEI dataset mainly includes data on movie clips and speeches, with approximately 1,000 clips. The data labels and modalities of this dataset are the same as those of MOSI. Considering the unstable characteristics of MOSI and MOSEI datasets and the characteristics of video tasks, this article only selected a small portion of approximately 500 data for dataset construction. In addition, the dataset will be mixed with relevant animation and digital media design company data from collaborators.

To optimize the use of the data, we enlisted the aid of volunteers to conduct data labeling refinement research. The videos were categorized into three emotion categories: positive, neutral, and negative. This classification was instrumental in conducting the relevant analysis as proposed in our framework. For the data labeling process, we engaged animation majors and producers, who performed two-person synchronous label division on a self-created dataset. During the labeling process, approximately 500 data samples in the dataset were classified and labeled. To ensure consistent model input, we truncated the data to equal lengths, removing any emotionally irrelevant or invalid data. Once the data preprocessing was complete, we commenced the training of the model and conducted relevant discussions based on the framework, as shown in Figure 2.

In this paper, to evaluate the model performance, we employ the precision, recall, and F1 score index, which can be calculated as follows:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \tag{14}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \tag{15}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

4.1. Model Training Loss and Classification Results. After confirming the data using the constructed dataset, we proceeded to train the model utilizing the loss function indicated in Equation (13). During training, we compared the loss function curves of different modal features, and the results are visualized, as shown in Figure 3.

As shown in Figure 3, it is evident that the video feature plays a crucial role in enhancing the model's recognition rate. Even when used in isolation, the video feature yields a final loss of less than 1. On the other hand, the multimodal feature exhibits a smaller initial loss, and the difference in final loss values between the two features is less than 0.1, underscoring the significance of video data in the recognition process. Moreover, the precision of the model is further improved

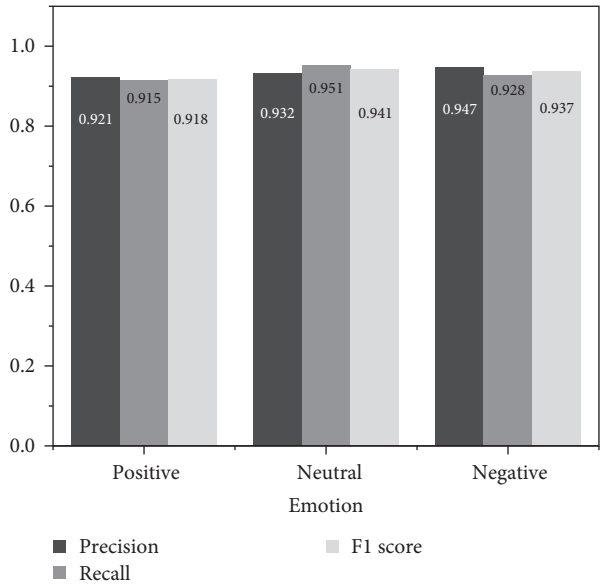

Figure 4: The result for the emotion classification.

after incorporating the attention mechanism. Recognition results for the three emotions are shown in Figure 4.

As shown in Figure 4, the emotion recognition precision achieved by the model in this paper is consistently above 90% for all emotions. Notably, the recognition accuracy for negative emotions is the highest, reaching an impressive 94.7%. The recognition accuracy for positive emotions is slightly lower but still meets practical usage requirements. Additionally, the F1 score, which balances precision and recall, also surpasses 90%, indicating a well-balanced recognition performance by the model.

We calculated the average values of the model's performance across the three metrics, and the results are shown in Table 1.

As shown in Table 1, the means of the three types of indicators are similar, which again illustrate the equilibrium performance of the model.

To provide a more intuitive representation of the fusion features' change patterns, we calculated the three types of indicators for the last 25 iterations and plotted the graph, with the precision indicator labeled. By observing the trend of the indicator changes, we can deduce that the precision indicator's data in the latter five iterations remain relatively flat, without showing significant improvement. However, over the course of training, the model's overall performance has shown improvement and attained a more balanced state. This highlights the efficacy of balancing the information from each modality through the attention mechanism, which is a crucial aspect of the model's effectiveness.

4.2. Comparison of Models with Different Features and Methods. After conducting a comprehensive comparative validation of the recognition accuracy of different features in the proposed model, we delved into further discussions regarding the recognition results obtained with various

Table 1: Result for the emotion classification in mean value.

Index	Mean
Precision	0.933
Recall	0.931
F1 score	0.932

classifiers and network structures. The outcomes of these experiments are shown in Figure 5.

As shown in Figure 6, focusing on the accuracy differences under different features, we observe that there is not a significant disparity in recognition results when using the audio feature and video feature independently, regardless of the method employed. However, when solely using the audio feature, the LSTM feature extraction method outperforms the GRU method. The model's advantages become evident when employing joint features, showcasing the efficacy of the proposed model in effectively utilizing related features and the attention mechanism to accomplish the emotion recognition task successfully. Further, in the model comparison, as shown in Figure 5, it becomes apparent that deep learning methods excel at leveraging the advantages of combined features, ultimately improving the overall recognition accuracy of the model.

4.3. The Practical Test Result. After completing the model construction and training analysis, we proceeded to conduct the actual test. As the previous training process utilized a public dataset as supplementary data, we required additional data collection for the actual test. To address this, we collaborated with an animation company to generate data meeting the specific data requirements. We then established labels through multiperson annotation to ensure a more accurate analysis during the actual test. For the actual test, emotions were still categorized into positive, neutral, and negative. The confusion matrix obtained from the actual test based on the model's predictions is shown in Figure 7.

As shown in Figure 7, we observe that the model's accuracy during practical application is slightly lower than the training test data. This is likely due to the presence of bias in the collected data, leading to results that are closer to real-world scenarios. In terms of recognition results, we find that the accuracy of positive emotion recognition is slightly lower than 90%, whereas neutral and negative expressions achieve recognition accuracies of more than 90%. Negative emotions usually have more expressions and vocabulary in the text. People tend to be more diverse in expressing negative emotions, while positive or neutral emotions may use relatively fewer vocabulary and expressions, making it easier for models to capture the characteristics of negative emotions. The misclassification mainly occurs in the positive emotion category, possibly because positive emotions are more subtle and require contextual semantic analysis for accurate recognition. This aspect could serve as a direction for further research in the future.

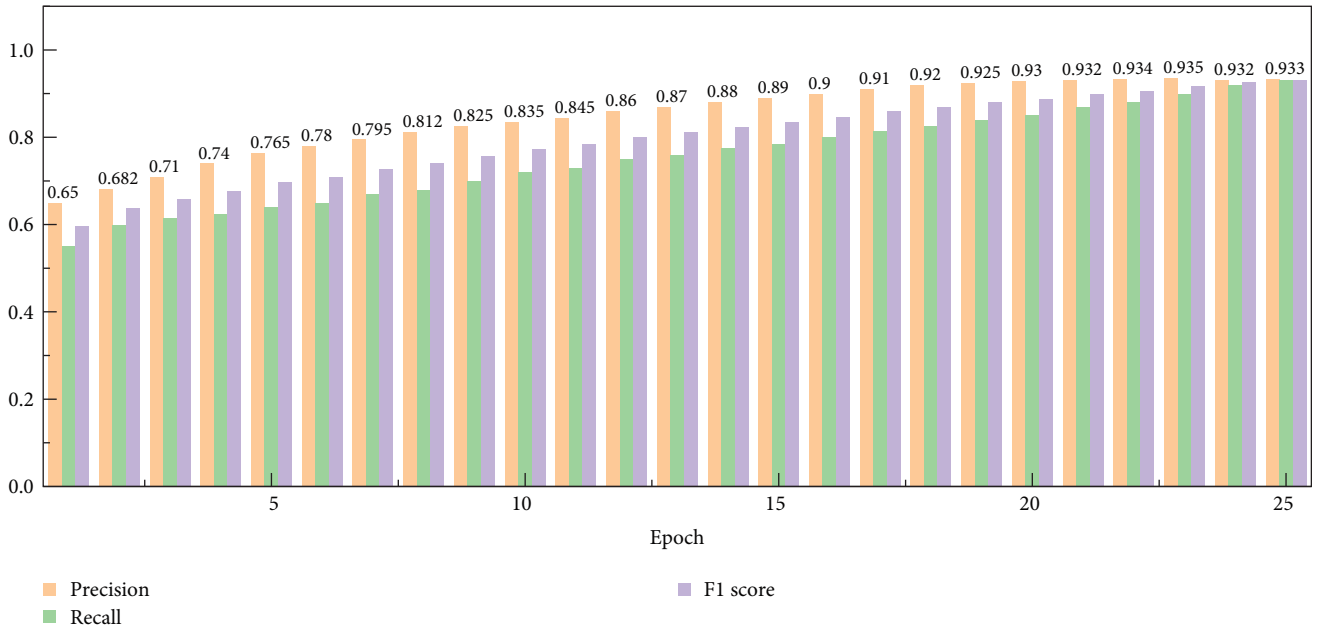

Figure 5: The index change trend in the training process.

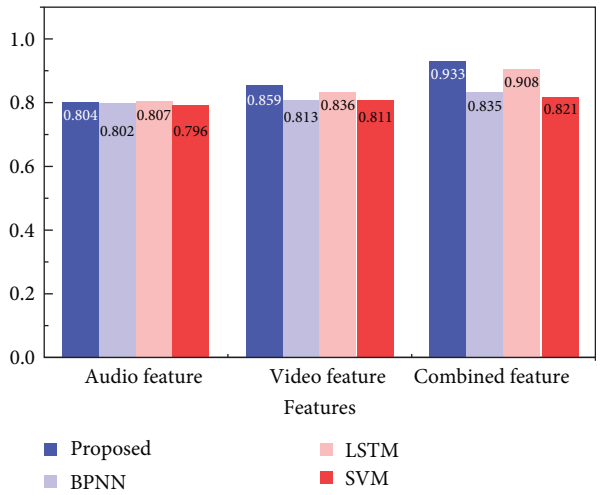

Figure 6: The result for different methods using different features.

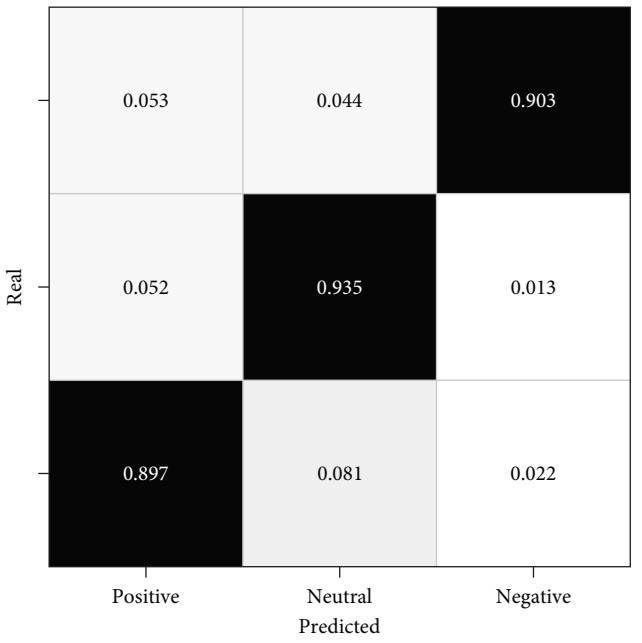

Figure 7: The confusion matrix for the different emotions in practical test.

5. Discussion

In today's era of rapidly growing video data, sentiment recognition in the animation industry plays a pivotal role for both animation creators and viewers. As a result, sentiment analysis in the field of animation has become a prominent research hot spot. In this study, we achieved successful fusion of audio and video image bimodal data and applied GRU as a feature extractor alongside the attention mechanism to accomplish the emotion recognition task. Comparing our approach with classical methods like BPNN and LSTM, we discovered notable advantages in our adopted method. The bimodal data fusion is a key highlight of our study. By combining audio and video image data, we capture the input data's features in a more comprehensive manner, consequently enhancing the performance of emotion recognition. This multimodal fusion approach compensates for the limitations present in single-modal data, yielding more accurate and comprehensive emotion classification results. Choosing GRU as the feature extractor was a well-considered decision. GRU, a RNN structure, exhibits strengths in processing sequential data. Compared to traditional RNNs and LSTMs, GRU has a simpler structure and fewer parameters, resulting in faster convergence and training in certain scenarios while retaining powerful modeling capabilities. With GRU as the feature extractor, we adeptly handle time-series data such as

audio and video images, thereby improving the accuracy of sentiment recognition. Furthermore, the introduction of the attention mechanism further enhances the model's performance. The attention mechanism dynamically learns and focuses on the most relevant parts of the input data, emphasizing crucial features. In bimodal fusion, we dynamically adjust the weights of the information based on the significance of different modalities for superior information fusion. This mechanism proves particularly vital in emotion recognition tasks, enabling the model to precisely capture key factors influencing emotional changes. In conclusion, the combination of bimodal fusion of audio and video images, GRU feature extraction, and the attention mechanism showcases significant advantages. Our method achieved satisfactory results in terms of accuracy, generalization performance, and time-series data processing when performing emotion recognition tasks. As a result, the method holds considerable potential for practical applications and shows promise in various domains, including sentiment analysis. It is important to note that in future research, we need to further optimize model training and tuning and validate its reliability and effectiveness on larger datasets. Additionally, exploring more ways of interaction between different modal data can further enhance the performance and robustness of sentiment recognition.

Emotional parsing of animation works through artificial intelligence methods enhances the viewing experience for animation viewers. Animation works often contain rich emotional elements, and emotional analysis allows viewers to gain a deeper understanding of the characters' emotions and the emotional logic behind their actions. This, in turn, enables viewers to connect with the storyline and characters on a more profound level, leading to increased emotional resonance and a stronger sense of love and identification with the animation work. For animation producers, emotional analysis serves as a valuable feedback tool. Producers face critical decisions in plot setting and characterization, and emotion analysis helps them gauge the audience's reception of emotional elements in the animation. By analyzing emotional parsing results, producers can gain insights into the audience's preferences and aversions to different emotional elements, empowering them to make precise adjustments and optimizations during the creation process. This, in turn, improves the overall quality and impact of animation works. Additionally, emotional analysis holds significant importance for learners studying animation production and creation. By understanding emotion analysis results, learners can grasp the intricacies and techniques of successful emotion expression in animation works. This is vital for developing learners' sensitivity to emotional expression and creative abilities. Through studying emotion analysis, learners can draw from the experiences of exceptional works, enabling them to better express and convey emotions in their own creations, making their works more compelling and captivating. Thus, refining the framework of emotion analysis for animation works and enriching the emotional connotation are pivotal for the future creation of animation works. Embracing emotion analysis in the animation industry has far-reaching benefits for both viewers and creators, fostering a deeper emotional connection between the audience and the animation works. In the future application process, using intelligent animation design methods, intelligent emotional analysis can complete corresponding work estimates before the animation goes public, improve the turnover efficiency of the animation market and people's viewing experience.

6. Conclusion

In this study, we have explored the application of GRU feature extraction and the attention mechanism for bimodal data fusion of audio and video images in the animation industry, specifically applied to the emotion recognition task. By comparing traditional methods like BPNN and LSTM, we have demonstrated the superiority of our proposed method in animation emotion recognition. Our approach leverages the correlation between multimodal information, such as audio and video images. With the GRU feature extractor, we efficiently handle time-series data, making it particularly suited for capturing emotions that involve temporal changes in animation works. Additionally, the attention mechanism enables automatic learning and focus on the most relevant parts of the input data, extracting and fusing crucial emotional features in a targeted manner. This amplifies the role of image features and enhances the model's accuracy in emotion recognition. The experimental results substantiate the effectiveness of our approach, with the multimodal feature fusion achieving recognition rates of 92.1%, 93.2%, and 94.7% for positive, neutral, and negative emotions, respectively. These results are detailed shown in Figures 4–7. These results not only offer robust data support for future multimedia animation emotion analysis and trainee training but also present innovative ideas for further research on multimodal fusion. By successfully applying GRU feature extraction and the attention mechanism to emotion recognition in animation works, we open up new possibilities for the animation industry and provide valuable insights into the potential of multimodal fusion in emotion analysis. This study contributes to the advancement of emotion recognition in animation and lays the foundation for future developments in this field.

Indeed, while this paper presents a promising approach for emotion recognition in animation works, there are some limitations that should be addressed. One of the key issues lies in the information fusion of multimodality, specifically with regard to audio data and facial expression data. The feature extraction methods for these two modalities are often based on prior knowledge, leading to potential loss of emotion information during the preprocessing stage. This can negatively impact the overall performance of the model. Future research should focus on employing more advanced feature processing methods for audio and facial expression modalities to enhance the quality of preprocessing features. By doing so, we can achieve more reliable and comprehensive emotion vector representations, resulting in improved recognition accuracy and performance. Furthermore, expanding the volume of available data and refining feature extraction methods are essential research directions for the future. Larger datasets

can provide a more robust foundation for training the model and can lead to better generalization capabilities. Additionally, continuously exploring and refining feature extraction methods can help to capture finer emotional nuances and make the model more adaptable to a wider range of animation works and emotions. By addressing these issues, the research in this area can advance further and provide more accurate and effective emotion recognition methods for animation works, leading to enhanced emotional experiences for both viewers and creators.

Data Availability

The data supporting the findings of this study are available upon reasonable request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication.

Acknowledgments

This work was supported by the National Social Science Foundation: "Goguryeo and Bohai" Research Special "Digital Protection and Utilization of Material Cultural Heritage in Shangcheng City of Bohai Country and its surrounding area-s"of Mudanjiang Normal University, the project number is 19VGH005 and also supported by the Fundamental Research Funds Project of Heilongjiang Education Department: "The Innovation Research and Practice of the '19th National Congress' Spirit Communication Form based on Digital media technology platform" of Mudanjiang Normal University, the project number is 1353MSYYB036.

References

[1] C. Kanellopoulou, K. L. Kermanidis, and A. Giannakoulopoulos, "The dual-coding and multimedia learning theories: film subtitles as a vocabulary teaching tool," *Education Sciences*, vol. 9, no. 3, Article ID 210, 2019.

[2] F. P. Rachmavita, "Interactive media-based video animation and student learning motivation in mathematics," *Journal of Physics: Conference Series*, vol. 1663, Article ID 012040, 2020.

[3] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, Article ID 81, 2021.

[4] R. M. Wong and O. O. Adesope, "Meta-analysis of emotional designs in multimedia learning: a replication and extension study," *Educational Psychology Review*, vol. 33, no. 2, pp. 357–385, 2021.

[5] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: a review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[6] W. Yu, H. Xu, F. Meng et al., "CH-SIMS: a chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Association for Computational Linguistics, 2020.

[7] R. W. Picard, *Affective Computing*, MIT Press, 2000.

[8] M. R. Amer, T. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, and S. Chai, "Deep Multimodal fusion: a hybrid approach," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 440–456, 2018.

[9] A. Ashfahani, M. Pratama, E. Lughofer, and Y.-S. Ong, "DEVDAN: deep evolving denoising autoencoder," *Neurocomputing*, vol. 390, no. 8, pp. 297–314, 2020.

[10] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[11] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[12] D. Bollegala, D. Weir, and J. Carroll, "Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 132–141, Association for Computational Linguistics, Portland, Oregon, USA, 2011.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86, Association for Computational Linguistics, 2002.

[14] X. Dong, D. Thanou, L. Toni, M. Bronstein, and P. Frossard, "Graph signal processing for machine learning: a review and new perspectives," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 117–127, 2020.

[15] F. Miwakeichi, E. Martínez-Montes, P. A. Valdés-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG data into space–time–frequency components using parallel factor analysis," *NeuroImage*, vol. 22, no. 3, pp. 1035–1045, 2004.

[16] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2892–2900, IEEE, Boston, MA, USA, 2015.

[17] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2012.

[18] H. T. P. Thao, D. Herremans, and G. Roig, "Multimodal deep models for predicting affective responses evoked by movies," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1618–1627, IEEE, Seoul, Korea (South), October 2019.

[19] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Association for Computational Linguistics, Copenhagen, Denmark, 2017.

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Association for Computational Linguistics, Florence, Italy, 2019.

[21] X. Shi, Z. Wang, H. Zhao et al., "Threshold-free phase segmentation and zero velocity detection for gait analysis using foot-mounted inertial sensors," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 176–186, 2022.

[22] M. J. Hamayel and A. Y. Owda, "A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms," *AI*, vol. 2, no. 4, pp. 477–496, 2021.

[23] S. Reza, M. C. Ferreira, J. J. M. Machado, and J. M. R. S. Tavares, "A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks," *Expert Systems with Applications*, vol. 202, Article ID 117275, 2022.

[24] C. Jin, C. Luo, M. Yan, G. Zhao, G. Zhang, and S. Zhang, "Weakening the dominant role of text: CMOSI dataset and multimodal semantic enhancement network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.