ISSN: 2688-8645

Research on Short Video Data Analysis Based on Multimodal Features

Yue Xie

Wenzhou Polytechnic, Wenzhou, Zhejiang, China

tse_0823@163.com

Abstract

As an emerging media format, short videos have become an important means for people to obtain information and entertainment. However, the complexity and diversity of short video content pose significant challenges for data analysis. This paper provides a reference for short video data analysis research by investigating multimodal data feature extraction, emotion recognition, and user interest prediction. Firstly, this study explores multimodal data feature extraction methods, utilizing deep learning models to extract image, audio, and textual features. Secondly, an emotion recognition method based on a user feature-guided attention mechanism is proposed, which enhances the feature representation of emotional analysis by fusing multimodal features. Finally, a user interest prediction model is designed by integrating multimodal features and user interest evolution patterns.

Keywords

Multimodal Features; Short Videos; Emotion Recognition; User Interest Prediction; Deep Learning.

1. Introduction

With the rapid development of Internet technology, short video platforms have gradually become an important medium for people to obtain information, entertain themselves, and socialize. Short videos have rich content, rapid dissemination, and strong interactivity, attracting a large number of users. In most cases, a short video contains information from multiple modalities, such as images, audio, and text. For example, the image modality in a video can reflect the visual content, while the audio modality conveys auditory information and the textual description provides more direct linguistic meaning. However, the diversity and complexity of short video content pose significant challenges for data analysis. Traditional analytical methods primarily focus on single-modality information, such as text or images, but these approaches fail to fully leverage the multimodal nature of short videos. Consequently, research on multimodal data analysis is crucial for enhancing user experience, optimizing content recommendations, and understanding user behavior.

In recent years, the fields of computer vision, natural language processing, and audio processing have made significant progress, providing technical support for the extraction and analysis of multimodal features in short videos. Meanwhile, the development of machine learning technologies has also provided powerful tools for processing complex short video data. Multimodal data analysis can fully utilize the multiple types of information in short videos, such as images, speech, and text, to provide more comprehensive and accurate analysis results.

Modality diversity implies diversity of information, but it also brings a large amount of noise interference and information redundancy. How to effectively obtain the consistency and complementary information among modalities and integrate them while preserving the original modality features to the greatest extent so that the integrated feature representation is

ISSN: 2688-8645

superior to that of a single modality is the difficulty and key of multimodal learning. It is also the focus of short video content understanding algorithm research.

As short video production continues to diversify, its application scenarios expand across various domains, including commercial advertising, educational knowledge sharing, and personal emotional expression. However, in practical applications, many organizations and individuals face the challenge of how to quickly and accurately understand the content of short videos. For example, news media need to monitor and interpret the emotional tendencies in short video news in real time to better guide public opinion. E-commerce platforms aim to optimize recommendation strategies by analyzing product demonstration videos. Educational institutions, on the other hand, wish to use short video analysis tools to assist in course design and teaching feedback.

However, existing analysis methods are generally inefficient and imprecise, which seriously limits the practical value of short video data. To address these challenges, an efficient and reliable short video data analysis method is urgently needed. By thoroughly analyzing multimodal features of short videos, platforms can gain deeper insights into video content, thereby refining recommendation algorithms to deliver more personalized and engaging content. This not only improves user experience and retention but also enhances platform activity and stickiness. Furthermore, multimodal-based analyses can provide creators with actionable guidance, fostering the production of high-quality content and elevating overall platform value. Additionally, such methods enable more effective identification and filtering of inappropriate content, ensuring healthy platform operations.

2. Literature Review

In the field of short video data analysis, accurate video feature extraction and representation serve as pivotal components, significantly influencing the precision of subsequent emotional analysis and user interest prediction models. Videos inherently encompass multimodal information, including static images, dynamic behavioral patterns, and audio signals. Feature extraction across these modalities often operates independently, with numerous scholars conducting research on feature extraction methodologies for different data modalities [1].

Textual feature extraction and representation involve quantifying word vectors derived from textual content to enhance computational recognizability [2]. Acoustic information conveyed through speech includes both semantic and phonetic features, with the latter containing critical auxiliary semantic information. Notably, audio modalities encompass diverse data types, each carrying distinct information. The Mel spectrogram, a feature inspired by human auditory perception, demonstrates superior performance in characterizing energy distribution patterns when adapted from speech recognition applications [3].

Image classification constitutes a foundational task in computer vision, evolving from early grayscale handwritten digit recognition (e.g., the 10-class MNIST dataset) to more complex benchmarks like CIFAR-10 and CIFAR-100. LeNet-5, a pioneering convolutional neural network (CNN), introduced a hierarchical architecture comprising input, convolutional, pooling, and fully connected layers. By leveraging local spatial correlations in images instead of traditional fully connected architectures, LeNet-5 significantly advanced neural network development [4]. ResNet addressed the degradation problem in deep networks through residual modules, enabling the training of profound network architectures [5]. Building upon ResNet, DenseNet enhances feature propagation and reuse by connecting each layer to all subsequent layers in the forward pass, thereby aggregating feature maps from all preceding layers while reducing parameter counts [6].

Multimodal data processing primarily involves three stages: feature extraction, feature fusion, and task modeling. During the feature extraction phase, specialized methods are typically

ISSN: 2688-8645

designed for each modality. For instance, convolutional neural networks (CNNs) are employed to extract local features from image data, while word embedding techniques like Word2Vec or BERT are utilized to capture semantic information in textual data. For audio data, Mel spectrograms combined with long short-term memory networks (LSTMs) are effective in capturing temporal features. The feature fusion stage focuses on integrating information from diverse modalities. Traditional approaches, such as simple feature concatenation or weighted averaging, often overlook the inherent correlations and heterogeneities between modalities. In recent years, attention mechanism-based fusion methods have gained significant attention. The MIFUIL model, proposed by scholars, employs an attention mechanism to learn the complementarity and correlations between different modalities, enabling efficient fusion of user multimodal information [7]. The Transformer architecture has also demonstrated exceptional performance in multimodal data fusion due to its strong capacity for modeling long-range dependencies. Researchers have developed the BRCTN model, which integrates textual, visual, and auditory information using Transformer networks, achieving an average accuracy improvement of approximately 15% compared to traditional methods on public datasets [8]. The task modeling stage aims to address specific downstream applications, such as sentiment analysis, object detection, and user behavior prediction. Morency LP et al. pioneered the use of visual, vocal, and textual modalities for sentiment analysis, introducing the first multimodal YouTube dataset. Veronica Perez-Rosas et al. advanced this work by conducting sentence-level sentiment analysis based on visual, vocal, and textual cues [9]. Chen Peiwen et al. constructed sentiment lexicons, employed statistical methods for feature selection, and utilized support vector machines (SVMs) for sentiment classification [10]. Ghosal D et al. combined LSTMs with attention mechanisms to capture emotional relationships between contextual utterances in videos, incorporating multimodal contextual dependencies to enhance the robustness of sentiment analysis systems [11]. Li Tingting highlighted that short video news achieves deep multimodal discourse integration through "parallel expression," blending emotional and rational elements using techniques like emotional metaphor, social narration, and scene construction [12]. This emotion-driven approach not only enhances information transmission effectiveness but also strengthens audience emotional resonance. Despite recent advancements in multimodal sentiment analysis, research integrating visual, vocal, and textual modalities remains relatively limited, and fusion models often lack robustness, necessitating further investigation.

Multimodal data processing also faces several challenges. Heterogeneity in data complicates feature extraction and fusion processes, while high annotation costs limit the availability of large-scale training datasets. Additionally, the poor interpretability of models, particularly in sensitive applications, may restrict their deployment. However, ongoing advancements in hardware computational capabilities and algorithmic optimizations are progressively mitigating these issues.

3. Research Methods

3.1. Multimodal Feature Extraction Methods

In multimodal feature extraction, we first identify the various modalities in short videos, such as text, images, and audio. Each modality has unique features and interacts with others in complex ways. Scientific methods are needed to extract and integrate these features effectively.

For text features, word embedding technology is commonly used. It converts text into continuous vector representations in a low-dimensional space, capturing semantic relationships. Pre-trained language models like BERT or RoBERTa can encode subtitles, generating high-quality text feature vectors. Fig. 1 shows the architecture diagram of the BERT

ISSN: 2688-8645

model. Specifically, given a text sequence of length T, {x_t} where t=1,...,T, BERT processes it to produce context-aware word embedding vectors {h_t}.

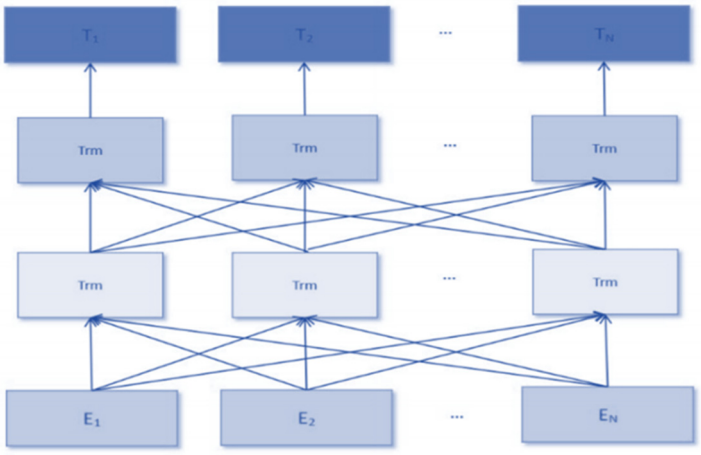

Fig 1. BERT model

Image feature extraction relies mainly on convolutional neural networks and self-attention mechanisms (SAM). CNNs excel at extracting local features from pixel matrices, while SAMs model global dependencies. An improved CNN-SAM hybrid model based on ResNet50 can be used. This model first extracts initial features with ResNet50, then enhances them with a SAM module.

Audio feature extraction is more complex due to multiple dimensions like frequency and duration. As shown in Figure 2, we can convert audio signals into Mel-spectrograms and model them with long short-term memory (LSTM) networks, which capture long-range dependencies. To boost robustness, dropout regularization is added to prevent overfitting.

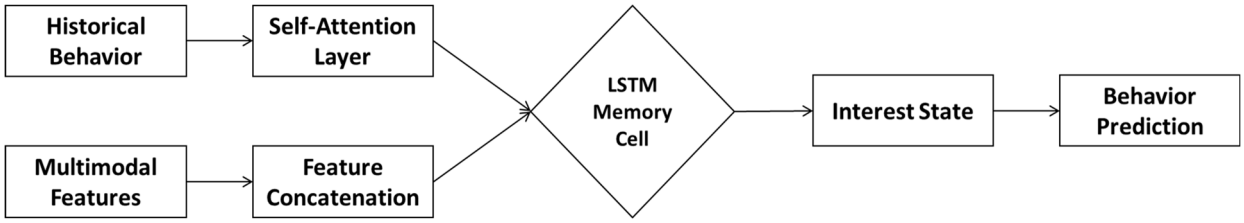

Fig 2. Self-Attention-driven Interest Modeling

Besides the three basic modalities, their interactions are also important. Multimodal feature fusion is key here. The main fusion strategies are early fusion and late fusion. Early fusion combines features from different modalities at a low level before further processing, while late fusion processes each modality separately and combines the results at a high level.

3.2. Construction of Data Analysis Model

In this study, to verify the effectiveness of the proposed short video data analysis method based on multimodal features, we designed a series of experiments to evaluate the model's performance on different tasks. The experimental design focuses on the following aspects: dataset selection and preprocessing, experimental environment configuration, model training strategies, and evaluation metric selection.

We chose a representative short video dataset for experiments. It consists of short videos from social media, featuring high diversity and complexity. The dataset includes about 50,00 short videos covering entertainment, news, education, etc., ensuring experimental diversity and a rich testing environment.

ISSN: 2688-8645

For the experimental environment, we used a high performance computing platform with the latest GPU - accelerated hardware to support large scale data processing and efficient deep learning model training. The operating system is Linux, the programming language is Python, and key libraries include mainstream machine learning frameworks like TensorFlow and PyTorch. To ensure experimental consistency and reproducibility, all experiments were conducted in the same environment, with system checks before each experiment to prevent external interference.

In terms of model training strategies, we adopted transfer learning, applying pre - trained deep - neural - network models to short video data analysis. Specifically, we used ResNet - 50 for image feature extraction, VGGish for audio features, and the BERT model for text processing. This approach allows quick adaptation to new tasks while reducing training time and resource consumption. To avoid overfitting, we incorporated Dropout regularization and Early Stopping. Iterations stop when the validation - set loss ceases to decline, yielding optimal model parameters.

When selecting evaluation metrics, we considered multiple aspects to comprehensively reflect the model's performance. First is accuracy, a common and intuitive metric for measuring the match between predicted and actual labels. Then there are precision and recall, which reflect the model's ability to correctly predict positive samples and cover all actual positive samples. These are especially useful for imbalanced class distributions. The F1 score, the harmonic mean of precision and recall, balances the two metrics. Additionally, we introduced the area under the ROC curve (AUC) to assess the model's overall performance across different thresholds.

To further validate the model's robustness and generalization ability, we conducted cross - validation. The dataset was divided into k mutually exclusive subsets. In each iteration, one subset served as the test set while the remaining subsets formed the training set. This process was repeated k times until each subset had been used as the test set. This method provides a more accurate estimate of the model's performance on unseen data.

4. Conclusion

4.1. Research Summary

This study systematically explored and implemented multimodal feature-based short video data analysis methods to address challenges in efficient multimodal feature fusion and deep analysis within the short video domain. Research findings demonstrate that, given the current high diversity and multimodal nature of short video content, adopting multimodal feature fusion techniques significantly enhances the analytical precision and applied value of short video data. Through comprehensive analysis of existing literature and research outcomes, it is evident that traditional single-modality approaches often fail to fully capture core features of short video content, whereas multimodal fusion methods exhibit clear advantages in tasks such as tag classification, sentiment analysis, and user behavior prediction.

Specifically, this study began with multimodal feature extraction and proposed a BiLSTM-Attention network-based multimodal feature fusion model. By aligning and integrating visual, auditory, and textual modalities, the model achieves precise representation of short video content. Research indicates that combining textual, visual, and auditory features for sentiment recognition not only enables more comprehensive understanding of emotional tendencies in short videos but also improves sentiment classification accuracy. In practical applications of multimodal fusion algorithms, the study identified correlations and complementarities between different modalities as critical factors affecting model performance. Through detailed analysis of feature interactions, a dynamic weight allocation strategy was developed to adaptively adjust contribution weights based on modality importance.

ISSN: 2688-8645

This study achieved significant progress in multimodal feature extraction, fusion, and analysis, providing new perspectives for short video data analytics while establishing foundational frameworks for broader multimodal applications. Future research will further expand the scope of multimodal fusion technologies to achieve breakthroughs in additional domains.

4.2. Research Innovations and Limitations

This study made substantial contributions to multimodal short video data analysis while identifying certain limitations. In terms of innovations, the proposed methodology integrates textual, visual, and auditory modalities for comprehensive short video content analysis. By incorporating convolutional neural networks (CNNs) and recurrent neural networks (RNNs), we successfully extracted visual and acoustic features, which were then combined with textual information to construct composite feature vectors. Additionally, an attention mechanism-based multimodal fusion framework was developed to dynamically allocate weights according to modality significance.

Despite the significant results achieved, there are also some limitations. Current research primarily focuses on static feature extraction with limited attention to dynamic characteristics. For instance, existing methods often neglect temporal dependencies between consecutive video frames, leading to suboptimal analysis accuracy. Future work could explore improved utilization of temporal information to enhance dynamic feature extraction. Furthermore, while current multimodal fusion methods improve accuracy, computational efficiency remains a challenge. Potential solutions include integrating parallel computing technologies or optimization algorithms to address this limitation.

4.3. Future Prospects

The rapid evolution of short video platforms has broadened the applicability of multimodal features in data analytics. This study addresses current challenges in short video complexity through multimodal fusion but acknowledges unexplored opportunities for further advancement.

In multimodal feature extraction, most existing research focuses on isolated modalities rather than synergistic interactions. For example, the complementary relationships between text, images, and audio remain underexploited. Future studies could develop advanced joint extraction algorithms that preserve modality-specific characteristics while maximizing inter-modality correlation capture.

For sentiment analysis tasks, varying emotional expression weights across modalities necessitate dynamic fusion strategies beyond traditional fixed-weight approaches. Potential directions include adaptive weight adjustment mechanisms or cross-modal knowledge transfer techniques to enhance model adaptability through domain-agnostic knowledge sharing.

Multimodal short video analytics holds substantial future potential. By refining feature extraction, fusion, and application strategies, practical requirements across scenarios can be better met to drive technological advancement. Concurrently, attention to data security and privacy protection remains critical to ensure ethical technology development. These research avenues will enrich theoretical frameworks and technical toolkits in multimodal analytics, injecting new vitality into the short video industry.

Acknowledgments

This paper was supported by Project R20240063 funded by the Wenzhou Science and Technology Bureau.

ISSN: 2688-8645

References

[1] REN Z Y, WANG Z C, KE Z W, et al. Survey of multimodal data fusion[J]. Computer Engineering and Applications, 2021,57(18): 49-64.

[2] Yao T, Zhai Z, Gao B. Text Classification Model Based on fastText[C]. 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), 2020: 154-157.

[3] Atliha V, Sesok D. Comparison of VGG and ResNet used as Encoders for Image Captioning[C]. 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020: 1-4.

[4] Henaff M , Bruna J , Lecun Y .Deep Convolutional Networks on Graph-Structured Data[J].Computer Science, 2015: 1-10.

[5] He K , Zhang X , Ren S ,et al.Deep Residual Learning for Image Recognition[J].IEEE, 2016.

[6] Huang G , Liu Z , Laurens V D M ,et al.Densely Connected Convolutional Networks[J].IEEE Computer Society, 2016.

[7] Fan Y, Zhou Q, Chen W, et al. User connection method based on multimodal information fusion [J]. Computer Engineering and Design, 2024, 45(9): 2641-2648.

[8] Xie X, Ding C, Wang X, et al. Multimodal emotion recognition integrating text, speech, and expression [J]. Journal of Qingdao University (Engineering & Technology Edition), 2024, 39(3): 20-30.

[9] Zadeh A, Zellers R, Pincus E. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv:1606.06259, 2016.

[10] Chen P, Fu X. Research on text sentiment polarity classification using SVM method [J]. Journal of Guangdong University of Technology, 2014, 31(3): 95-101.

[11] Ghosal D, Akhtar M, Chauhan D. Contextual inter-modal attention for multi-modal sentiment analysis [C]. Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 3454-3466.

[12] Li T. Multimodal discourse analysis of short video news driven by emotion [J]. News Lovers, 2024, (12): 28-30.