MMSF: A Multimodal Sentiment-Fused Method to Recognize Video Speaking Style

Beibei Zhang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
zhangbb@smail.nju.edu.cn

Yaqun Fang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
fangyq@smail.nju.edu.cn

Fan Yu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
yf@smail.nju.edu.cn

Jia Bei
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
beijia@nju.edu.cn

Tongwei Ren*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rentw@nju.edu.cn

ABSTRACT

As talking takes a large proportion of human lives, it is necessary to perform deeper understanding of human conversations. Speaking style recognition is aimed at recognizing the styles of conversations, which provides a fine-grained description about talking. Current works focus on adopting only visual clues to recognize speaking styles, which cannot accurately distinguish different speaking styles when they are visually similar. To recognize speaking styles more effectively, we propose a novel multimodal sentiment-fused method, MMSF, which extracts and integrates visual, audio and textual features of videos. In addition, as sentiment is one of the motivations of human behavior, we first introduce sentiment into our multimodal method with cross-attention mechanism, which enhance the video feature to recognize speaking styles. The proposed MMSF is evaluated on long-form video understanding benchmark, and the experiment results show that it is superior to the state-of-the-arts.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Speaking style recognition, multimodal analysis, sentiment analysis, long-form video understanding

ACM Reference Format:
Beibei Zhang, Yaqun Fang, Fan Yu, Jia Bei, and Tongwei Ren. 2023. MMSF: A Multimodal Sentiment-Fused Method to Recognize Video Speaking Style.

*Corresponding author.

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592219

In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592219

1 INTRODUCTION

Each person on average takes part in talkings of 14,878 words each day, which lasts about an hour and fifteen minutes. This is what happened in a case study about human conversations [21], which proves that talking accounts for a large proportion in human daily lives. To further understand human activities, it is necessary to conduct deep studies on human conversations. Speaking style recognition (SSR) aims to identify the conversation styles among characters in videos. As shown in Figure 1, speaking styles can reflect the state of human conversations. Recognizing speaking styles enables us to understand "how" the characters talk rather than just "talk", which helps us to construct a more fine-grained understanding of human conversations in videos.

In addition to contributing to more accurate understanding of conversations among characters, speaking style recognition can further assist other video understanding tasks by providing speaking

Figure 1: Examples of different speaking styles of human conversations.

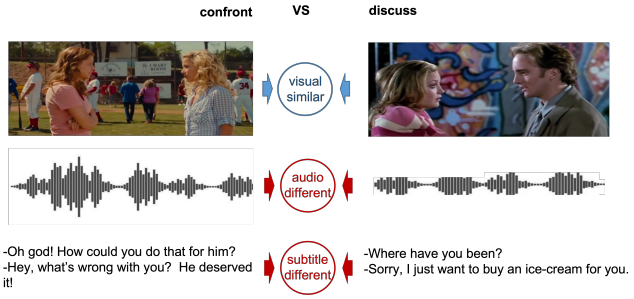

Figure 2: An example of comparison among different speaking styles on different modalities.

style contexts, such as video description generation, visual question answering and video retrieval.

Most of the solutions to SSR focus on analyzing videos from the visual perspective, such as ViS4mer [15] and STAN [9]. However, it is of high possibility that videos of different speaking styles seem similar because most of them are photographed in the form of "someone talking to the other one". Speaking style, as an attribute of human conversation, necessarily has a strong correlation with audio and text. Especially when visual differences are not obvious, audio and text can provide more concrete information. For example in Figure 2, two speaking styles, confront and discuss, both visually seem like talking between a couple, but they are obviously different in the audiograms. At the same time, contents of their conversations are different. Compared to discuss, people tend to speak more modal particles such as "Hey!" when speaking style of their conversation is "confront". Therefore, we design a multimodal pipeline to predict the speaking styles by integrating visual, audio and textual features, which helps distinguish similar speaking styles more effectively.

Current works on SSR conduct analysis on movie clips. A movie clip is a segment cut from a movie and records a complete plot, which lasts one to three minutes. It is hard to analyze a movie clip because it contains various information. It is required to capture feature expressions of high distinctiveness to help recognize speaking styles. Sentiment is one of the motivations of human behavior, which has a close relationship with the style of human conversations. As shown in Figure 3, people whose sentiments are anger and fear are more likely to be associated with speaking style

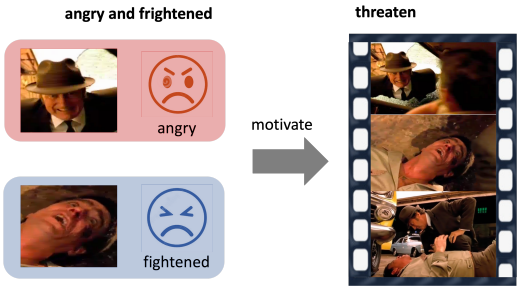

Figure 3: An example of showing the relationship between speaking styles and sentiments of characters.

like "threaten" rather than other speaking styles like "teach". It is obvious that the sentiments of characters can help identify different speaking styles. As a result, we propose MMSF, a multimodal sentiment-fused method, which introduces sentiment influence into multimodal pipeline to enhance video feature.

Our methods firstly extracts multimodal and sentiment features of each modality of movie clips. It then fuses multimodal and sentiment features of corresponding modalities with the help of cross-attention mechanism. Multimodal sentiment-fused features is finally used to recognize the speaking style of movie clips.

We evaluate MMSF on long-form video understanding (LVU) dataset [24], where SSR is firstly proposed. The performance of MMSF achieves the best one.

In general, the main contributions of our work can be summarized as follows: (1) We propose a multimodal pipeline, which makes use of vision, audio and text modalities, to identify the speaking style of conversations between characters in movie clips. (2) We first introduce the effect of sentiment into multimodal method by calculating cross attention between sentiment and common multimodal features.

2 RELATED WORK

2.1 Multimodal Video Analysis

Early video classification mainly use Convolutional Neural Networks (CNN), which is highly effective at video classification tasks but are computationally intensive to train. To address computational overhead in training convolutional video networks, Feichtenhofer *et al.*[8] propose an efficient CNN to improve the performance of video classification, which utilizes a fast and slow temporal sampling stream. Recently, some transformer-based video classification models have been proposed to work effectively for video classification, such as ViViT [2], Video Swin Transformer [17] and MeMViT [25]. Most of these methods tend to model the visual spatial and temporal correlation of videos, which is of high effectiveness in video analysis tasks that are closely related to sequential information, such as action recognition. However, to understand the storyline of movie clips, it is necessary to pay more attention to all the behaviors of characters, including their actions and words. So multimodal analysis is necessary.

There have been many multimodal video models so far. Multimodal Transformer [23] proposes the directional pairwise cross-modal attention to resolve alignment and dependencies between different modalities. Videobert [22],which is a joint visual-linguistic model, is built upon the BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens. VideoAudio-Text Transformer [1] takes raw signals as inputs and extracts multimodal representations for video analysis. I-code [27] is a framework where users may flexibly combine the modalities of vision, speech, and language into unified and general-purpose vector representations.

2.2 Video Sentiment Analysis

There are also some models of sentiment analysis that focus on the design of single-modal features. Xu *et al.* [26] used a pre-trained CNN as a provider of high-level visual attribute descriptors in order to train two sentiment classifiers based on logistic regression.

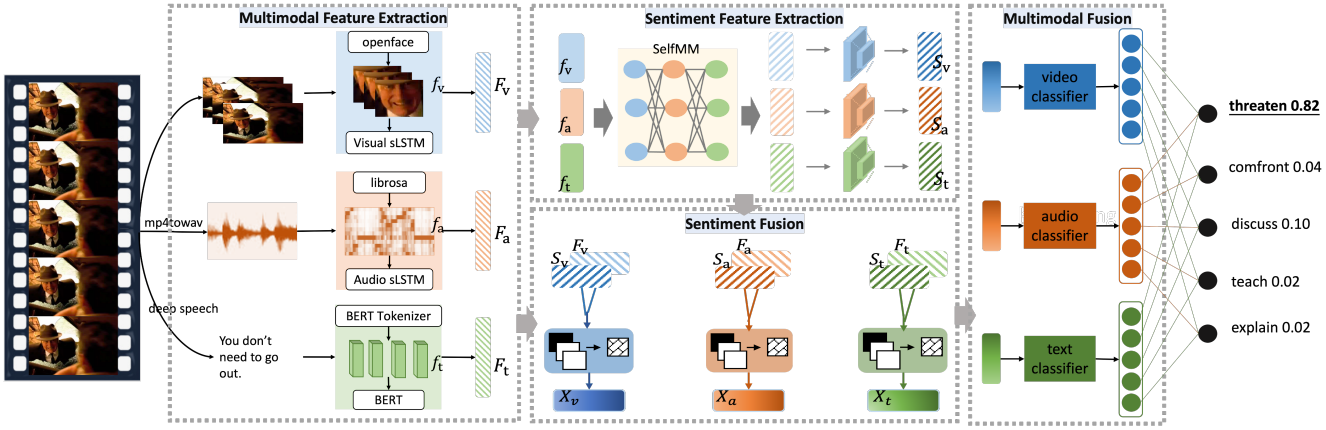

Figure 4: An overview of MMSF. Here, v, a and t are original visual, audio and textual features; f_v, f_a and f_t are multimodal features which have been updated with timing information; s_v, s_a and s_t are multimodal sentiment features; v_s, a_s and t_s are multimodal sentiment-fused features.

Progressive CNN [28] proposed to use a progressive approach for training a CNN in order to perform visual sentiment analysis. Cao *et al.* [5] focused on textual sentiment analysis where only words are used to analyze the sentiment, ignores the interdependencies and relations among the utterances of a video.

Multimodal sentiment analysis is a developing area of research. Morency *et al.* [19] aim at identifying the sentiment of a speaker by gaining clues from multimodal signals, including textual, visual and acoustic channels. Generally, different modalities are often complementary to each other, providing extra cues for semantic and sentimental disambiguation [20]. MISA [13] projects each modality to two distinct subspaces, and provides a holistic view of the multimodal data, which is used for fusion that leads to task predictions. Multimodal-informax [11] synthesizes fusion results from multi-modality input through a two-level mutual information maximization. BBFN [10] learns two text-related pairs of representations, which is text-acoustic and text-visual, enforcing each pair of modalities to complement mutually. Compared with single-modal sentiment analysis, multi-modal sentiment analysis can integrate more contextual information and achieve better performance.

Although sentiment is one of the inspirations of human behaviors, overall information is acquired to analyze the comprehensive content of a video. MMSF firstly extracts spatial-temporal features of videos, which are then integrated with sentiment features by cross-attention encoder as the final classification features.

2.3 Long-Form Video Understanding

LVU [24] benchmark is proposed to understand the full picture of a movie clip. Wu *et al.* [24] collect movie clips from publicly available MovieClips [29] and introduce the LVU benchmark, which contains nine diverse tasks covering a wide range of aspects of long-form video understanding. The nine tasks are of three types: *Relationship*, *Scene/Place* and *Speaking Style Classification* are of content understanding type; *"Like" Ratio* and *"Popularity" Prediction* is of user engagement type; *Director*, *Writer*, *Genre* and *Release Year Recognition* are of movie metadata type.

There are some work studying on LVU, SSR certainly included. The Object Transformer [24] adopts a Transformer architecture and a variety of external modules to handle LVU tasks. Recently, ViS4mer [15] is proposed, which is an efficient long-range video model that combines the strengths of self-attention and the structured state-space sequence layer. STAN [9], which models dependencies between static image features and temporal contextual features using a two-stream transformer architecture, achieves good performance on LVU tasks. Movie2Scenes [6] uses all available information to generate a general-purpose scene-level representation by contrastive learning.

These methods are designed to deal with all tasks of LVU benchmark. However, tasks of LVU are of totally different types. Some focus on contents while others pay attention to metadata of videos. MMSF pay more attention to the characteristics of tasks, rather than applying the same processing steps to quite different tasks. Due to speaking style is an attribute of conversations among characters, MMSF not only takes multimodal information into consideration but also introduces sentiment features to help recognition.

3 METHOD

As shown in Figure 4, the input of MMSF is a movie clip and the output is the speaking style classification results. Original visual, audio and textual features are firstly extracted with the help of feature extraction tools, and then they are respectively input into video LSTM, audio LSTM and BERT to model temporal information. At the same time, a pre-trained multimodal sentiment recognition model is applied to obtain sentiment features of different modalities. Sentiment features are then input to processing layers to obtain the most related sentiments and filter out irrelevant noise. After that, multimodal features are fused with the corresponding sentiment features with the help of cross-attention encoder. Multimodal classification results are obtained after handling multimodal sentiment-fused features with multiple classifiers. Multimodal results are integrated by late fusion strategy with different weights to generate the final recognition results.

3.1 Multimodal Feature Extraction

Movie clips are firstly transformed into frames, which are then input to openface [3] to generate original visual features f_v. Audio is extracted from videos by ffmpeg command tools [4], and then are input to librosa [18] to obtain original audio features f_a. As for text, subtitle generation tools, which are Deep Speech [12] and YouTube Transcript API [16], are applied to generate subtitles of movie clips. A pretrained BERT tokenizer is then used to tokenize the speech texts to obtain the corresponding tokens f_t.

To extract timing information of multimodal features, f_v and f_a are input to single directional Long Short-Term Memory models (sLSTM) [14], which helps generate visual features F_v and audio features F_a that bring temporal information. Since NLP models have made great success, a BERT [7] is applied to extract the sequential features of texts from f_t. We choose the outputs of the last layer as textual representations F_t:

$$F_i = \mathcal{F}_{sLSTM}\left(f_i; \theta_i^{lstm}\right), i \in \{v, a\}, \tag{1}$$

$$F_t = \mathcal{F}_{BERT}\left(f_t; \theta^{bert}\right), \tag{2}$$

where $\mathcal{F}_{sLSTM}$ and θ^{lstm} represent LSTM network and its parameters, $\mathcal{F}_{BERT}$ and θ^{bert} represent BERT network and its parameters, v, a and t refer to vision, audio and text modalities, respectively.

3.2 Sentiment Feature Extraction

We apply a multimodal sentiment analysis model, SelfMM [30], which focuses on the information complementation among different modalities, as our multimodal sentiment feature extraction model. The product of penultimate layer before final result will be taken as sentiment features:

$$S_i' = \mathcal{F}_{SelfMM}\left(f_i; \theta^{selfmm}\right), i \in \{v, a, t\}, \tag{3}$$

where $\mathcal{F}_{SelfMM}$ and θ^{selfmm} represent the network and pretrained parameters of SelfMM, v, a and t refer to vision, audio and text modalities.

The sentiment model we used is pretrained on sentiment datasets. Considering that not all sentiments detected are related to the target speaking style, it is necessary to pay more attention to those effective ones and filter out noise. As a result, a 1D convolutional layer followed by a max pooling layer is used to process the outputs of SelfMM to enhance sentiment features:

$$S_i = g(\phi(S_i')), i \in \{v, a, t\}, \tag{4}$$

where ϕ is a 1D convolutional layer and $g(\cdot)$ is a max pooling operator, and v, a and t refer to vision, audio and text modalities.

3.3 Sentiment Fusion

We design a cross-attention encoder to fuse multimodal features and their corresponding sentiment features. The cross-attention encoder consists of multiple layers and each attention layer consists of multiple attention heads. Figure 5 shows the multi-head attention layer of cross-attention encoder. The multiplication of query and key represents the similar map between them. Then the other multiplication operation on the similar map and value will generate the attention map which is from key/value to query. As a result, when key, value and query are set from different data sources, the effect of one source can be made to the other one by generating attention map and adding it to the target data source.

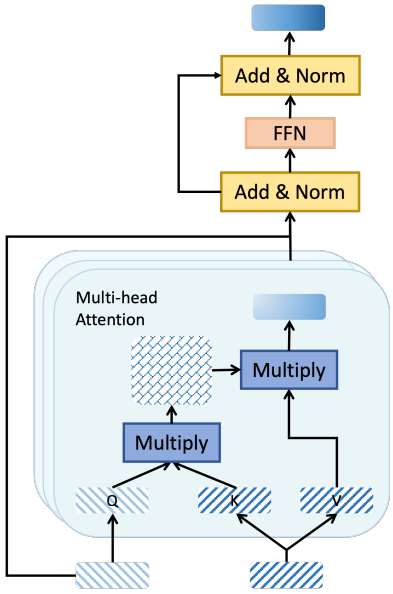

Figure 5: The illustration of the cross-attention encoder we used to fuse sentiment features and multimodal features. Here, Q, K and V refer to query, key and value; Multiply, Add, Norm respectively denote matrix multiplication, matrix addition and layer normalization operations; FFN represents a feed forward network.

For example, to pass the influence of visual sentiment features S_v to visual temporal features F_v, in each attention head, the attention from S_v to F_v can be calculated as follows:

$$Q = F_v \times W_q + b_q, \tag{5}$$

$$K = S_v \times W_k + b_k, \tag{6}$$

$$V = S_v \times W_u + b_u, \tag{7}$$

$$A_j = \text{softmax}\left(\frac{Q_j^{\mathrm{T}} \times K_j}{\sqrt{d_j}}\right) \times V_j^{\mathrm{T}}, j \in \{1, ..., n\}, \tag{8}$$

where W_q, b_q, W_k, b_k, W_u and b_u are the weights and bias of linear layers in each head to transform multimodal and sentiment features to key, value and query, n is the number of attention heads, T is a matrix transposition operation, d is scaled factor. Then the attention outputs of all heads will be fused together by concat and a linear layer will be used to update them:

$$A = (A_1^{\mathrm{T}} \oplus A_2^{\mathrm{T}} \oplus ... \oplus A_n^{\mathrm{T}}) \times W_o + b_o, \tag{9}$$

where n is the number of attention heads, $\oplus$ is a concat operation, T is a matrix transposition operation, W_o and b_o are the weights and bias of the linear layer to update the concat of attention maps.

Sentiment-fused features are obtained after adding the attention map A to the visual features F_v. The layer normalization is then conducted to process the original sentiment-fused features. Finally, a feed forward network is applied to perform updates and generate

the final outputs X_v:

$$X'_v = \mathcal{F}_{LN}(A + F_v), \tag{10}$$

$$\begin{aligned} X''_v &= \mathcal{F}_{FFN}(X'_v) \\ &= \max(0, X'_v \times W_1 + b_1) \times W_2 + b_2, \end{aligned} \tag{11}$$

$$X_v = \mathcal{F}_{LN}(X''_v + X'_v), \tag{12}$$

where $\mathcal{F}_{LN}$ is the layer normalization operation, $\mathcal{F}_{FFN}$ is the feed forward network, W_1, b_1, W_2 and b_2 are the weights and biases of the feed forward network. In this way, sentiment-fused multimodal features X_v, X_a and X_t are finally obtained.

In order to explore the performances of different sentiment fusion methods, we also take dot product of multimodal features and sentiment features as another fusion strategy. Taking visual feature as an example:

$$\widetilde{X_v} = \phi_f(F_v) \cdot \phi_s(S_v), \tag{13}$$

where F_v are visual temporal features, S_v are visual sentiment features, *cdot* is a dot product operation, ϕ here is aimed at aligning the dimensions of temporal and sentiment features by a convolutional layer. The comparison between using cross-attention encoder and direct dot product to fuse sentiments can be seen in the ablation experiments in shown in Section 4.2.

3.4 Multimodal Fusion

Generally, early fusion strategies, such as concat, are used as the final fusion strategy among multimodal features. As shown in Equation 15, fusing all features by concat and then adding a softmax classifier to get final classification results:

$$C = X_v \oplus X_a \oplus X_t, \tag{14}$$

$$y = \text{softmax}(\mathcal{R}(C \times W + b)), \tag{15}$$

where X_v, X_a and X_t are sentiment-fused visual, audio and textual features, $\oplus$ is a concat operation, $\mathcal{R}(\cdot)$ is a relu activation operation, W and b are the weights and bias of the linear layer in the classifier.

However, there are great differences among features of different modalities. The differences may become even more obvious after fusing multimodal sentiment features. As a result, a late fusion strategy is applied to merge multimodal classification results, which are generated independently of each other. Since different modalities have different effects on the SSR, after getting multimodal results from corresponding softmax classifiers, we multiply the multimodal results and different modality weights k_i to obtain the final modal-weighted classification result as follows:

$$y_i = \text{softmax}(\mathcal{R}(X_i \times W_i + b_i)), i \in \{v, a, t\}, \tag{16}$$

$$y_c = k_v y_v + k_a y_a + k_t y_t, \tag{17}$$

where $\mathcal{R}(\cdot)$ is a relu activation operation, X_i are sentiment-fused multimodal features, W_i and b_i are weights and bias of the classifier of each modality, v, a and t refer to vision, audio and text modalities. Early fusion and late fusion strategy are compared in the ablation experiments, which can be seen in Section 4.2.

3.5 Loss Function

Due to SSR is a classification task, we apply cross entropy loss as our basic loss function:

$$\mathcal{L} = -\sum_{i=1}^{n} h_i \log(y_i), \tag{18}$$

where n is the number of categories, h is one hot representation of ground truth label of each sample and y is the classification result, which consists of the probability of each speaking style category. When we apply late fusion strategy, we need to pay attention to not only the combination result but also each single modality result. The final loss is the sum of these results:

$$\widetilde{\mathcal{L}} = \mathcal{L}_c + \mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_t, \tag{19}$$

where $\mathcal{L}_c$ is calculated with the outputs of late fusion which fuses multimodal results by weights, $\mathcal{L}_v$, $\mathcal{L}_a$ and $\mathcal{L}_t$ are the losses that are calculated with independent visual, audio and textual classification results.

4 EXPERIMENTS

4.1 Dataset and Experimental Settings

Dataset. We conduct experiments on LVU [24] dataset, which contains 1,339 movie clips in total, comprising 937 training videos, 203 validation videos, and 199 test videos. LVU [24] is proposed to build comprehensive understanding of long-form videos from various perspectives. There are nine tasks of LVU benchmark and SSR is one of them. In this task, each movie clip lasts one to three minutes and corresponds to one speaking style label. There are five speaking style categories, namely *Explain*, *Confront*, *Discuss*, *Teach* and *Threaten*. The distribution of different speaking styles in training set is shown in Table 1. The amount of "teach" and "threaten" samples is obviously smaller than others.

Evaluation Metrics. Similar to general classification tasks, Top-1 Accuracy and F1-score are used as the evaluation metrics, which are shortened to Acc and F1 in Table 3 and 5. Top-1 Accuracy measures the accuracy of our model on speaking style classification, which is calculated as the ratio of true predicted samples in all the samples. F1-score measures the robustness of our model, which is calculated as the formula below:

$$F1 = \frac{2 * recall * precision}{recall + precision}, \tag{20}$$

where *recall* is the ratio of the number of true positive samples to the number of all ground-truth samples and *precision* is the ratio of

Table 1: The quantitative distribution of different speaking styles in LVU training set.

Speaking Style	Quantity
Explain	260
Confront	215
Discuss	247
Teach	105
Threaten	110

Table 2: Ablation results of recognizing speaking styles with unimodal features *vs.* multimodal features on LVU dataset. Here, V means video features, A means audio features, T means textual features.

Method	Top-1 Accuracy	F1-score
V	27.3	16.6
A	32.3	28.5
T	45.5	41.6
V+A	31.3	18.9
V+T	42.4	36.2
A+T	46.5	41.6
V+A+T	**50.0**	**44.1**

the number of true positive samples to the number of all predicted results.

Implementation Details. Since videos of LVU dataset are collected from MovieClip [29], whose data source is YouTube, we firstly cut the last 30 seconds of videos in advance to remove advertisements.

The multimodal sentiment recognition model we used, SelfMM, is pretrained on MOSI [31] dataset for 8 iterations. The pretrained BERT we used is of 12-layers, 768-hidden and 12-heads. There are 5 layers in cross-attention encoder and each layer contains 5 attention heads.

In the training stage, we adopt Adam optimizer with initial learning rate 0.001. The model is trained for 100 iterations on LVU dataset and the batch size is 16. All experiments are implemented using Pytorch on one GPU of RTX 3090.

4.2 Ablation Study

To verify the effectiveness of multimodal information on SSR, we conduct comparative study on multimodal features and unimodal features. We also compare the performance of using early fusion and late fusion strategy to integrate multimodal features, since multimodal fusion is a great challenge for multimodal methods.

To evaluate the effectiveness of introducing the influence of sentiments, we compare the recognition results of applying sentiment-fused features and those without sentiment features. Moreover, to explore appropriate sentiment-fusion strategies, we experiment on two sentiment-fusion strategies. One is to fuse them by a cross-attention encoder and the other one is to conduct a direct dot product on multimodal features and their corresponding sentiment features.

By default, we apply cross-attention encoder as sentiment fusion strategy and use weighted late fusion strategy as multimodal fusion strategy because this combination can achieve the best performance.

Multimodal Features. In Table 2, the performance of only using visual features is the worst. The reason is that it is hard for the model to converge when we only use visual features which are highly complex. Audio and text are easier to be understood because they contain more clear and intensive information.

When we fuse visual and audio features or fuse visual and textual features, the performances are better than those of the methods that only using visual features. When we fuse audio and textual features, the combination performs better than that only using audio features. The performance of the varient that fuses visual, audio and textual features performs best, which is superior to all combinations. The results of these varients demonstrate that multimodal information can complement each other, which improves the performance on SSR.

We also think that subtitles are essential when understanding long-form videos, considering that varients including textual features generate desired results. It is noted that the varient fusing visual and textual features performs worse than that of only textual features, we think it is because the differences between visual and textual features bring in noise when a direct concat is conducted to fuse them.

Sentiment Fusion. As shown in Table 3, the performances of both unimodal and multimodal varients are improved after fusing corresponding sentiment features, indicating that the introduction of sentiments is effective. It also proves that sentiments are indeed correlated with speaking styles and can enhance the feature expressions to recognize speaking styles.

When more modalities are used, the performance is improved more after multimodal features are fused with sentiments. It proves that sentiment-fusion is appropriate for multimodal architecture.

Multimodal Fusion Strategies. It can be seen from the line 1 and 3 of Table 4 that when we use multimodal features without fusing sentiment features, early fusion is better than late fusion in almost all varients. We introduce late fusion because we think that multimodal features are different from each other and direct concat will bring noise to different features. When we fuse features of two or three modalities with by concat, the noise seems not that serious.

However, from the line 2 of Table 4, we can see that the influence is clear when we fuse sentiment features to multimodal features. The noise is too large for model to converge, and it can be explained as this is because the number of features fusing sentiments is twice

Table 3: Ablation results of recognizing speaking styles with *vs.* without sentiment features on LVU dataset. Here, V represents visual features, A represents audio features, T represents textual features; w/ and w/o sentiment means if fusing multimodal and sentiment features or not.

Method	V		A		T		V+A		V+T		A+T		V+A+T	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
w/o sentiment	26.8	11.3	31.3	**29.1**	43.4	36.1	28.8	**23.6**	40.9	36.1	37.4	31.5	41.4	37.1
w/ sentiment	**27.3**	**16.6**	**32.3**	28.5	**45.5**	**41.6**	**31.3**	18.9	**42.4**	**36.2**	**46.5**	**41.6**	**50.0**	**44.1**

Table 4: Ablation results of fusing multimodal results by early fusion *vs.* late fusion strategy to recognize speaking styles of LVU dataset. Here, V means visual features, A means audio features, T means textual features; early fusion means fusing multimodal features by concat directly and late fusion means conducting a weighted average of multimodal results; w/ and w/o sentiment means if fusing multimodal and sentiment features or not.

Method		Top-1 Accuracy				F1-score			
		V+A	V+T	A+T	V+A+T	V+A	V+T	A+T	V+A+T
early fusion	w/o sentiment	37.9	43.4	39.9	44.9	34.8	37.1	30.2	44.2
	w/ sentiment	29.3	29.3	26.8	26.8	13.3	13.3	11.3	11.3
late fusion	w/o sentiment	28.8	40.9	37.4	41.4	23.6	36.1	31.5	37.1
	w/ sentiment	31.3	42.4	46.5	50.0	18.9	36.2	41.6	44.1

Table 5: Ablation results of fusing sentiments by Dot Product *vs.* Cross-Attention Encoder to recognize speaking styles on LVU dataset. Here, V means visual features, A means audio features, T means textual features.

Method	V		A		T		V+A		V+T		A+T		V+A+T	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Dot Product	26.8	11.3	29.8	24.9	41.9	38.0	**34.8**	**31.3**	**43.4**	**38.1**	40.4	36.4	39.4	30.7
Cross-Attention Encoder	**27.3**	**16.6**	**32.3**	**28.5**	**45.5**	**41.6**	31.3	18.9	42.4	36.2	**46.5**	**41.6**	**50.0**	**44.1**

of that without sentiments. When dealing with more data sources, the differences among different modalities will be larger and the noise will be enhanced.

As we expected, the line 4 of Table 4 shows that late fusion can reduce negative effects of large differences among multimodal features fused with sentiments. When we fuse more modalities, the advantage of late fusion is more obvious. The best performance is achieved when we apply late fusion strategy to integrate sentiment-fused visual, audio and textual features.

Sentiment Fusion Strategies. It can be seen in Table 5 that, for most unimodal and multimodal varients, cross-attention encoder performs better than dot product on fusing multimodal features and sentiment features. This proves that the multi-head cross-attention layer is effective in terms of introducing the influence of sentiments, which is helpful to recognize speaking styles.

However, dot product does well in some varients that include visual features, due to the high complexity of visual features and the plain processing of visual features. As a result, the differences between visual and other features are not enhanced, which leads to less negative influence of feature fusion.

4.3 Comparison with the SOTA

To prove the effectiveness of MMSF, we compare it with the typical video analysis methods. Due to SSR is an up-to-date task proposed by LVU [24], there are few methods targeted at it. As a result, except for SOTA methods on SSR, we add a lot of outstanding video classification methods for comparison. As shown in Table 6, Object Transformer [24], ViS4mer [15] and STAN [9] are designed for LVU tasks while SlowFast [8], VideoBERT [22], MulT [23], VIVIT[2], MMIM [11], Swin Transformer [17] and MeMViT [25] are video classification models, and the performance of MMSF is superior to all of them on SSR.

We use seven as baselines, namely SlowFast [8], VIVIT[2], Object Transformer [24], Swin Transformer [17], ViS4mer [15], STAN [9] and MeMViT [25]. All of them only takes visual information into consideration. MMSF is obviously superior in both Top-1 Accuracy and F1-score, which is due to its integrated utilization of various information of movie clips. Even of different speaking styles, conversations are often visually similar. In such cases, audio and text contain more helpful information that enables us to recognize targeted speaking styles. The influence of applying multimodal features can also be seen in the aforementioned ablation experiment.

Table 6: Comparison results of our method *vs.* different state-of-the-art methods on LVU dataset.

Method	Top-1 Accuracy	F1-score
SlowFast [8]	35.8	23.79
VideoBERT [22]	37.9	-
MulT [23]	38.9	32.2
VIVIT [2]	28.14	23.69
MMIM [11]	32.8	24.2
Object Transformer [24]	38.4	38.2
Swin Transformer [17]	32.16	31.19
ViS4mer [15]	40.8	-
STAN [9]	41.41	-
MeMViT [25]	34.67	35.06
MMSF (Ours)	**50.0**	**44.1**

The other reason that MMSF achieves the best performance is the introduction of sentiment. Among all the baselines, MulT [23], VideoBERT [22] and MMIM [11] are multimodal methods. Although they have utilized multimodal information to facilitate video understanding, the performances of these methods are still worse than MMSF. For one thing, sentiments, which consist an important part of MMSF, have strong impacts on SSR. For another, it is challenging to fuse multimodal features as the mutual influences of different

Figure 6: Qualitative comparisons of speaking style recognition results using different methods on LVU dataset. Here, the left part of each example are the frames of one movie clip; the right part of each example are the speaking styles of the movie clip predicted by different methods; results in green color represent correct predictions while those in red color represent wrong ones.

modalities are required to be balanced with a specific multimodal fusion strategy.

In addition, among all the baselines, Object Transformer [24], ViS4mer [15] and STAN [9] are aimed at LVU tasks. Although they achieve excellent performances compared to other methods, they are still far worse than MMSF on SSR. These pretrained visual feature extractors can be applied to a variety of downstream tasks, but may not work for all the video understanding tasks, especially for those of quite different types. To handle with specific tasks like SSR, it is necessary to design more specific methods. Because sounds and subtitles are important parts of human conversations, MMSF focuses on not only visual but also audio and textual information of videos. Moreover, as sentiments are one of the motivations of human behaviors, MMSF takes sentiment fusion as a part of pipeline. It is these designs that enable MMSF to recognize speaking styles more effectively.

Figure 6 shows qualitative results of the proposed MMSF on LVU dataset. We can see that MMSF can distinguish different speaking styles well while other methods tend to confuse among different speaking styles. Videos of explain, confront and discuss visually seem similar, and thereby they are prone to be mistaken for each other.

When the label of video is "teach" or "threaten", it is hard to be recognized. For one thing, the unbalanced data distribution limits the ability of MMSF to give correct predictions about these two categories. For another, "teach" often involves multi-person conversation, while other speaking styles tend to describe one-to-one conversations. Despite the limited training samples of "threaten", MMSF can distinguish "threaten" from other speaking styles owing to its comprehensive analysis of sentiment information, which enhances feature expressions and magnifies the differences among different speaking styles.

5 CONCLUSION

In this paper, we proposed a multimodal sentiment-fused method, MMSF, to recognize speaking styles. By extracting visual, audio and textual features by multiple processing steps and integrating these features with a weighted late fusion strategy, MMSF made full use of multimodal information of human conversations to analyze the speaking styles of them. With the help of cross-attention encoder, the effects of sentiments were introduced to multimodal features, which enhanced the ability of MMSF to understand human conversations. To evaluate the effectiveness of MMSF, we conducted extensive experiments on LVU dataset. The experimental results showed that our method significantly outperforms the state-of-the-art methods and confirmed the effectiveness of applying multimodal features and sentiments.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation of China (62072232) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: transformers for multimodal self-supervised

learning from raw video, audio and text. In *Advances in Neural Information Processing Systems*. 24206–24221.
[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: a video vision transformer. In *IEEE International Conference on Computer Vision*. 6836–6846.
[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition*. 59–66.
[4] Fabrice Bellard. [n. d.]. *FFmpeg*. http://ffmpeg.org
[5] Runqing Cao, Chunyang Ye, and Hui Zhou. 2020. Multimodel sentiment analysis with self-attention. In *Future Technologies Conference*. 16–26.
[6] Shixing Chen, Xiang Hao, Xiaohan Nie, and Raffay Hamid. 2022. Movies2Scenes: learning scene representations using movie similarities. In *arXiv preprint arXiv:2202.10650*.
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.
[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *IEEE International Conference on Computer Vision*. 6202–6211.
[9] Edward Fish, Jon Weinbren, and Andrew Gilbert. 2022. Two-stream transformer architecture for long video understanding. In *British Machine Vision Conference*.
[10] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *International Conference on Multimodal Interaction*. 6–15.
[11] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*. 9180–9192.
[12] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: scaling up end-to-end speech recognition. In *arXiv preprint arXiv:1412.5567*.
[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM International Conference on Multimedia*. 1122–1131.
[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural Computation*. 1735–1780.
[15] Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*. 87–104.
[16] jdepoix. [n. d.]. *YouTube Transcript API*. https://github.com/jdepoix/youtube-transcript-api.git
[17] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3202–3211.
[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: audio and music signal analysis in python. In *Python in Science Conference*. 18–25.
[19] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *International Conference on Multimodal Interfaces*. 169–176.
[20] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *International Conference on Machine Learning*. 689–696.
[21] Paul Rayson, Geoffrey N Leech, and Mary Hodges. 1997. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. In *International Journal of Corpus Linguistics*. 133–152.
[22] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: a joint model for video and language representation learning. In *IEEE International Conference on Computer Vision*. 7464–7473.
[23] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Conference. Association for Computational Linguistics. Meeting*. 6558.
[24] Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1884–1894.
[25] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Memvit: memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 13587–13597.
[26] C Xu, S Cetintas, KC Lee, and LJ Li. 2014. Visual sentiment prediction with deep convolutional neural networks. In *arXiv preprint arXiv:1411.5731*.
[27] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. 2022. I-code: an integrative and composable multimodal learning framework. In *arXiv preprint arXiv:2205.01818*.
[28] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI Conference on Artificial Intelligence*. 381–388.
[29] YouTube. [n. d.]. *MovieClips*. https://www.movieclips.com/
[30] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence*. 10790–10797.
[31] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *arXiv preprint arXiv:1606.06259*.