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

Beibei Zhang, Yaqun Fang, Fan Yu, Jia Bei, Tongwei Ren

State Key Laboratory for Novel Software Technology, Nanjing University

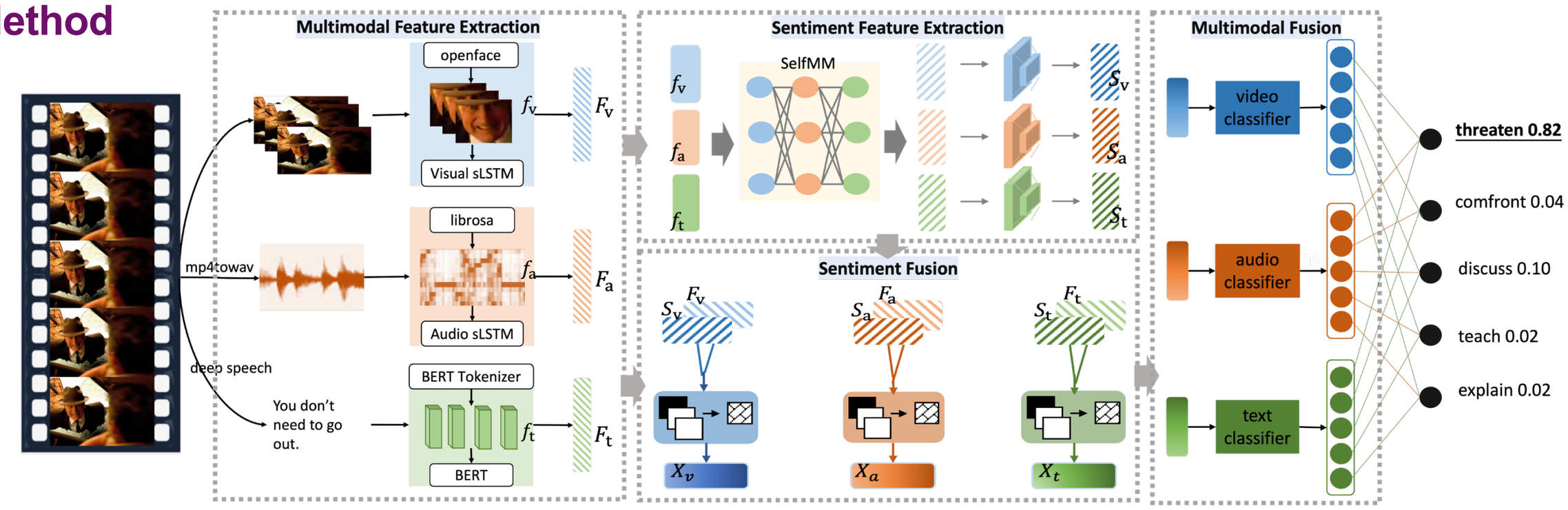
ICMR 2023

Introduction

Speaking style recognition (SSR) is aimed at recognizing the styles of conversations, which provides a fine-grained description about talking.

To recognize speaking styles, we propose a novel multimodal sentiment-fused method, MMSF, which extracts and integrates visual, audio, textual and corresponding sentiment features of videos. 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**.

Method



The input of MMSF is a movie clip and the output is the speaking style classification result. Original visual, audio and textual features are firstly extracted by feature extraction models. Then, a pre-trained multimodal sentiment recognition model is applied to obtain multimodal sentiment features. After that, multimodal features are fused with sentiment features with the help of cross-attention encoder. Multimodal classification results are obtained after handling multimodal sentiment-fused features with softmax classifiers, which are integrated by late fusion strategy to generate the final recognition result.

Experiments

Dataset: Long form understanding dataset (LVU)

- Training : Validation : Test = 937 : 203 : 199
- Categories: Explain, Confront, Discuss, Teach and Threaten

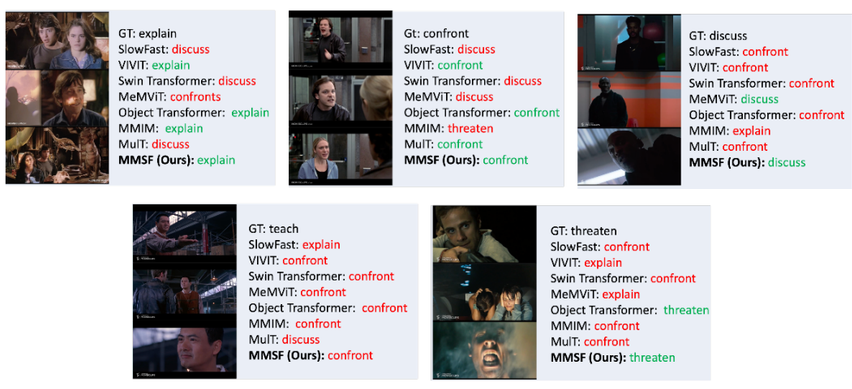
Metrics: Top-1 Accuracy and F1-score

Ablation Study: We conduct multiple ablation experiments. The result shows that the varient which integrates visual, audio and textual features and fuses sentiment features performs best. This demonstrates that applying multimodal information and introducing sentiment are effective.

Comparison with the SOTA: MMSF is superior to all the video analysis methods in both accuracy and F1-score.

Qualitative Analysis: MMSF can distinguish different speaking styles well while other methods tend to confuse different speaking styles.

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**



南京大学 NANJING UNIVERSITY

zhangbb@smail.nju.edu.cn

rentw@nju.edu.cn

MAGUS MediA recoGnition and UnderStanding