ICMR 2023
12-15, June
Thessaloniki, Greece

Poster ID: 17

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

- **Motivation:** Talking accounts for a large proportion in human daily lives. **Speaking style recognition** (SSR) aims to identify the talking styles among characters in videos, which can help other video understanding tasks.

- **Novelty:**

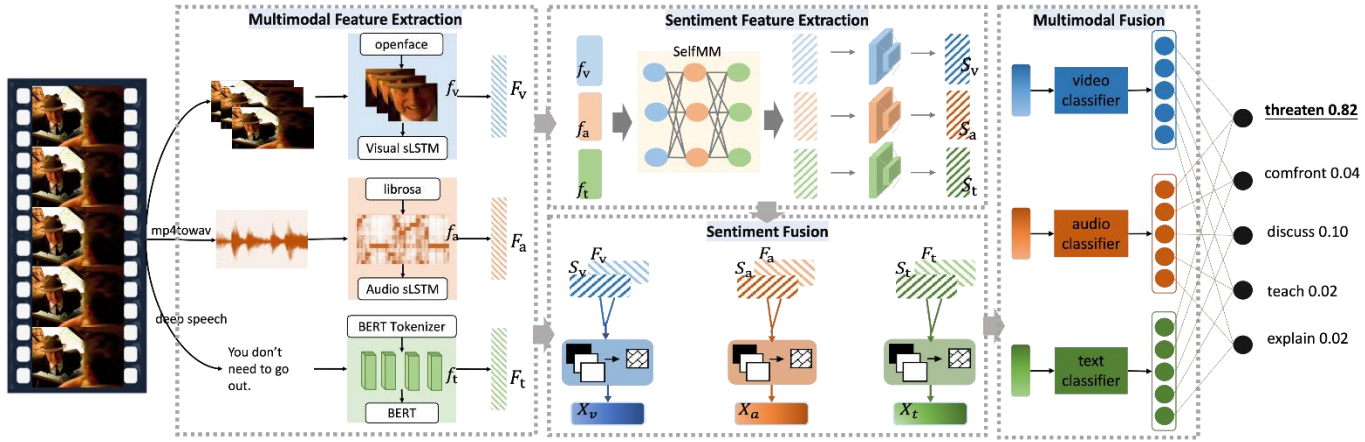
  Visual feature ➔ **Multimodal** feature + **Sentiment** fusion

- **Solution:** We propose **MMSF**, a multimodal sentiment-fused method to recognize video speaking style.

People “talk” to each other

Speaking styles

Applications of SSR



Framework of MMSF

- **Experiment:** We conduct experiments on LVU dataset, which involves 1,399 movie clips and 5 speaking styles. Both quantitative and qualitative analysis results show that the performance of **MMSF is superior to all of state-of-the-arts**.

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**
	(+8.59)	**(+5.90)**

Qualitative Analysis

Quantitative Analysis

ICMR 2023
12-15, June
Thessaloniki, Greece

Poster ID: 17

Beibei Zhang, Yaqun Fang, Fan Yu, Jia Bei, Tongwei Ren
Email: zhangbb@smail.nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University

MAGUS
MediA recoGnition and UnderStanding