A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations

Wenjie Zheng[1], Jianfei Yu[1], Rui Xia[1], and Shijin Wang[2]

[1]Nanjing University of Science and Technology, China
[2]iFLYTEK AI Research, China

ACL 2023, Toronto, Canada
May.28 2023

Background

- **Multimodal Emotion Recognition in Multi-party Conversations (MERMC)**
  - Goal: recognize the emotion of the real speaker in an utterance.

anger

Background

- **Multimodal Emotion Recognition in Multi-party Conversations (MERMC)**
  - Goal: recognize the emotion of the real speaker in an utterance.

Motivation of Our Work

- **Limitation of Existing MERMC Approaches**

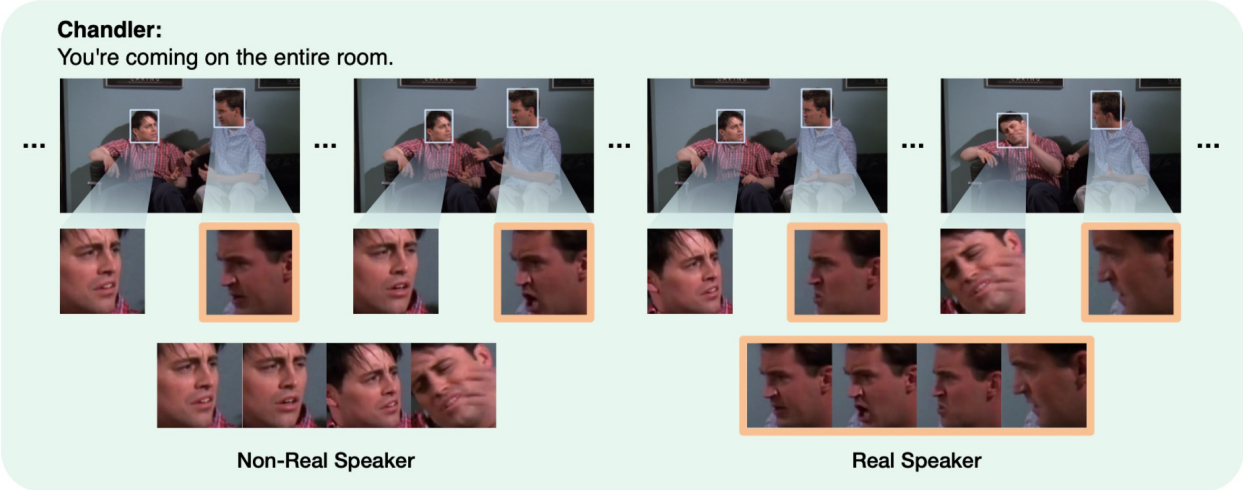
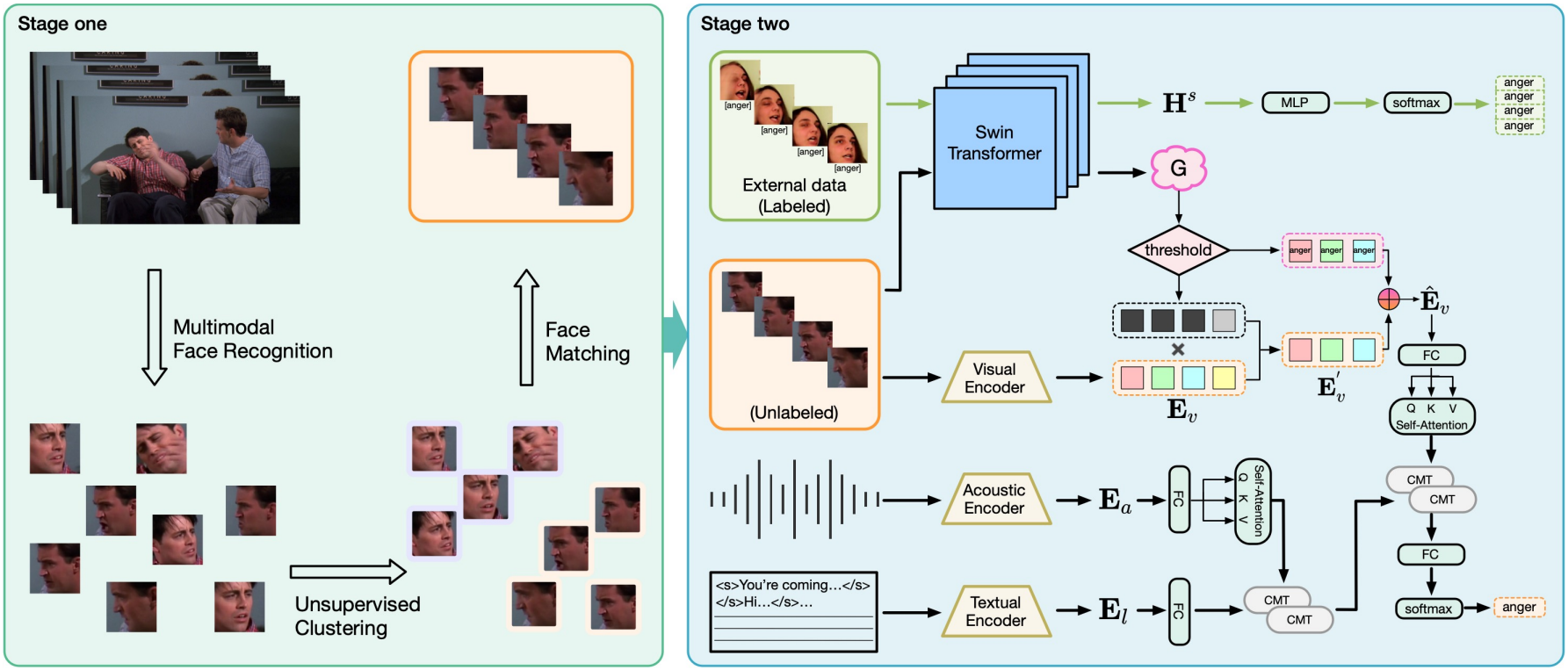


Figure 1: An example of MERMC task where an utterance contains two individuals with different facial expressions. One (*Joey* on the left side of the frame) expresses *disgust*, while the other (*Chandler* on the right side of the frame) expresses *anger*, and the latter is the real speaker whose emotion is annotated as the emotion of the utterance.

Methods	*#Col*	*#Seg*	*#Rec*
MELD (Poria et al., 2019)	✗	✗	✗
UniMSE (Hu et al., 2022b)	✗	✗	✗
MMGCN (Hu et al., 2021)	✓	✗	✗
MESM (Dai et al., 2021)	✓	✗	✗
M^{3}ED (Zhao et al., 2022b)	✓	✗	✗
FacialMMT (Ours)	✓	✓	✓

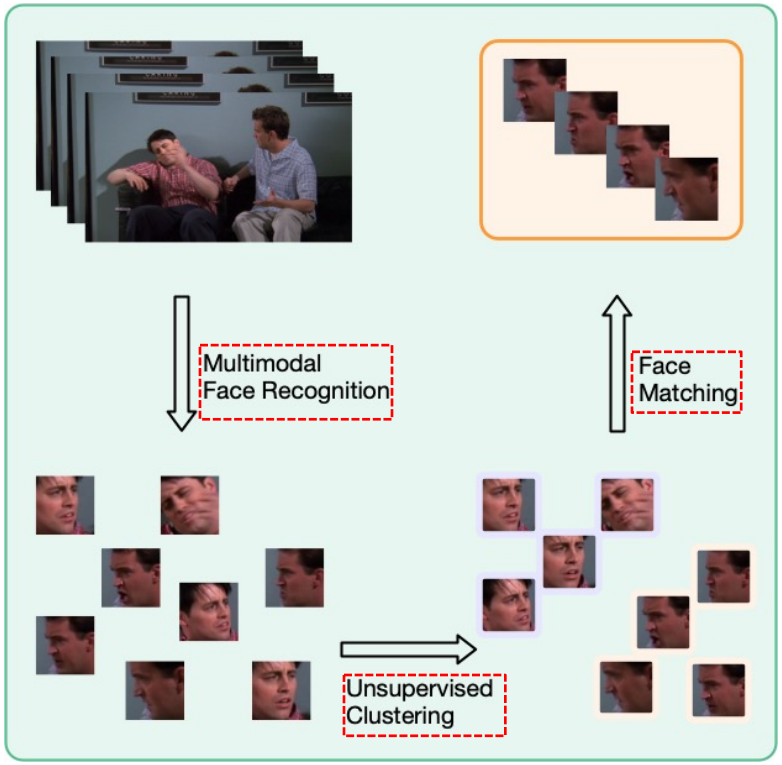
Our Proposed Framework

- **Overview of the proposed framework**
  - A two-stage framework



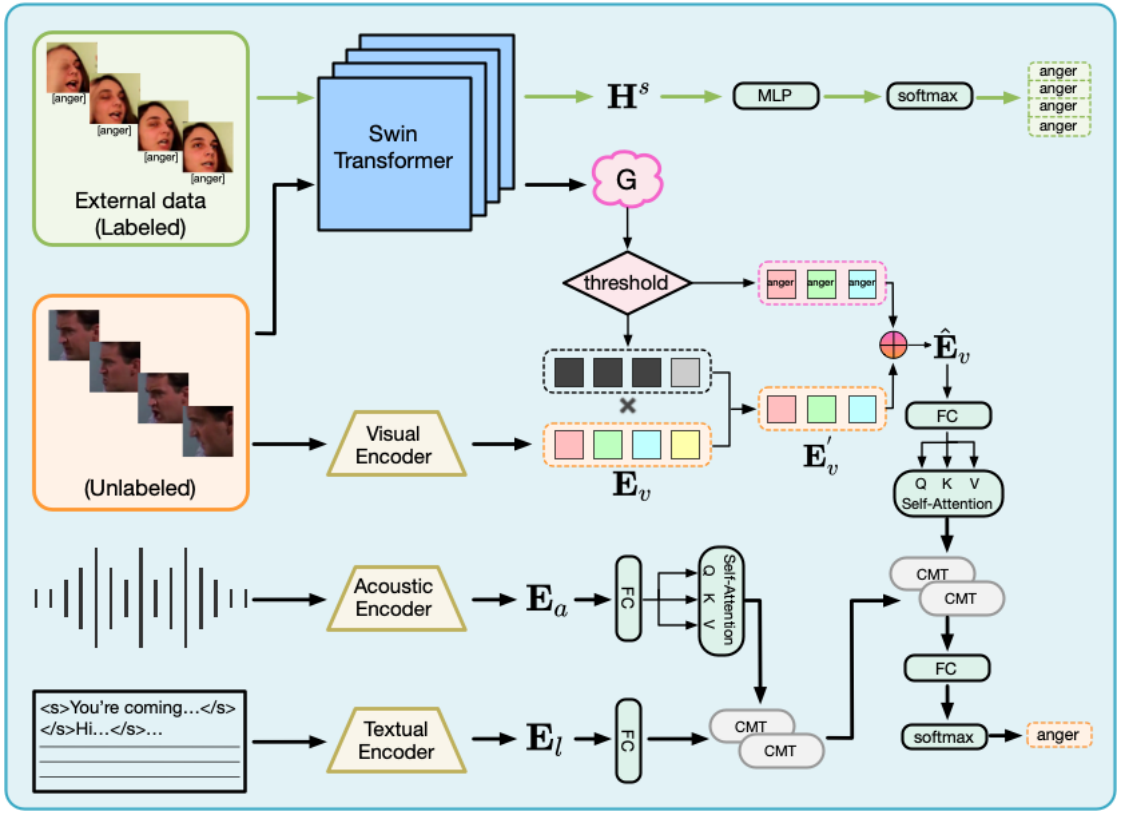
Our Proposed Framework

- **Stage One:** extracting facial sequences of the real speaker in an utterance
  - extract facial sequences of all possible speakers
  - indentity the number of face clusters in the sequences
  - determine the facial sequences of the real speaker



Our Proposed Framework

- **Stage Two:** proposing a multimodal facial expression-aware multi-task learning model
  - frame-level facial expression recognition task
  - utterance-level emotion recognition task



Experiments

- **Main results**
  - Multimodal Emotion Recognition

Models	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	F1
DialogueRNN (Majumder et al., 2019)	73.50	49.40	1.20	23.80	50.70	1.70	41.50	57.03
ConGCN (Zhang et al., 2019)	76.70	50.30	8.70	28.50	53.10	10.60	46.80	59.40
MMGCN (Hu et al., 2021)	-	-	-	-	-	-	-	58.65
DialogueTRM* (Hu et al., 2021)	-	-	-	-	-	-	-	63.50
*DAG-ERC** (Shen et al., 2021)	-	-	-	-	-	-	-	63.65
MM-DFN (Hu et al., 2022a)	77.76	50.69	-	22.94	54.78	-	47.82	59.46
EmoCaps⋆ (Li et al., 2022b)	77.12	**63.19**	3.03	**42.52**	57.50	7.69	**57.54**	64.00
UniMSE▲ (Hu et al., 2022b)	-	-	-	-	-	-	-	65.51
GA2MIF (Li et al., 2023)	76.92	49.08	-	27.18	51.87	-	48.52	58.94
FacialMMT-BERT	78.55	58.17	13.04	38.51	61.10	**30.30**	53.66	64.69
FacialMMT-RoBERTa	**80.13**	59.63	**19.18**	41.99	**64.88**	18.18	56.00	**66.58**

Table 2: Comparison results of the MERMC task on the MELD dataset. The baselines with italics only use textual modality. ▲ indicates the model uses T5 (Raffel et al., 2020) as the textual encoder. The baselines tagged with ⋆ and * respectively use BERT and RoBERTa as textual encoders. The best results are marked in bold.

Experiments

- **Main results**
  - Visual Modality Emotion Recognition

Models	Composition of visual information	F1
EmoCaps	Video frames	31.26
MM-DFN	Video frames	32.34
MMGCN	Possible speakers' face sequences	33.27
FacialMMT	Real speaker's face sequence	**36.48**
	- w/o UC, FM	34.36
	- w/o MFR, UC, FM	32.27

Table 4: Comparison of single visual modality emotion recognition results. MFR represents multimodal face recognition, UC represents unsupervised clustering, and FM represents face matching.

Experiments

- **Case study**

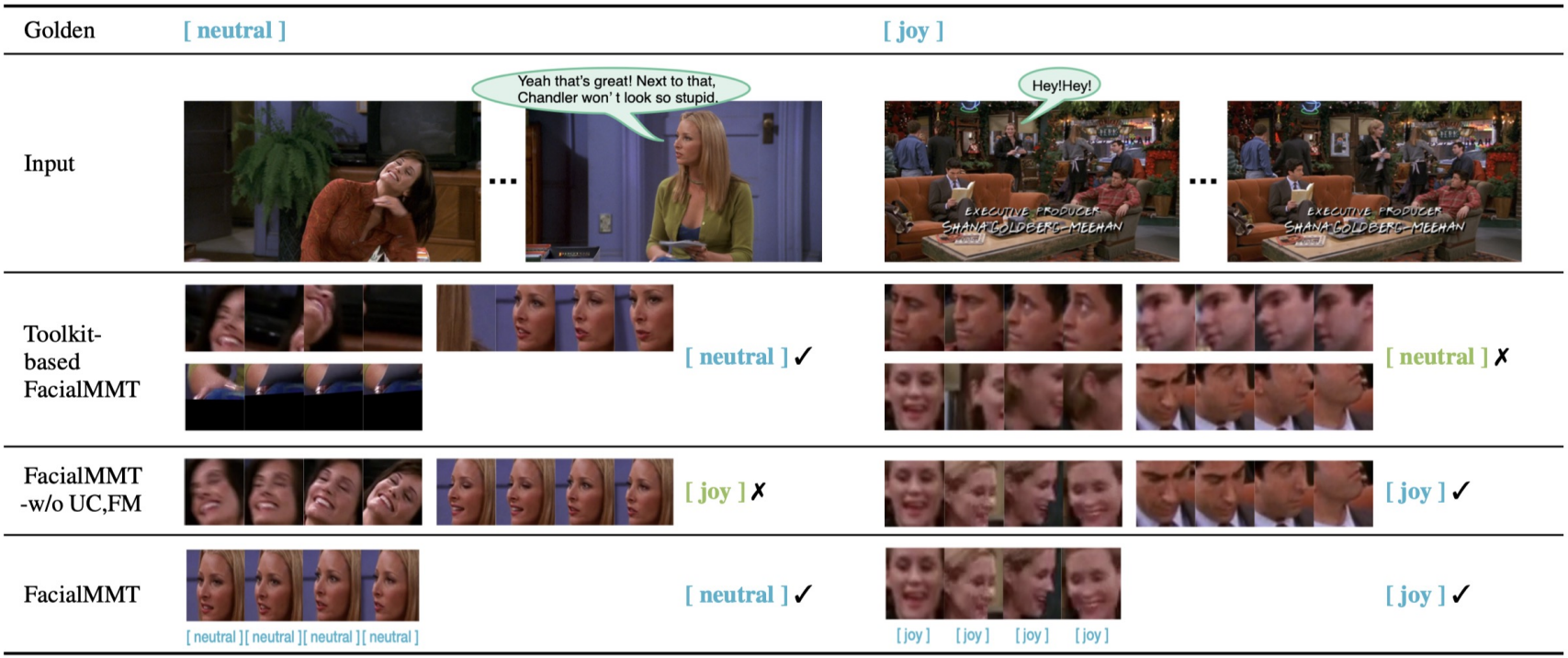


Figure 3: Prediction comparison between different methods on two test samples for the MERMC task.

Conclusions and Future Work

- **Conclusions**
  - A multimodal face sequence extraction method
    - Obtain the facial sequences of the real speaker in an utterance in a pipeline manner
  - A multi-task learning Framework
    - Leverage Frame-level task to help utterance-level task
  - Experiments
    - Achieve the SOTA performance on the benchmark MELD dataset
- **Future work**
  - Build an end-to-end framework instead of a two-stage approach
  - Explore better cross-modal alignment and multimodal fusion mechanisms

Thanks for your attention!

The source code is publicly available on
https://github.com/NUSTM/FacialMMT