ACM multimedia 2020 Seattle

mmgc2627

Deep Relationship Analysis in Video with Multimodal Feature Fusion

Fan Yu, Dandan Wang, Beibei Zhang, Tongwei Ren*

南京大学 NANJING UNIVERSITY

MAGUS MediA recoGnition and UnderStanding

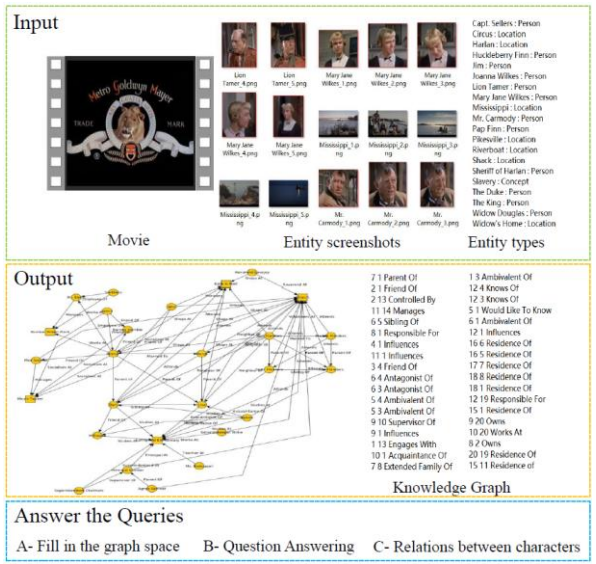
Task & Dataset

Deep video understanding (DVU)

- requires the analysis of known information to reason about hidden information
- populates a knowledge graph of a long duration video with example screenshots and types of the entities

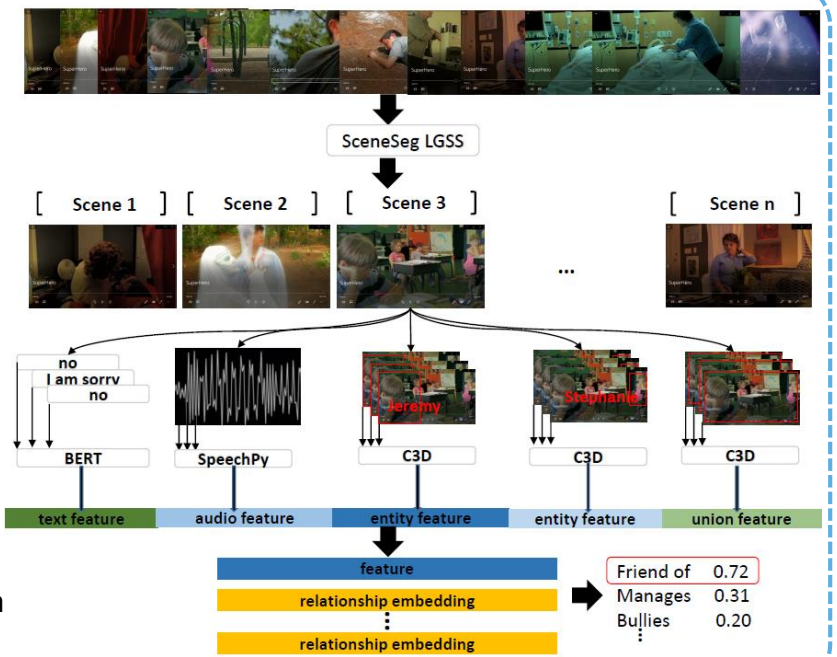
HLVU dataset

- 10 videos, 6 for development and 4 for test
- entity name, type and screenshot
- 59 predefined relationships
- knowledge graph



Solution

- Scene segmentation
- Person tracking & recognition
- Location recognition
- Auto subtitles->text features
- Audio features
- Vision features
- Multimodal features fusion
- Zero-shot learning
- Entity-relationship graph
- Query answering



Experiments

Method	R@50	R@100	R@θ
T+E/+C/+C+N	2.959/5.325/8.876	9.467/10.651/17.160	2.367/3.550/9.467
T+E+U/+C/+C+N	8.876/9.467/14.201	**10.651**/**11.834**/17.160	7.692/**8.284**/**11.834**
A+E/+C/+C+N	2.367/2.959/8.284	8.284/11.243/17.160	2.367/2.959/8.284
A+E+U/+C/+C+N	0.592/1.775/8.284	2.959/8.284/13.018	0.592/1.183/7.692
T+A+E/+C/+C+N	**9.467**/9.467/**14.793**	10.059/**11.834**/**17.751**	**8.284**/**8.284**/10.651
T+A+E+U/+C/+C+N	**9.467**/**10.059**/**14.793**	**10.651**/11.243/**17.751**	5.917/7.101/**11.834**
C+N	8.284	13.609	7.692

T: text feature, A: audio feature, E: entity vision feature, U: union vision feature, C: category rule, N: name rule

	Fill in the graph space	Question Answering	Relations between characters
Shooter	0.5555555	1.0	0.0,0.0,NaN,0.0,0.0,NaN
Sophie	0.3616558	0.875	0.0,0.0,NaN,0.0,0.0,NaN,0.0,0.0,NaN, 0.0,0.0,NaN,0.0,0.0,NaN,0.0,0.0,NaN
Time Expired	0.33025533	0.5	0.0,0.0,NaN,0.0,0.0,NaN,0.0,0.0,NaN, 0.0,0.0,NaN,0.0,0.0,NaN,0.0,0.0,NaN
The Big Something	0.5083333	0.75	0.0,0.0,NaN,0.0,0.0,NaN

Fan Yu: yf@smail.nju.edu.cn Tongwei Ren: rentw@nju.edu.cn