ACM multimedia 2020 Seattle

Deep Relationship Analysis in Video with Multimodal Feature Fusion

Fan Yu [1,2], DanDan Wang [1], Beibei Zhang [1], Tongwei Ren [1,2,*]

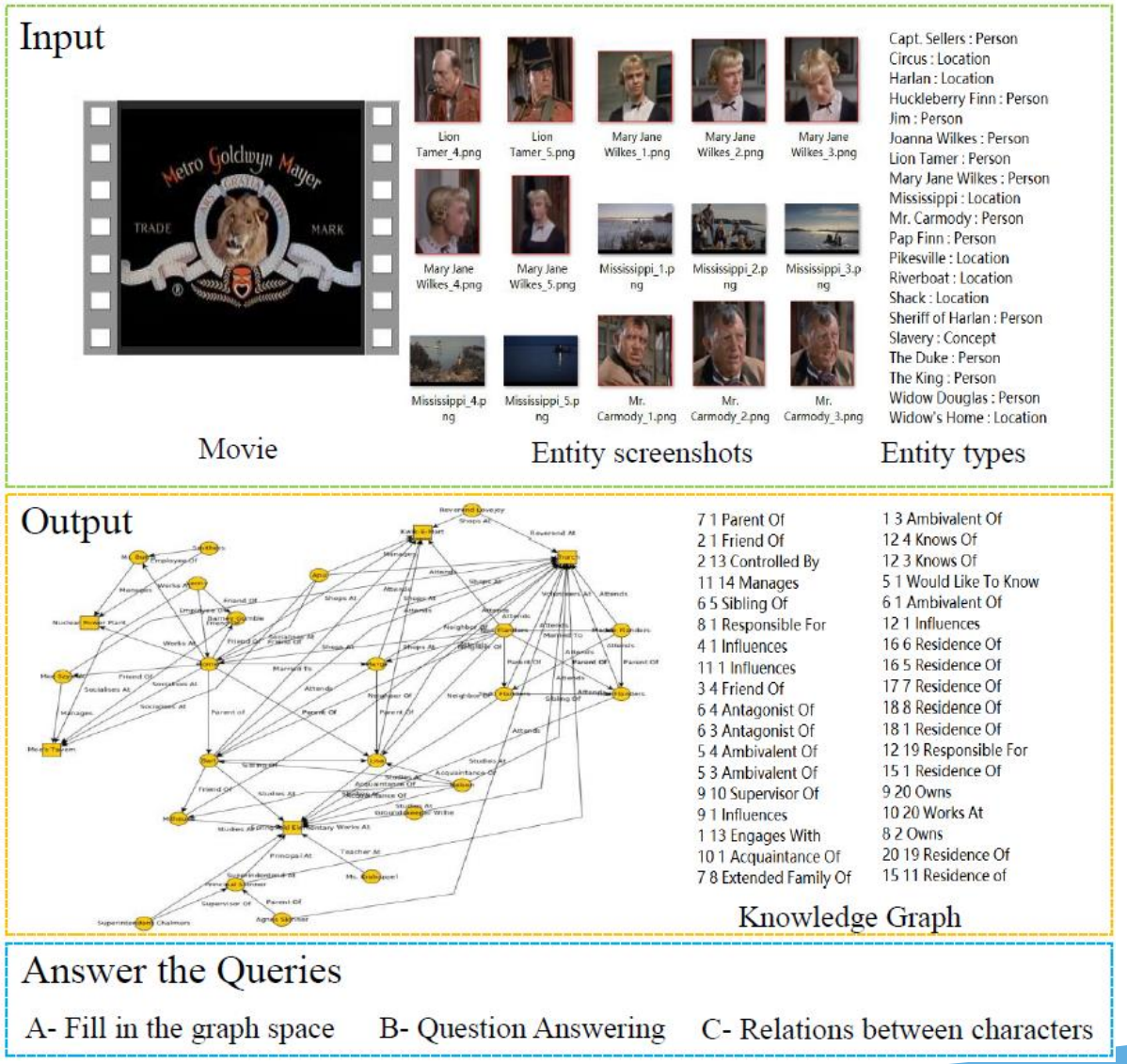
[1] State Key Laboratory for Novel Software Technology, Nanjing University

[2] School of Computer Science, Nanjing University of Science and Technology

MAGUS
MediA recoGnition and UnderStanding

Task & Dataset

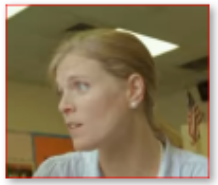
- **Deep video understanding (DVU)**
  - requires the analysis of known information to reason about hidden information
  - populates a knowledge graph of a long duration video with example screenshots and types of the entities
- **HLVU** dataset
  - 10 videos, 6 for development and 4 for test
  - entity name, type and screenshot
  - 59 predefined relationships
  - knowledge graph

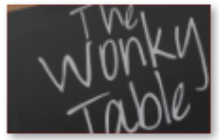


Analysis

- **Challenges**
  - Relationships between entities may change over time in long duration videos.
  - It is hard to extract and mix validated multimodal features.
  - Additional common sense might be used for predicting relationships between entities, especially those that have not co-occurred or even not appeared in sight.
  - It is difficult to distinguish between the relationships that have similar meaning.
  - There are only 6 samples in the development set.


Ms. Johnson_2.png

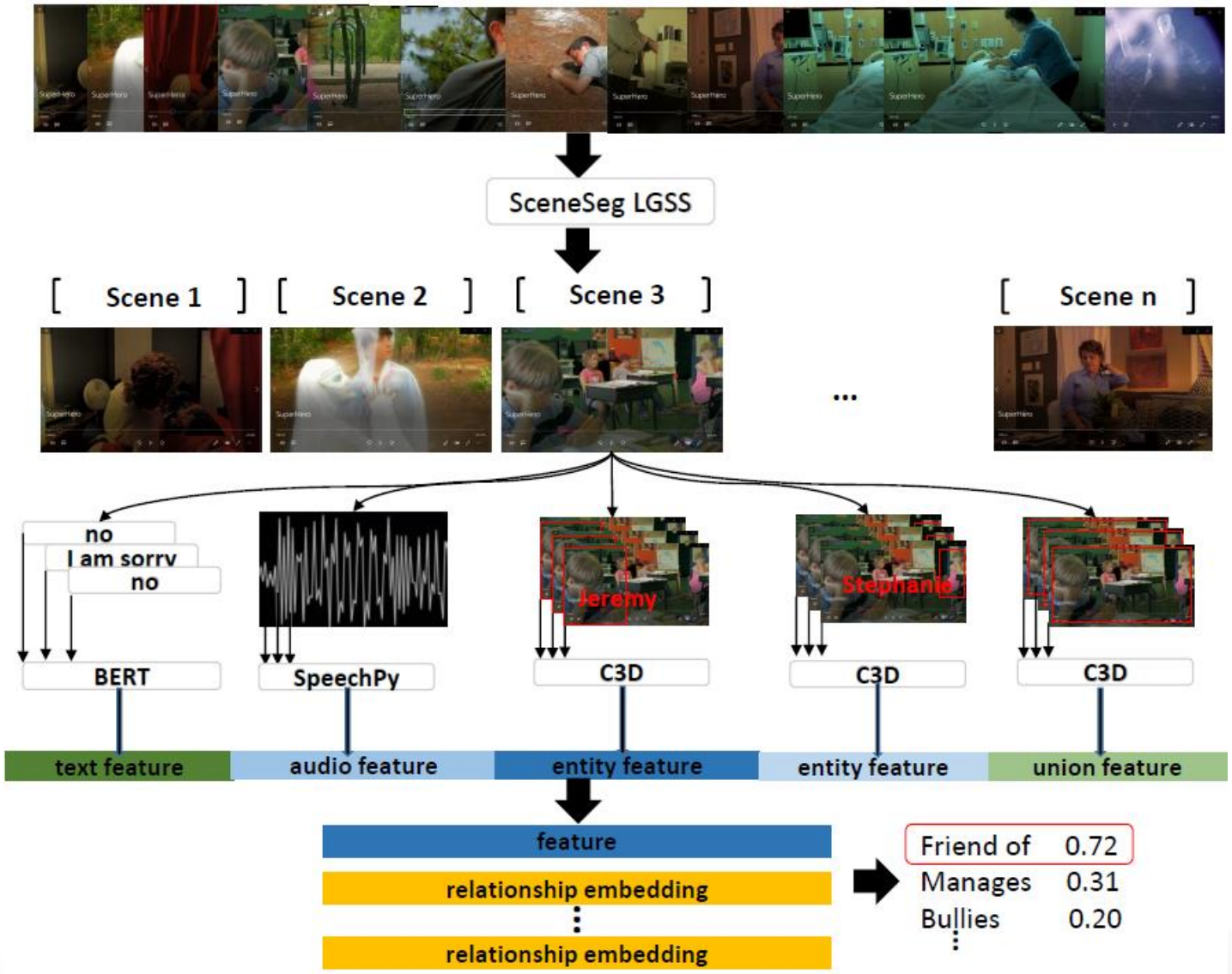

Ms. Johnson_4.png



Cafe_1_1.png


Cafe_1_2.png

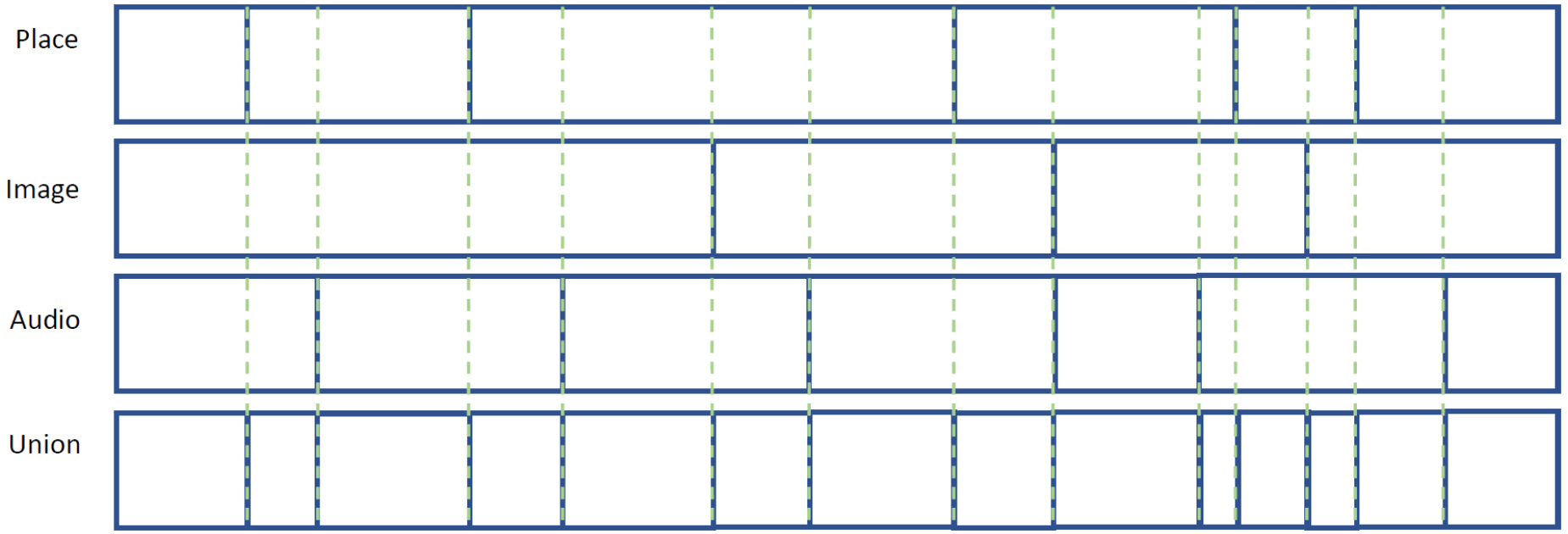
Solution



- 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

Solution

- Scene segmentation——SceneSeg LGSS
  - shot detection
  - scene segmentation by place, image and audio

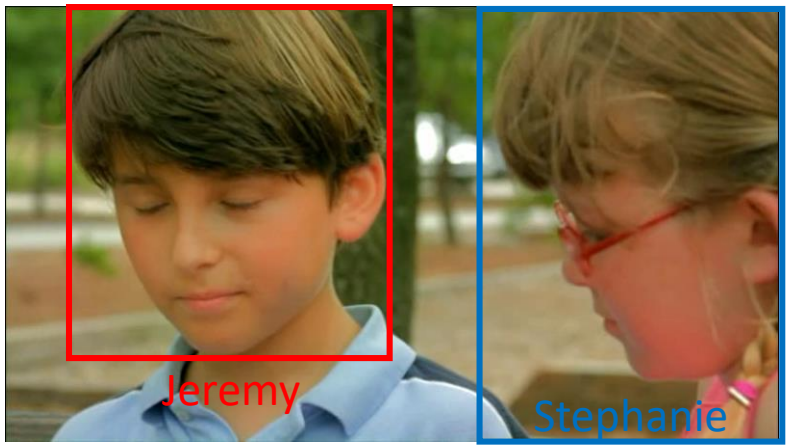


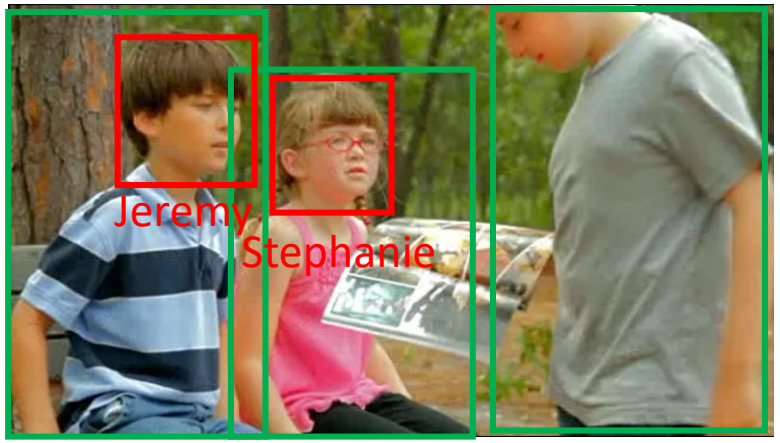
Solution

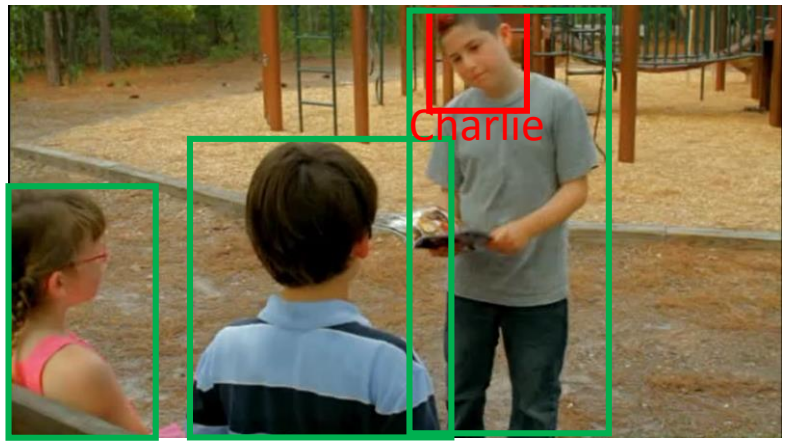
- Location recognition
  - SURF template matching

Solution

- Person tracking & recognition
  - InsightFace——face recognition
  - CenterTrack——person tracking
  - SURF template matching







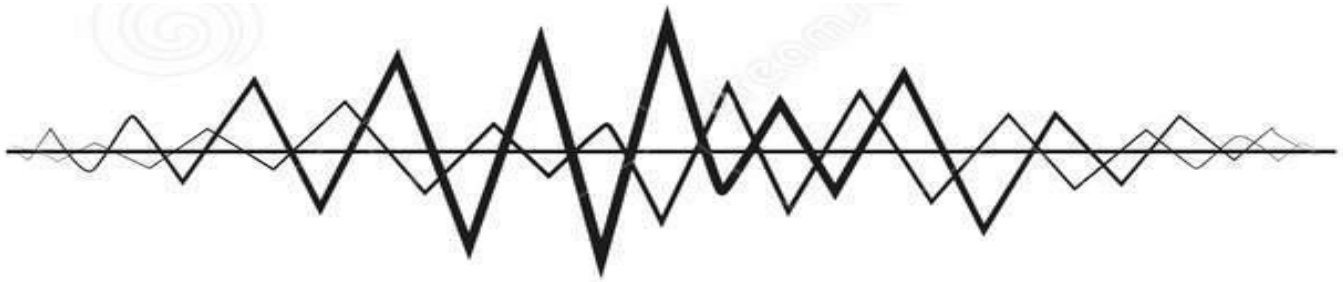
Solution

- Auto subtitles->text features
  - Youtube, aliyun, autoSub
  - BERT model

```
1
00:00:01,060 --> 00:00:53,280
[Music]

2
00:00:53,280 --> 00:00:56,290
the mighty Celestials powers were too

3
00:00:56,290 --> 00:00:58,480
much for the Beast and with a wave of

4
00:00:58,480 --> 00:01:02,170
her hand she lifted the beast up off the

5
00:01:02,170 --> 00:01:04,089
ground and threw him back into the dark

6
00:01:04,089 --> 00:01:07,890
realm from which he came

7
00:01:07,890 --> 00:01:14,880
[Music]
```

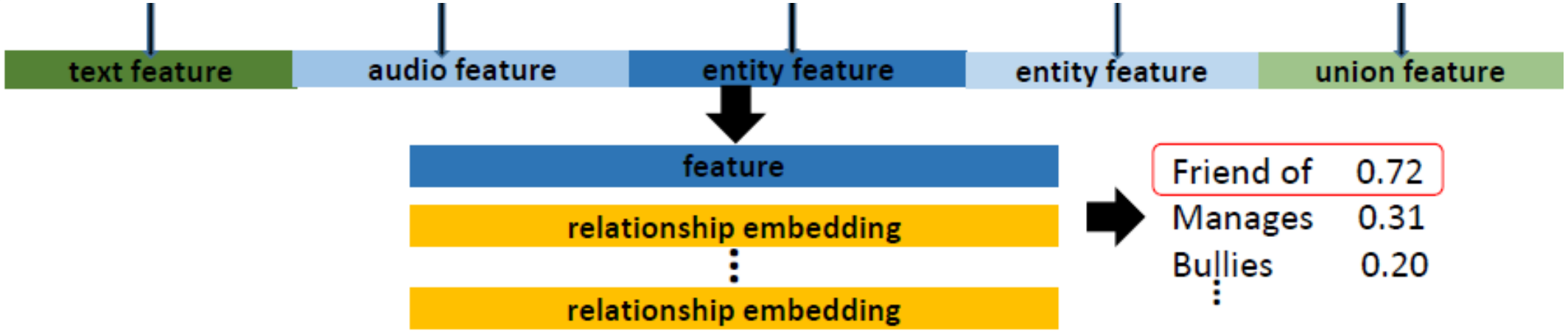
Solution

- Audio features
  - MFCC features
  - LMFE features
  - First and second differential feature



- Vision features
  - entity features
  - union features
  - C3D model

Solution



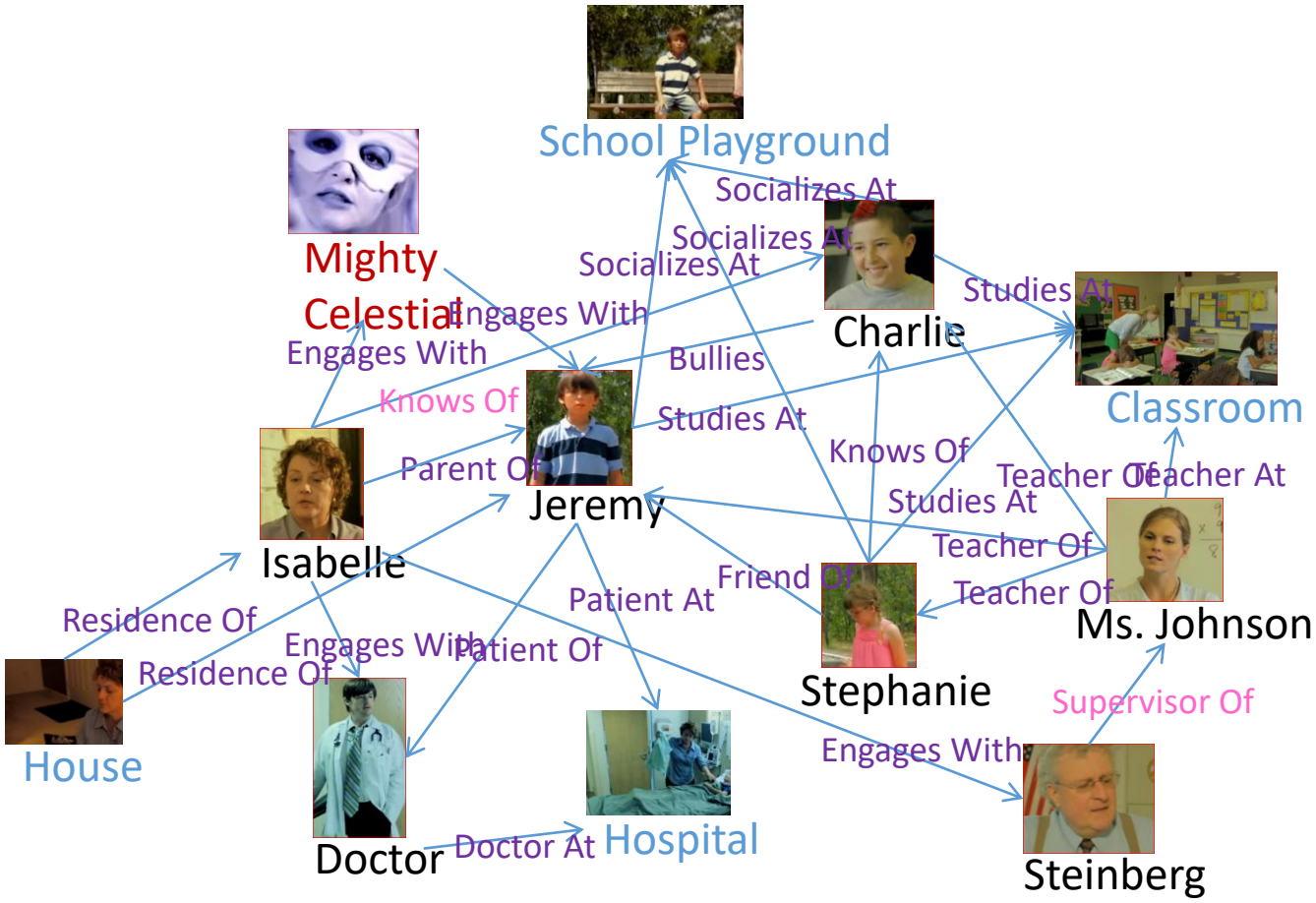
- Multimodal features fusion
- Zero-shot learning

$$L = (1 - cos(\beta, \gamma))^2 + \frac{\sum_{i \in U} (cos(\beta, \mu_i) + 1)^2}{n}$$

- L denotes the total loss
- β denotes the feature of pair
- γ denotes the feature of the positive relationship
- U denotes the set of negative relationships
- μi denotes the feature of relationship i
- n denotes the number of negative relationships

Solution

- Entity-relationship graph
  - predict relationships between each two entities in a scene
  - construct entity-relationship graph in a scene
  - merge graphs in scenes and prune by threshold
  - complete graph by rules about category and name
    - person-location, person-person, location-location
    - xxx's father – xxx
    - xxx Zimmerman – xxxx Zimmerman
    - ……



Solution

- Query answering
  - **Fill in the graph space**
    - sort the candidates in the entity-relationship graph according to scores generated by our method.
  - **Question answering**
    - plug each choice into question and check whether the graph is satisfied.
    - If none of the choices can fit our graph, choose a reasonable answer based on the types of entities and relationships.
  - **Relations between characters**
    - collect the paths between two entities by depth-first searching through the graph.

Experiments

- Entity-relationship graph
  - recall@50, recall@100, recall@θ
  - θ: number of relation triplets in ground truth
  - component analysis: difference cases using different features and rules

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

Experiments

- Query answering
  - **Fill in the graph space**: mean reciprocal rank

$$MRR = \frac{1}{\lambda}\sum_{i=1}^{\lambda}\frac{1}{\mu_i},$$

  - λ denotes the number of unknown variables
  - μi denotes the rank of right answer of ith unknown variable in the answer list

  - **Question Answering**: correct answers/number of total questions
  - **Relations between characters**: recall, precision, f1

	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

THANK YOU

MAGUS

MediA recoGnition
and UnderStanding