WHU-NERCMS @ TRECVID 2023: DEEP VIDEO UNDERSTANDING TASK

Ruizhe Li

2020300004016@whu.edu.cn

Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University

November 13, 2023

Outline

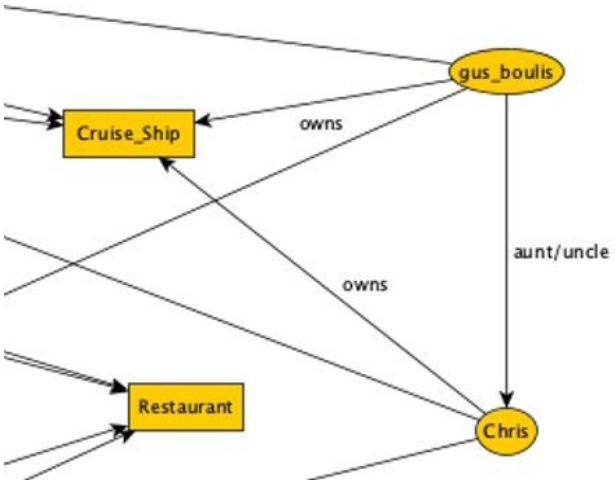
- Introduction
- Approach
- Results
- Conclusion

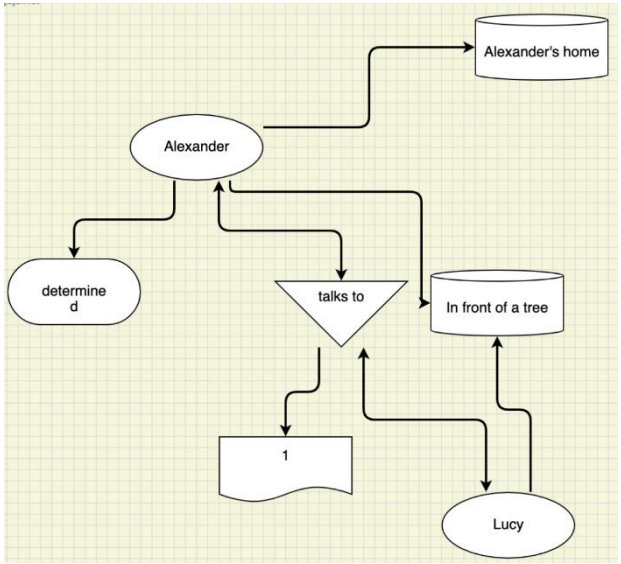
Outline

- Introduction
- Approach
- Results
- Conclusion

Introduction

- Deep Video Understanding(DVU)
  - Movie KG, entities pic, scene seg, scene KG, scene sum, vocab
  - 2 Movie-Level Groups & 2 Scene-Level Groups



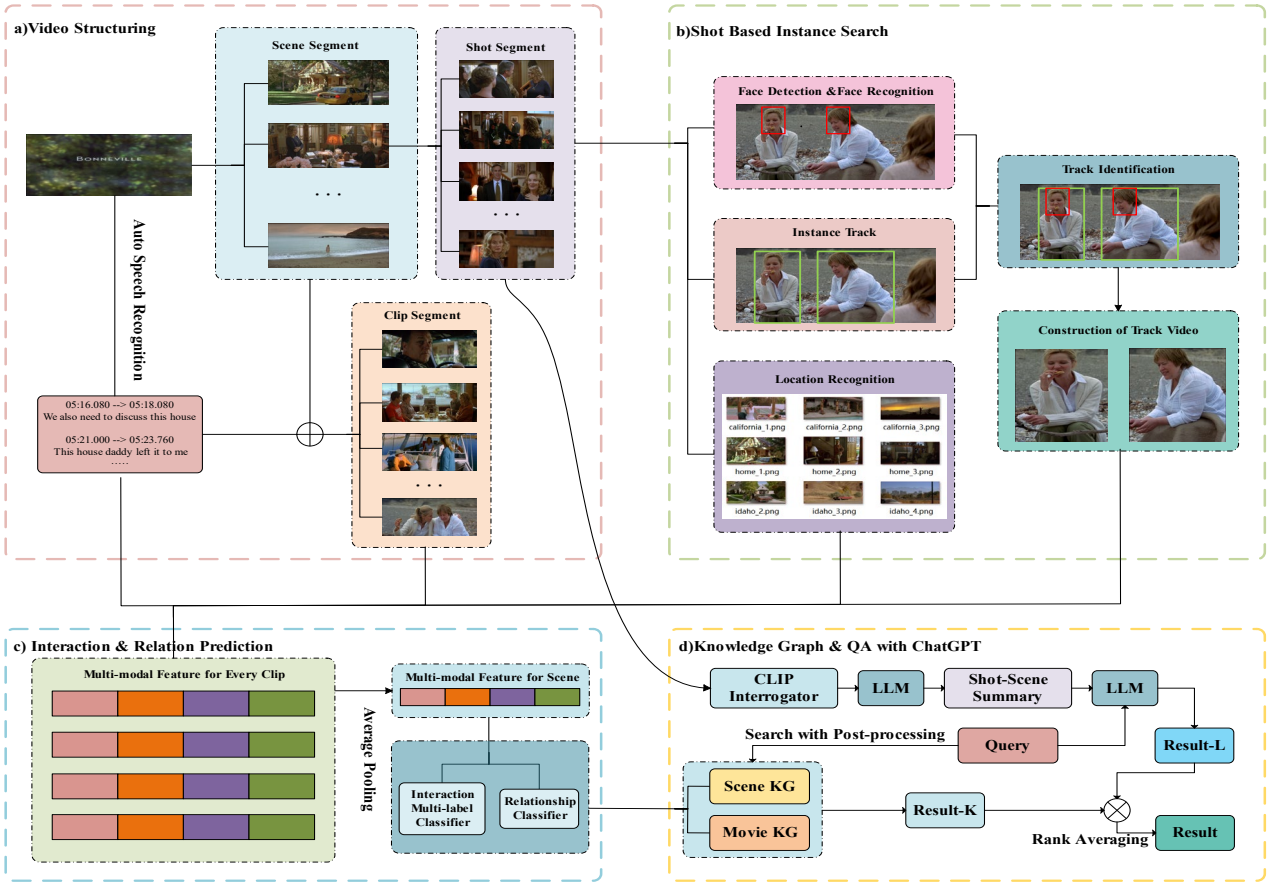


Outline

- Introduction
- Approach
- Results
- Conclusions

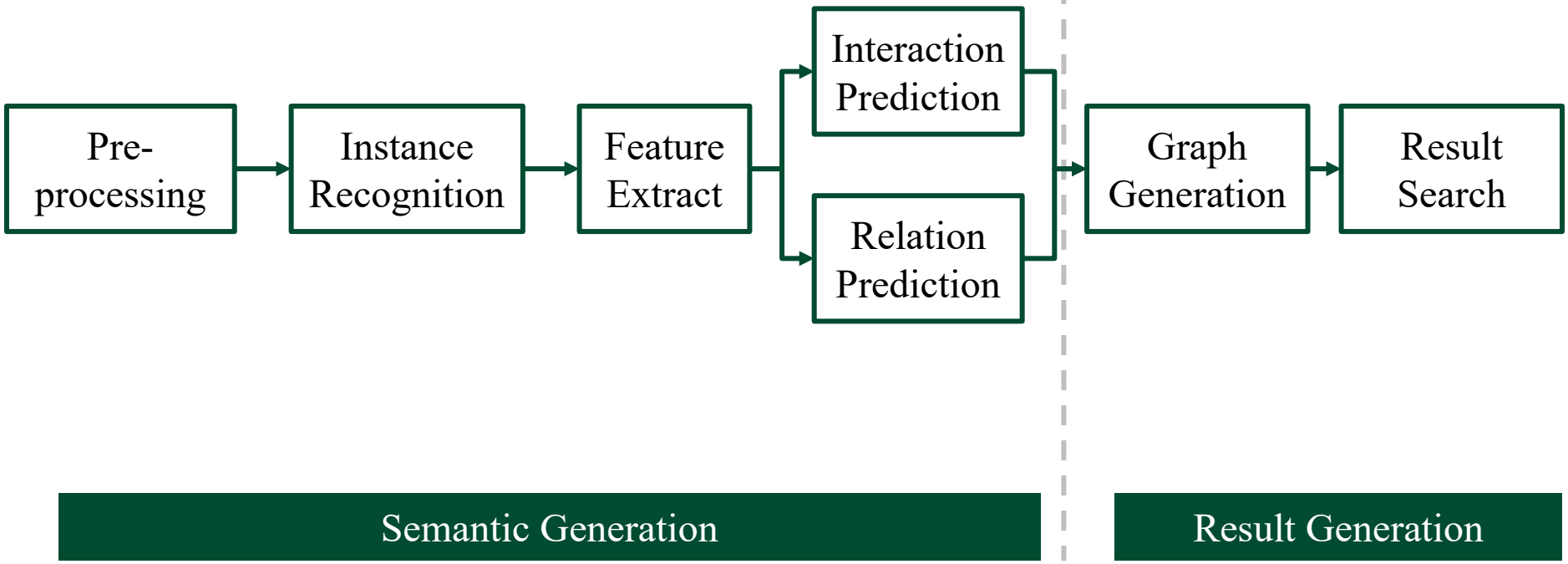
Approach

Framework

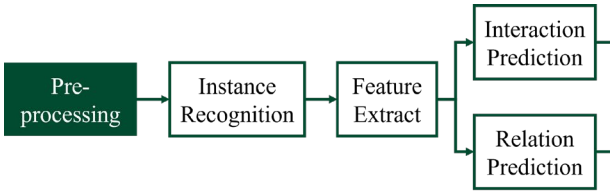


Approach

- Framework

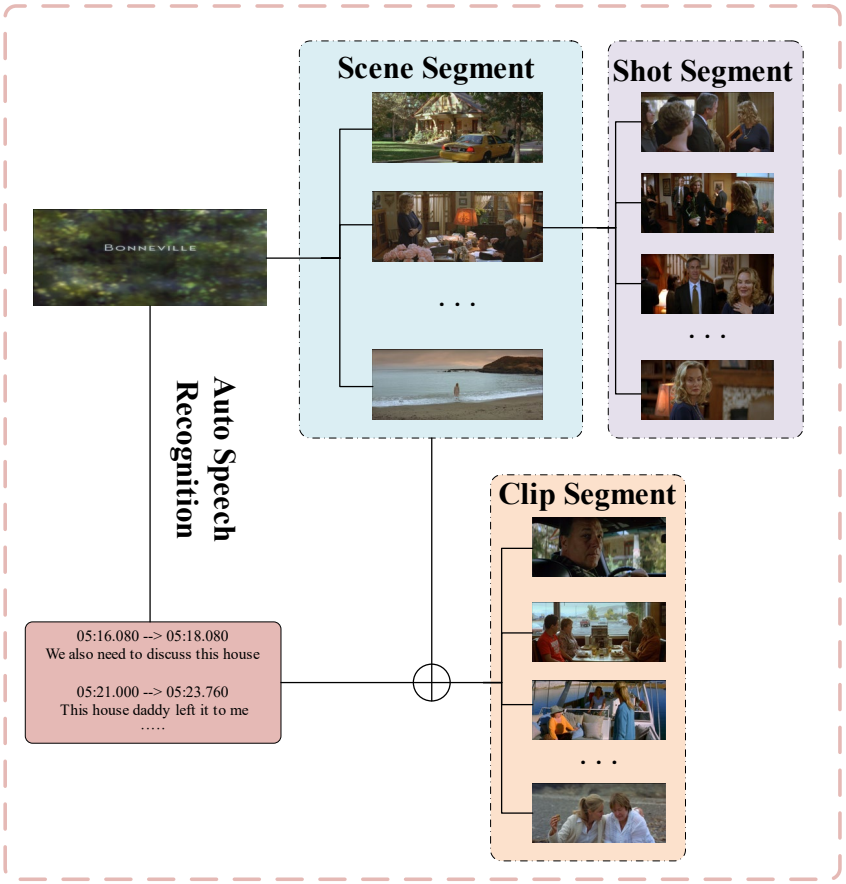


Approach

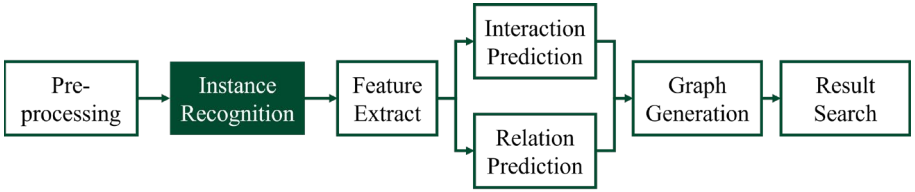


Step 1: Pre-processing

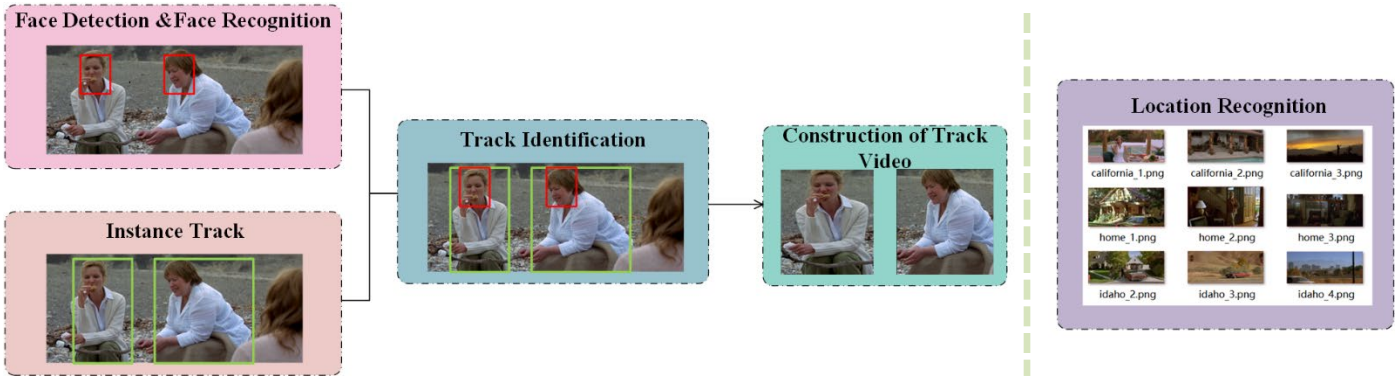
- Segmentation
  - ➢ Scene segmentation
    - a. Download scene files.
    - b. Seg with timestamps locally.
  - ➢ Shot segmentation
    - a. Shot detection & seg
  - ➢ Clip segmentation
    - a. Use YouTube ASR to generate subtitles.
    - b. Seg with timestamps of subtitles.



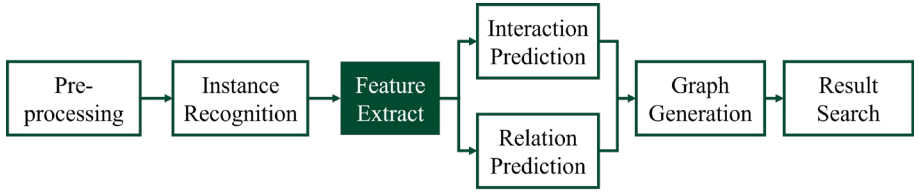
Approach



- Step 2: Instance Recognition (shot-based)
  - Person Recognition and Track(Shot2Scene)
    - Person recognition: SCRFD + Arcface + Extended Face Database
    - Person Track: faster RCNN + Deepsort
    - Trajectory in Scene: Track + Face Indentity(Voting Mechanism)
  - Construction of Track Video
    - Seg Person Track Video with Clip Timestamp
  - Location Recognition
    - Resnet + extended location database

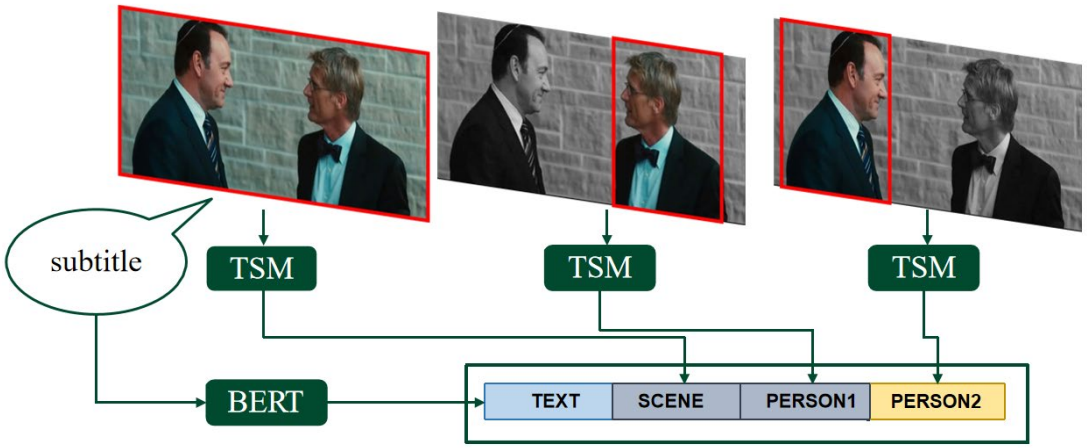


Approach

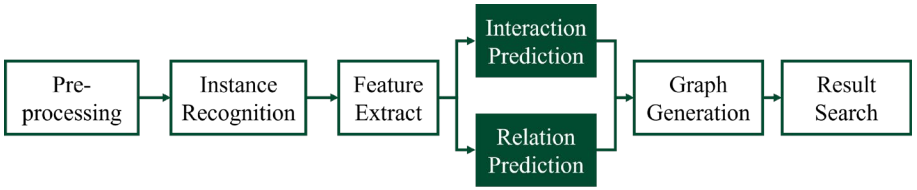


Step 3: Feature Extract

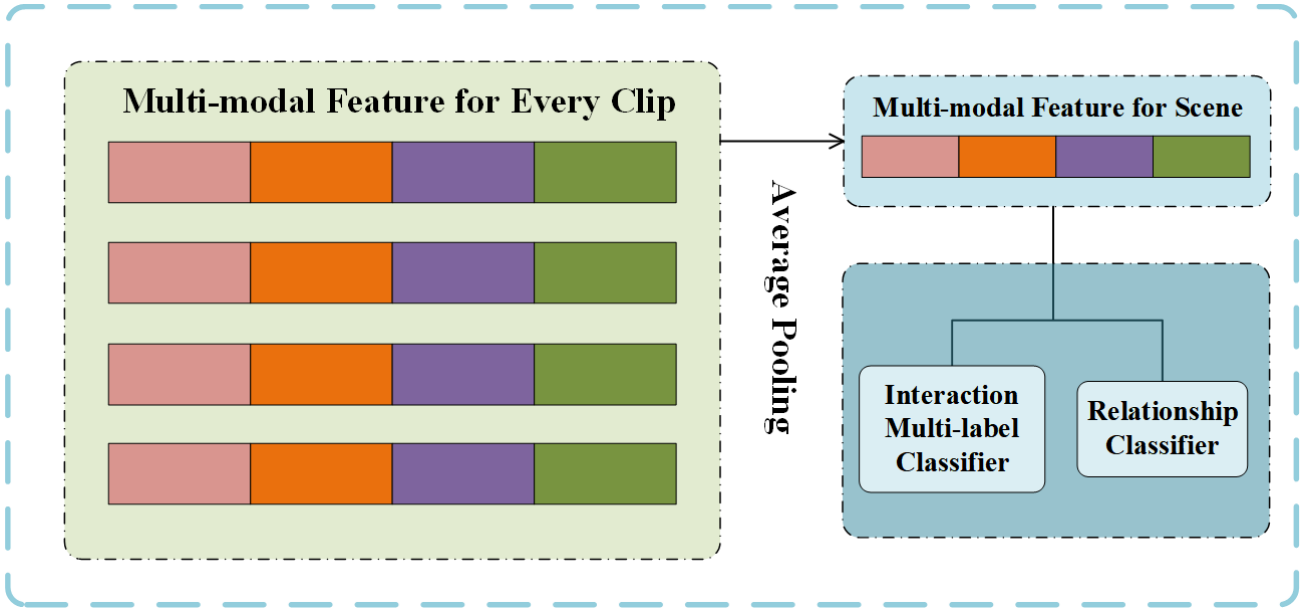
- Feature Extract
  - Text feature: Bert-base extracts a feature of 768 dimensions for a clip.
  - Visual feature: TSM extracts a feature of 2048 dimensions for a clip.
  - Track feature: Unite results to generate a feature of 2048*2 dimensions for a Person-Person/Person-Location pair in a clip.



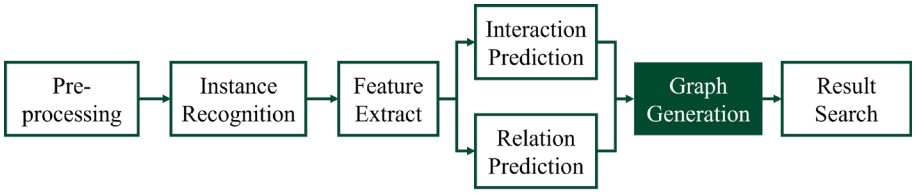
Approach



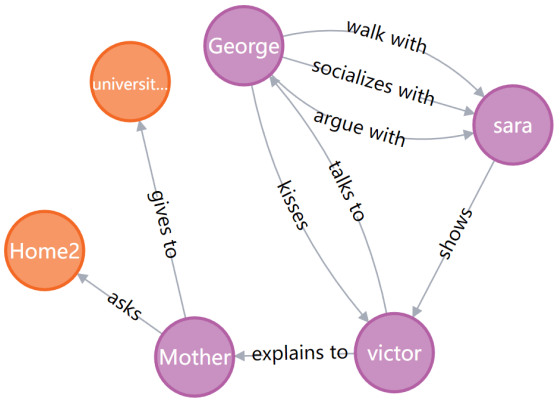
- Step 4: Interaction & Relation Prediction

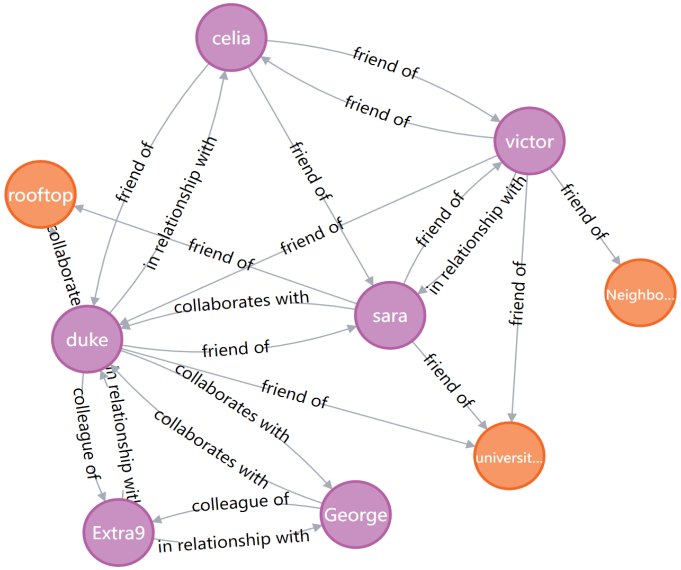


Approach

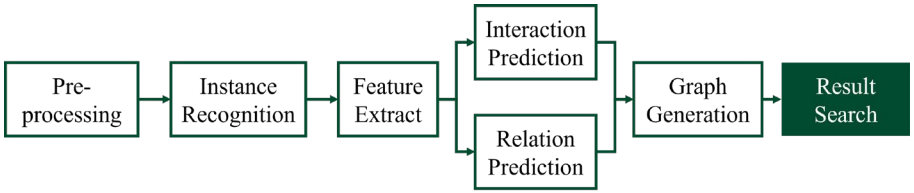


- Step 5: Graph Generation
  - KG Tool: neo4j
    - The interactions and relations are saved in the different graphs
    - **Nodes:** the person node and location node
    - **Lines:** recognized interaction in interaction part; recognized relation in relation part

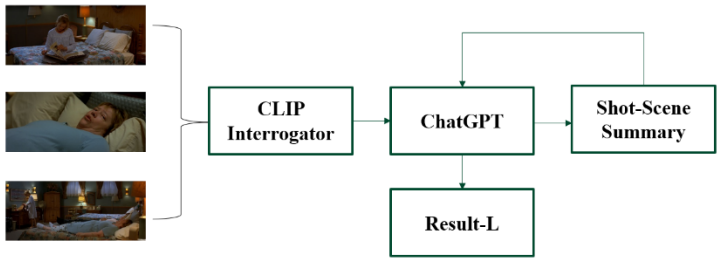


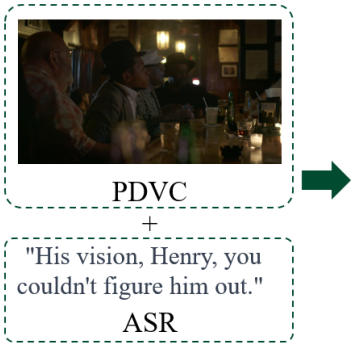


Approach



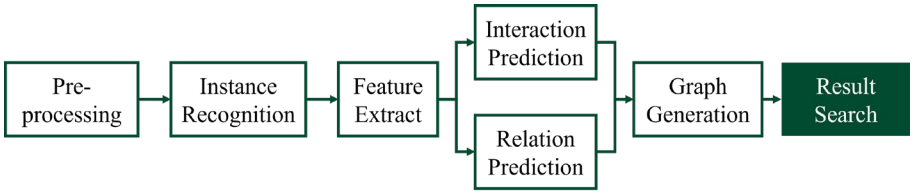
- Step 6: Result Search (KG search + LLM)
  - ChatGPT delete wrong candidates in P2P or P2L questions
  - ChatGPT(shot-scene summary)
  - ChatGPT(video caption, ASR) generates scene summary





Plot Summary: In this scene, Willis Earl Beal is seen walking around a sidewalk and the surrounding area. He interacts with the camera and shows off the ground. There are no specific speeches or dialogues mentioned in the captions. The scene also includes a boy who appears briefly.

Approach



- Step 6: Result Search
  - Movie-level Track
    - Group 1

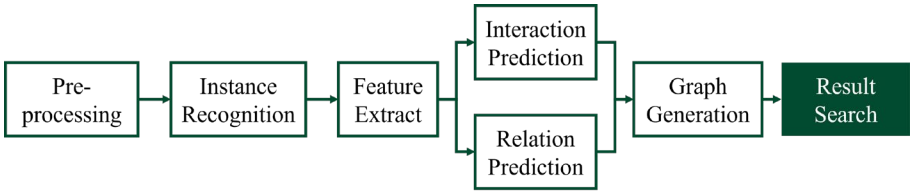
```
<DeepVideoUnderstandingTopicQuery question="2" id="1">
<item subject="Person:Rabbi_Brookstein"
predicate="Relation:Unknown_1" object="Person:Debbie"/>
<item description="What is the relation / connection
from Rabbi_Brookstein to Debbie?"/>
<Answers>
<item type="Person" answer="Apprentice Of"/>
<item type="Person" answer="Has Met"/>
<item type="Person" answer="Parent Of"/>
<item type="Person" answer="Takes Care Of"/>
<item type="Person" answer="Child Of"/>
<item type="Person" answer="Mentor Of"/>
</Answers>
</DeepVideoUnderstandingTopicQuery>
```

KG SQL search sequences:

```
MATCH (a:person_id)-[r]->(b:person_id) where a.name='Rabbi_Brookstein' and b.name='Debbie' and r.type='rela'
    return type(r),r.score
order by toInteger(r.score)DESC
```

Using ChatGPT to narrow down the answer space

Approach



Step 6: Result Search

- Movie-level Track
  - Group 2

```
<DeepVideoUnderstandingTopicQuery question="2" id="17">
<item description="Why did Christopher accompany the family on the holiday?"/>
<Answers>
<item answer="he is a cousin of Patricia"/>
<item answer="he is an artist and was hired to give painting lessons"/>
<item answer="he is a friend of Edward"/>
<item answer="he is William's brother"/>
<item answer="he is a neighbor of Cynthia"/>
<item answer="he is Cynthia's romantic partner"/>
</Answers>
</DeepVideoUnderstandingTopicQuery>
```

GPT solves the problem:

scene_1 summary: ……
scene_2 summary: ……

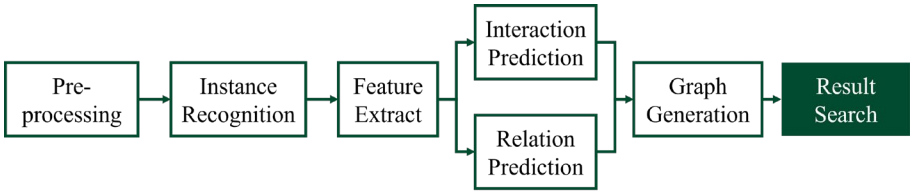
scene_n summary: ……

question_k: <DeepVideo……

GPT

answer_k: "he is William's brother"

Approach



- Step 6: Result Search
  - Scene-level Track
    - Group 1

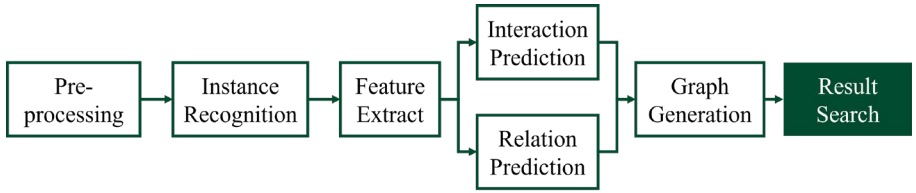
```
<DeepVideoUnderstandingTopicQuery question="2" id="1">
<item subject="Person:Debbie" scene="18" predicate="Interaction:talks to"
object="Person:Co-Worker"/>
<item description="In Scene 18, Debbie talks to Co-Worker. What is the  immediate
next / following interation between Co-Worker and Debbie, in  scene 18?"/>
<Answers>
<item type="Interaction" scene="18" answer="greets"/>
<item type="Interaction" scene="18" answer="hits"/>
<item type="Interaction" scene="18" answer="asks"/>
<item type="Interaction" scene="18" answer="shoots"/>
<item type="Interaction" scene="18" answer="talks to"/>
<item type="Interaction" scene="18" answer="yells at"/>
</Answers>
</DeepVideoUnderstandingTopicQuery>
```

KG SQL search sequences:

```
match(a:person_id)-[r]->(b:person_id) where r.scence='18' and a.name='Debbie' and b.name='Co-Worker'
return id(r),type(r) order by id(r)
```

Using shot-scene summary for ChatGPT to generate another answer list

Approach



- Step 6: Result Search
  - Scene-level Track
    - Group 2

```
<DeepVideoUnderstandingTopicQuery question="5" id="6">
<item subject="Scene:Unknown" predicate="Description"/>
<item description="Patricia has an argument with her husband on the phone before joining Edward
and Cynthia for dinner."/>
<Answers>
<item type="Integer:Scene" answer="12"/>
<item type="Integer:Scene" answer="47"/>
<item type="Integer:Scene" answer="44"/>
<item type="Integer:Scene" answer="35"/>
<item type="Integer:Scene" answer="32"/>
<item type="Integer:Scene" answer="28"/>
<item type="Integer:Scene" answer="22"/>
<item type="Integer:Scene" answer="1"/>
<item type="Integer:Scene" answer="18"/>
<item type="Integer:Scene" answer="15"/>
</Answers>
</DeepVideoUnderstandingTopicQuery>
```

Treat it as a video retrive task

Outline

- Introduction
- Approach
- Results
- Conclusion

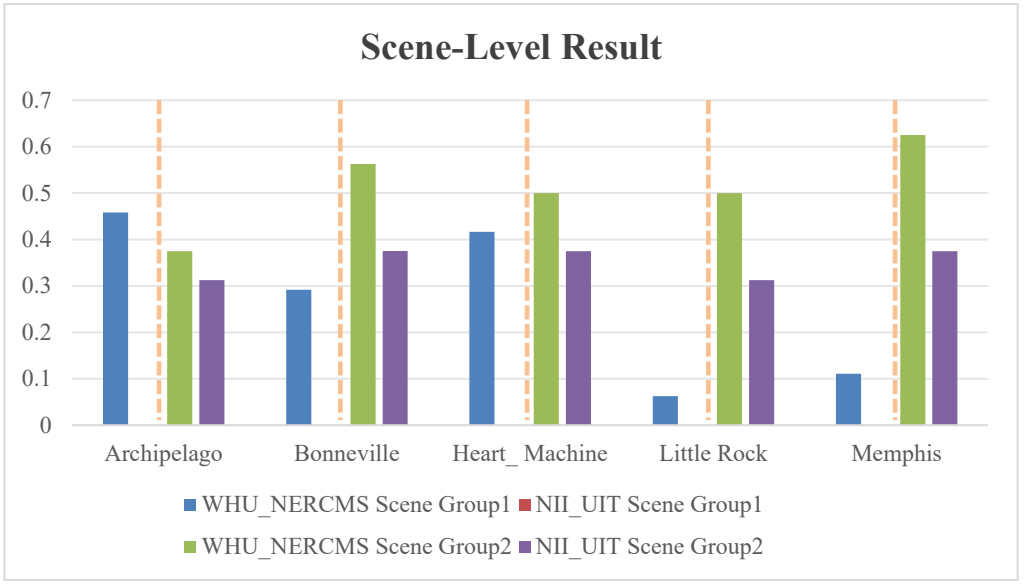
Results

- Overall Result
  - Movie-level & Scene-level

Result in TRECVID 2023 Grand DVU Task(%)						
Movie	Scene-level				Movie-level	
	s1(MRR)	s2&s3(ACC)	s4(ACC)	s5(ACC)	s1(MRR)	s2(ACC)
Archipelago	12.5	62.5	40	33.3	81.25	33.3
Bonneville	37.5	25	50	66.7	62.5	39.1
Heart_ Machine	25	50	60	33.3	37.5	61.5
Little Rock	18.8	0	60	33.3	62.5	30
Memphis	8.32	12.5	60	66.7	54.2	59.1
total	26.8(Group1)		51.2(Group2)		59.6(Group1)	43.7(Group2)

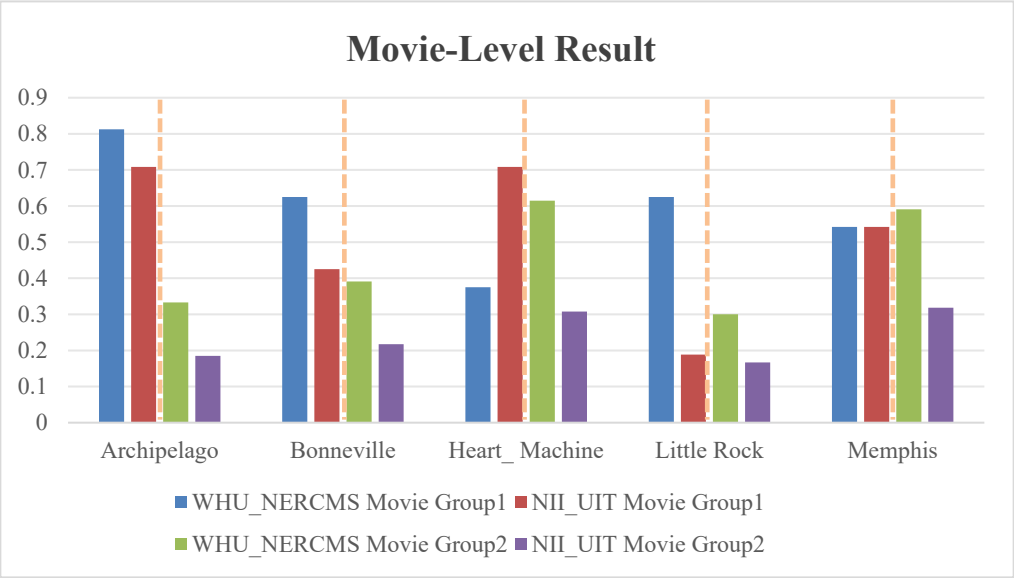
Results

- Overall Result
  - Compare with Other team



Results

- Overall Result
  - Compare with Other team

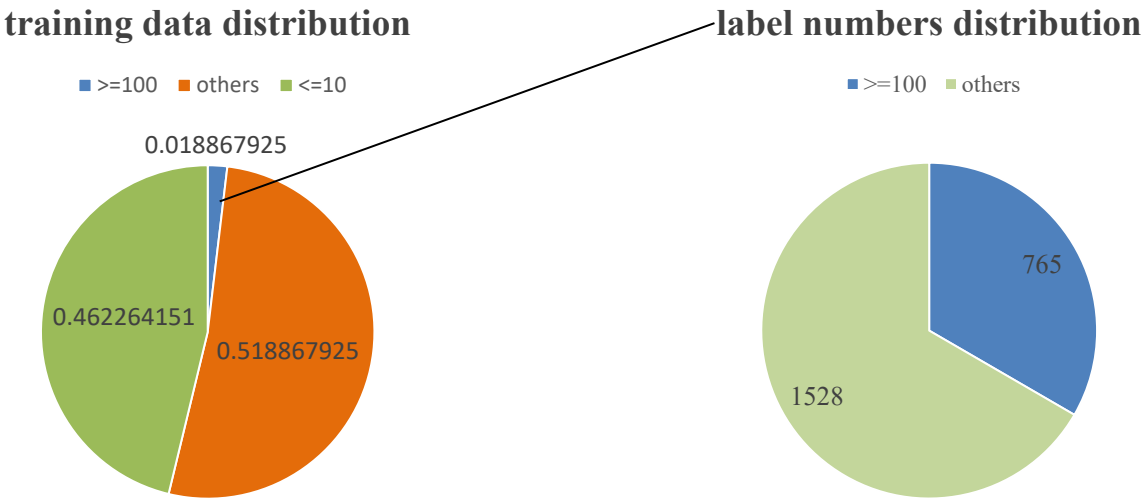


Outline

- Introduction
- Approach
- Results
- Conclusions

Conclusion

- Some actions are hard to distinguish
  - Such as “talks to” “ask” “yell at”
- Annotation is inadequate
  - The labels in the training data have a long-tailed distribution



Thanks for your time!

Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University