gch3339

ACM multimedia
Chengdu, China OCT 20-24 2021

Joint Learning for Relationship and Interaction Analysis in Video with Multimodal Feature Fusion

BeibeiZhang[1] , Fan Yu,[1,2] Yanxin Gao,[1]
Tongwei Ren,[1,2,*] Gangshan Wu[1]

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

[2]Shenzhen Research Institute of Nanjing University

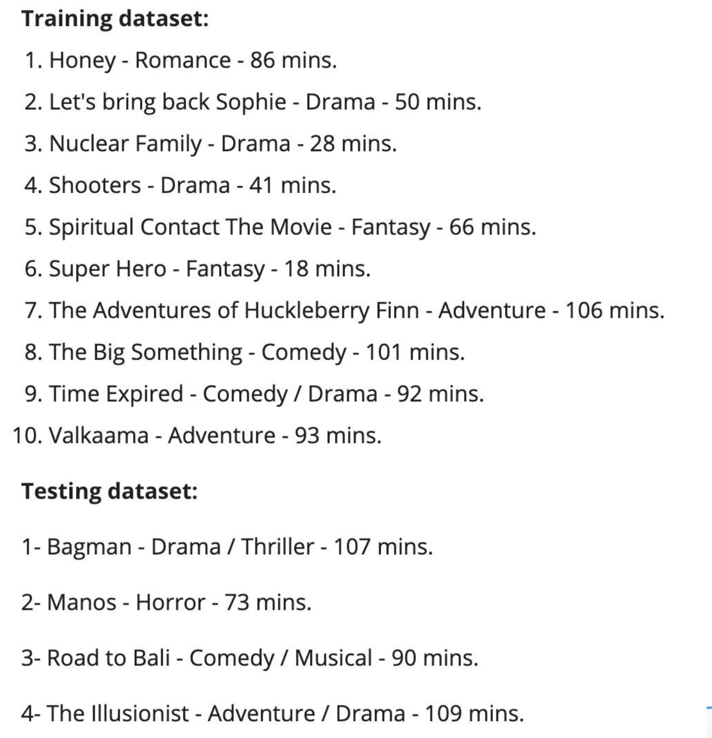
南京大學
NANJING UNIVERSITY

MAGUS
MediA recoGnition and UnderStanding

gch3339

Introduction

- **Deep video understanding (DVU)**
  - requires systems to develop a deep analysis and understanding of long video.
  - use known information to reason about other, more hidden information, and to populate a knowledge graph (KG) with all acquired information.
- **HLVU** dataset
  - 14 videos
    - 10 for development
    - 4 for test
  - 1h/video in average
  - shot, entity name, entity type, screenshots



Training dataset:

1. Honey - Romance - 86 mins.
2. Let's bring back Sophie - Drama - 50 mins.
3. Nuclear Family - Drama - 28 mins.
4. Shooters - Drama - 41 mins.
5. Spiritual Contact The Movie - Fantasy - 66 mins.
6. Super Hero - Fantasy - 18 mins.
7. The Adventures of Huckleberry Finn - Adventure - 106 mins.
8. The Big Something - Comedy - 101 mins.
9. Time Expired - Comedy / Drama - 92 mins.
10. Valkaama - Adventure - 93 mins.

Testing dataset:

1- Bagman - Drama / Thriller - 107 mins.

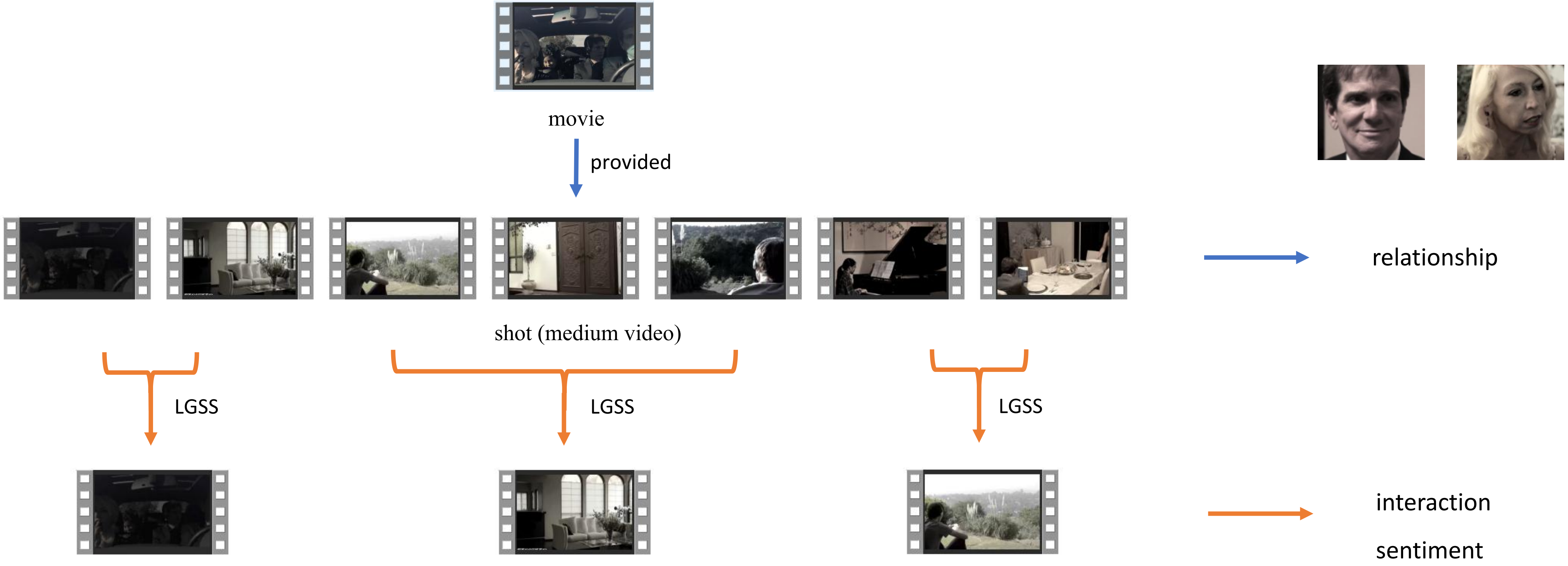
2- Manos - Horror - 73 mins.

3- Road to Bali - Comedy / Musical - 90 mins.

4- The Illusionist - Adventure / Drama - 109 mins.

Solution

- Video segmentation



movie

provided

relationship

shot (medium video)

LGSS

LGSS

LGSS

interaction

sentiment

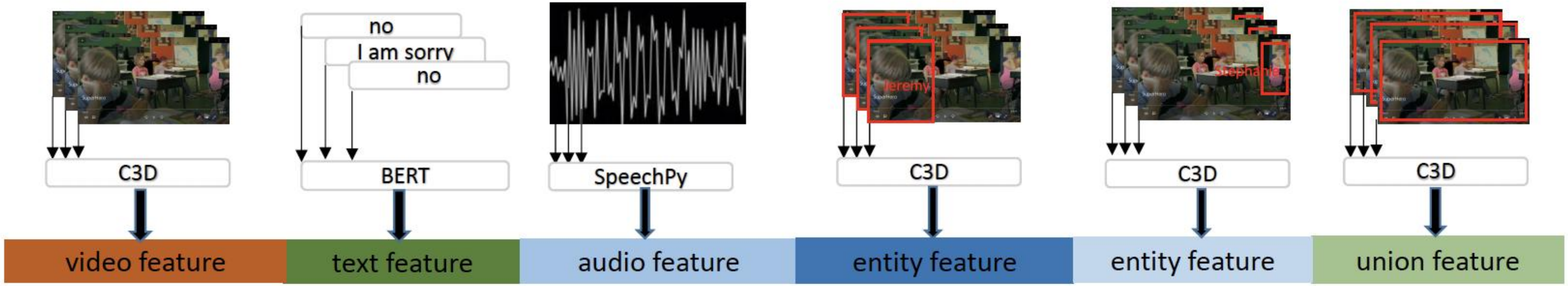
long video

gch3339

Solution

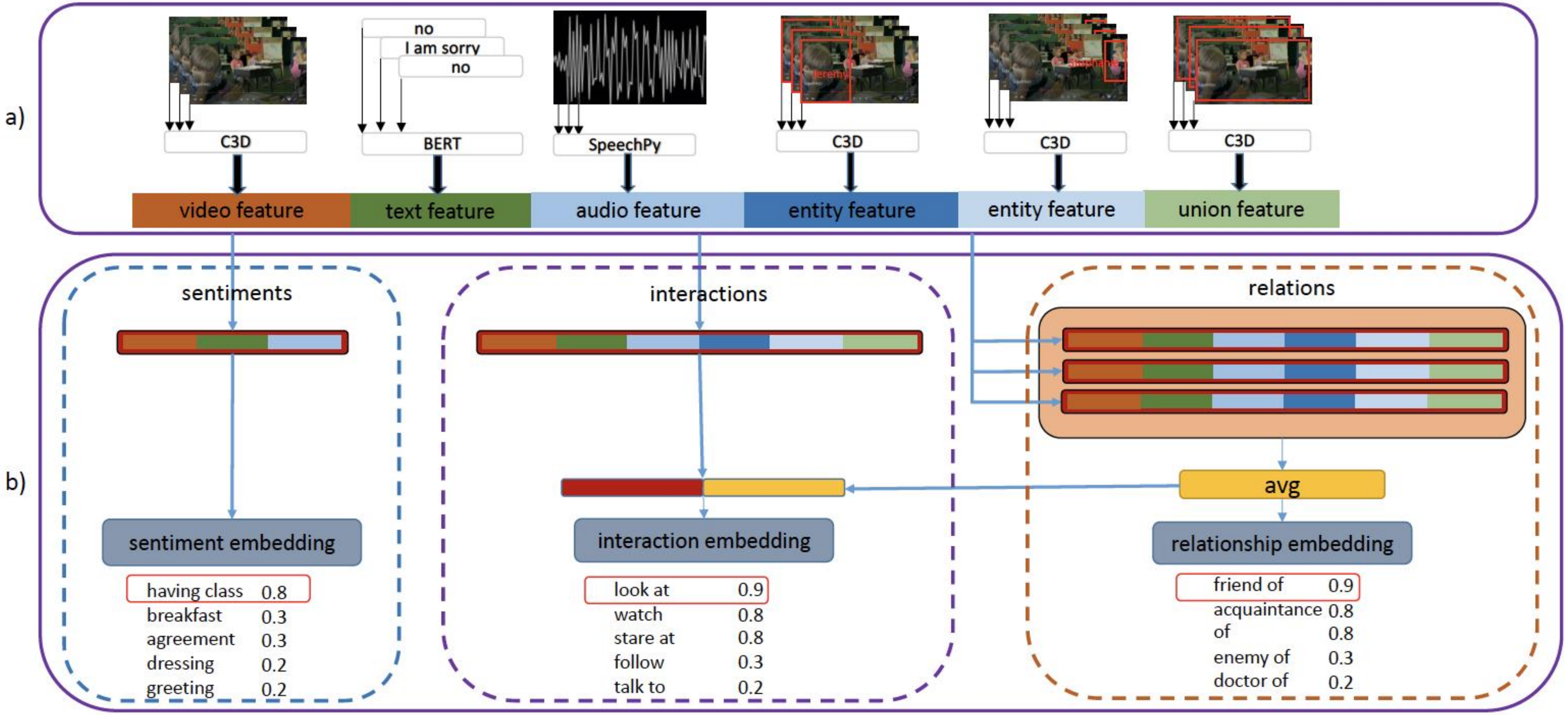
- **Video feature**
  - C3D
- **Audio feature**
  - MFCC, LMFE
- **Text feature**
  - BERT

- **Entity feature (subject, object, union)**
  - CenterTrack
  - InsightFace
  - C3D



Solution

- Joint learning architecture
  - relationship: average of medium video feature
  - interaction: medium video feature + average feature



gch3339

Solution

- Low-shot, Zero-shot learning
- Joint learning

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

- l denotes loss
- denotes the feature of pair
- denotes the feature of the positive relationship
- denotes the set of negative relationships
- denotes the feature of relationship
- denotes the number of negative relationships

$$L = l_R + \frac{1}{n}\sum(l_I + l_S)$$

- denotes the total loss
- l_R denotes the loss of relationship
- $l_I$denotes the loss of interaction
- $l_S$denotes the loss of sentiment

Query answering

- **movie-level**
    - Find all possible paths question.
    - Fill in the part of graph question.
    - Multiple choice questions.
        - relationship knowledge graph
- **scene-level**
    - Find the unique scene.
    - Fill in the graph space.
        - interaction knowledge graph
    - Find next/previous interaction in scene X between person Y and person Z.
        - split medium video into shot videos
    - Find the 1-to-1 relationship between scenes and natural language descriptions
        - match with predicted interactions and sentiments.
    - Classify scene sentiment from a given scene.
        - sentiment model

gch3339

THANK YOU

南京大學
NANJING UNIVERSITY

MAGUS
MediA recoGnition
and UnderStanding