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

Beibei Zhang[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, Nanjing, China
[2] Shenzhen Research Institute of Nanjing University, Shenzhen, China
{zhangbb,yf,gyx}@smail.nju.edu.cn,{rentw,gswu}@nju.edu.cn

ABSTRACT

To comprehend long duration videos, the deep video understanding (DVU) task is proposed to recognize interactions on scene level and relationships on movie level and answer questions on these two levels. In this paper, we propose a solution to the DVU task which applies joint learning of interaction and relationship prediction and multimodal feature fusion. Our solution handles the DVU task with three joint learning sub-tasks: scene sentiment classification, scene interaction recognition and super-scene video relationship recognition, all of which utilize text features, visual features and audio features, and predict representations in semantic space. Since sentiment, interaction and relationship are related to each other, we train a unified framework with joint learning. Then, we answer questions for video analysis in DVU according to the results of the three sub-tasks. We conduct experiments on the HLVU dataset to evaluate the effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Deep video understanding; relationship analysis; interaction analysis; multimodal feature fusion

1 INTRODUCTION

Videos are typical unstructured data that are hard to be managed and exploited. However, along with the development of video acquisition equipment, the number of videos increases rapidly and the requirement for managing and exploiting videos is also ever-growing. Many applications require structurization and analysis of videos, such as video retrieval [15] and video summarization [13].

The Deep Video Understanding (DVU) task [4] is proposed to deeply analyze long duration videos and extract knowledge graphs in videos, whose nodes are persons, locations and concepts. The DVU task segments long duration videos into several scenes and proposes to answer a series of questions on movie level and scene level to analyze movies with several snapshots for each entity. The three types of movie-level questions are follows: 1) Find all the possible paths from the source person to the target person. 2) Fill in the unknown part of graph. 3) Choose correct answer for the unknown part from multiple options. There are six types of questions on the scene level: 1) Find the unique scene. 2) Fill in the graph space. 3) Find the next interaction between given persons in a given scene. 4) Find the previous interaction between given persons in a given scene. 5) Find the relationship with given natural language descriptions. 6) Classify scene sentiment from a given scene. The description of the DVU task is shown in Figure 1.

There also exist some tasks for structuring and analyzing videos, such as video visual relation detection [14], human-object interaction detection [7] and social relation recognition [16]. However, all these tasks focus on relationships between instances especially persons but the entities that contain locations and concept for DVU are not included. Moreover, these tasks are commonly based on shot videos while DVU requires not only interaction graphs on video clips but also a relationship graph on the whole video.

In this paper, we propose a method using joint learning and multimodal feature fusion to address the DVU task. We classify the questions proposed by DVU into three types of sub-tasks: scene sentiment classification, scene interaction recognition and super-scene video relationship recognition. Two preprocesses need to be performed before the three sub-tasks. Firstly, we segment the given scene clips into several shot clips and merge the given scene clips into super-scene clips. Then, the corresponding entities are tracked in shots, scenes and super-scenes. All the three sub-tasks fuse visual features, text features and audio features. For scene sentiment classification, visual features, audio features and text features of the scene video clip are combined to predict scene sentiment representation in semantic space. For scene interaction recognition, in addition to the above-mentioned features, visual features of entities and entity pairs are integrated to predict scene interaction representation in semantic space. For super-scene video relationship recognition, mean pooling is performed on the integrated features of the corresponding scenes and super-scene relationship representation is predicted in semantic space. The movie-level questions are answered according to the merged super-scene relationship recognition results and the scene-level questions are answered according to the results of scene interaction recognition and scene sentiment classification.

2 PRELIMINARY

Shot and scene segmentation. A long duration video is usually composed of many scenes, which describe a complete event, and a scene is composed of many shots, which contain a complete camera motion. Segmenting videos into shots is usually based on both camera and scene motion, especially the combinations of static vs. dynamic camera and static vs. dynamic scene [9]. Scene segmentation is more challenging because scenes in videos often contain abundant temporal structures and complex semantic information. Rao *et al.* [12] propose a local-to-global scene

*Corresponding author.

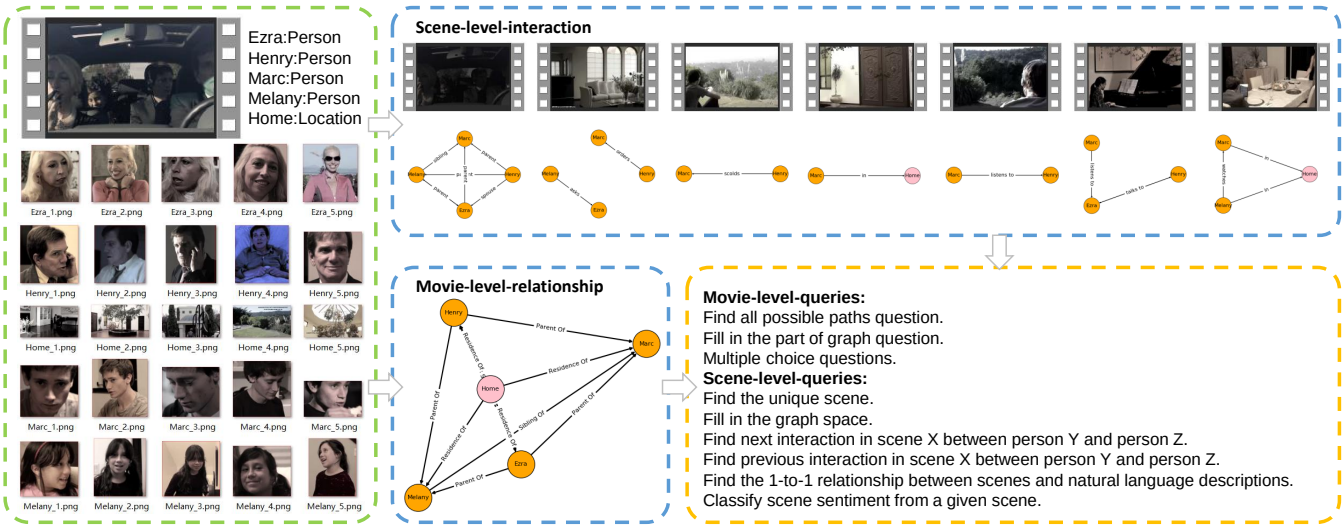

Figure 1: The description of deep video understanding task.

segmentation framework that can integrate multimodal information across three levels: clip, segmentation and movie. Chen *et al.* [3] propose a self-supervised shot contrastive learning approach to learn a shot representation that maximizes the similarity between nearby shots compared to randomly selected shots.

Entity search in video. A straightforward method to search entities in video is to match entity templates and track the entities. Some template matching methods, e.g., cross correlation matching, correlation coefficient matching and square difference matching, are easy to make mistakes if the template has compression distortion. Feature matching methods use scale invariant features like speed up robust features (SURF) [1] to detect local feature of an image. To track objects in videos, methods that combine static and adaptive template tracking have been proposed [11] as well as methods that recognize "reliable" parts of the template [8]. Later, some methods propose to track objects by detection, such as [2] and [21]. Specially, for searching persons in videos, InsightFace [5] matches detected faces with face samples by integrating face detection and face alignment into a framework like the MTCNN model [20], and OMS [18] achieves online searching by exploiting a dynamic memory bank to store face, body and audio features of persons.

Feature extraction from video. Since videos contain different modals, videos can be represented with multimodal features, among which visual features and audio features are straightforward. Similar to image feature extraction, visual features of videos can also be extracted by convolutional networks, e.g., deep 3-dimensional convolutional networks (C3D) [17] are proposed for spatiotemporal feature learning. To extract semantic information from audio, MFCC and and LMFE are two important audio features that could be extracted by SpeechPy, a useful tool for speech processing and feature extraction. In addition to audio features, features from speech text are also essential. After transforming speech in videos to text, models like BERT [6] can be used to extract text features of words and sentences.

Interaction and Relationship recognition. A general task for interaction and relationship recognition in video is video visual relation detection [14], which aims to detect the interactions between general instances. Another typical task named human-object interaction detection [7] focuses on recognizing human actions with objects. Since interactions between persons imply relationships between them, social relationship recognition is proposed [16]. Specially, Kukleva *et al.* [10] focus on the interactions and social relationships between characters in a movie and predict interactions and social relationships with visual and language cues by joint learning. Yu *et al.* [19] propose a solution to the previous DVU task with multimodal feature fusion.

3 OUR METHOD

We think that the relationships and interactions between characters, and sentimens of the video are all influenced by each other, and therefore we propose a joint learning method to simultaneously train and infer relationships, interactions and sentiments. As the relationship between characters is inferred on the basis of long-time video which consists of multiple short-time videos, we divide the video into multiple super-scenes to obtain the relationship between characters. Each super-scene can be divided into multiple scenes from which we can obtain interaction and sentiment. We take the official video segment as scene and aggregate them into multiple super-scenes according to the multi-modal features of scene with the help of SceneSeg LGSS [12]. Compared to [19], we reuse multimodal features, but we propose a new joint learning method to predict relationships, interactions and sentiments in order to deal with more complicated questions.

3.1 Multimodal Features

As shown in Figure 2 (a), We use multi-modal features, including visual features, audio features and text features, as the scene feature to predict relationships and interactions between characters.

Visual features Video features, entity features and union features are all visual features and they are obtained as follows:

- We use C3D to extract video features directly.

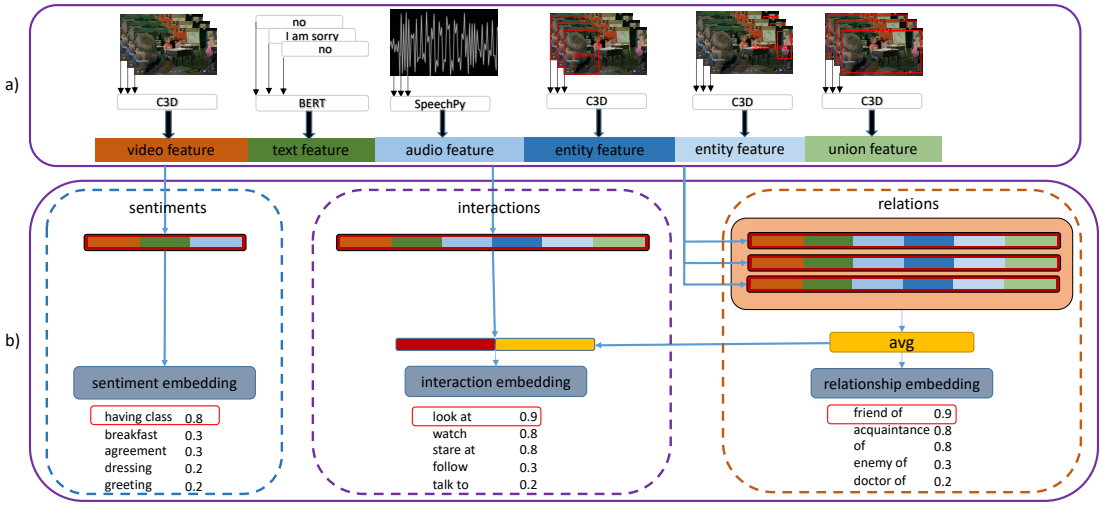

Figure 2: The pipeline of our method. We use a joint learning method to obtain relationships and interactions between characters and sentiments at the same time. Multimodal features of the video are the input and the output is the relationships between characters and sentiments of the super-scene and the interactions between characters of the scene.

- We use SURF to match locations and people with the screen shots, and then entity bounding boxes can be tracked by face recognition method InsightFace [5] and multiple object tracking method CenterTrack [21]. According to the algorithm of average interval sampling from entities' traces in each scene, the visual features of entities in the corresponding scenes are generated from the C3D model.
- We can also compute the union bounding boxes of two entities and take an average sample in each scene as the input of C3D to obtain the union features.

Audio features We extract MFCC and LMFE features by SpeechPy, and then calculate their first and second differential features respectively. Finally, the feature cubes of MFCC and LMFE are joint together to represent the audio features.

Text features We use the speech-to-text tools provided Youtube to obtain the lines of a movie. Then, we match the lines to the scene to obtain sentences according to the time they appear. Finally, we use BERT [6] to convert sentences to vectors as text features.

All the features are then concatenated together and transformed into a feature whose dimension is the same as that of a text feature for relationship prediction between entities.

3.2 Joint learning

In Figure 2(b), we average the features of multiple scenes as the feature of super-scene to predict the relationship between characters. After that, we concatenate the super-scene feature and scene feature as the final scene feature to predict interaction between characters, which shows the influence of character relationship on character interaction. Since sentiment is about the whole scene, we only use the fusion of video, audio and subtitle features to predict sentiment. Finally, we encode the feature and calculate the similarity to target relationship, interaction and sentiment. In this way, we get the relationships and interactions between characters, and sentiments to answer questions.

3.3 Training and Inference

For the absence of some labels in the training set, we use zero shot learning during training. The final descriptions of relationship, interaction and sentiment are used to generate features in the same way as the generation of text features from subtitles. Cosine similarity of the feature representing scene/shot and the target relationship/interaction/sentiment features are computed. The loss function is computed as follows:

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

where l denotes the loss, β denotes the feature of scene/shot; γ denotes the feature of the positive relationship/interaction/sentiment; U denotes the set of negative relationship/interaction/sentiment; μ_i denotes the feature of relationship/interaction/sentiment i; n denotes the number of negative relationships/interactions/sentiments. Since each super-scene consists of multiple scenes, we take the average of the sum of interaction and sentiment loss of each scene to add the relationship loss of the super-scene as the total loss of the super-scene, which is computed as follows:

$$L = l_R + \frac{1}{n} \sum (l_I + l_S), \tag{2}$$

where L denotes the total loss, l_R denotes relationship loss of super-scene; l_I denotes interaction loss of scene; l_S denotes sentiment loss of scene; n denotes the number of scenes in the super-scene. During inference, the cosine similarity of super-scene/scene feature and relationship/interaction/sentiment feature is the final score.

3.4 Query Answering

In order to obtain the sequence of interactions between characters in each scene, we divide scenes into shorter shots and specify that there is only one interaction between a pair of characters in each shot. We used the same method in Figure 2a) to predict the person relationships for each scene and the interactions for each shot.

Table 1: Experiments on different variants, where R represents relationship branch, I represents interaction branch, S represents sentiment branch, *s* represents average of scene features, *ss* represents complete super-scene feature, *u* represents union feature.

Method	Recall_R	Recall_I	Recall_S
R_s+I+S	37.3	**39.9**	48.5
R_{ss}/I/S	35.5	35.9	**62.3**
R_{ss}+I+S	**38.5**	39.6	44.1
R_s+I+S_u	36.1	28.8	41.7
R_{ss}+I+S_u	35.5	28.5	40.7

We have built the final movie-level entity-relation graph and scene-level entity-interaction graphs. For the movie-level tasks, we solve them using the method that we proposed last year [19], on the basis of the entity-relation graph. For "Finding the Unique Scene question", we traverse all the entity-interaction graphs to find the most suitable scene. For "Fill in the Graph question", we adopt the same way in answering the movie-level Fill-in questions. For "Predicting the Next and Previous Interaction question", we first locate the interaction in a certain scene, and then traverse all the choices to find the best answer. For "Match Scene with Natural Language Description question", we traverse all the scene-level interaction-graphs in the options, and use the entities and interaction to match the words in the description to find the best match scene. For "Classify Sentiment Label question", we have obtained the scores of each sentiment in each scene, then we use them to find the best choice of sentiment.

4 EXPERIMENTS

4.1 Dataset and Experimental Settings

All the experiments are conducted with E5-2680 v4 2.40GHz 14 cores CPU, 64GB memory and one GeForce RTX 3090 GPU, on the HLVU dataset [4].

The HLVU dataset contains 14 movies from public websites lasting 16 hours in total. The dataset provides each movie in the development set with manually annotated knowledge graphs that contains entities and their relationships, interactions, sentiments and descriptions about the movie. The dataset also provides a set of image examples of different actors and entities including important locations, with a name ID for each entity.

In our experiments, we evaluate the performance of knowledge graphs generation using metric *Recall@k*, which is usually applied in visual relation detection. The metric *Recall@k* is computed by

$$Recall@k = \frac{TP_k}{TP_k + FN_k}, \tag{3}$$

where TP_k and FN_k denote the number of correct label predicted and unpredicted in the top k confident predictions, respectively. k is set to the number of ground truth relationships, interactions and sentiments.

4.2 Component Analysis

In order to verify the effectiveness of the joint learning architecture, we construct different combinations of relationship, interaction and sentiment branch for comparison.

In order to check the influence of character interaction on sentiment prediction, we add union feature to help predicting sentiment. For the relationship branch, we also compared the sources of the features, which are the complete super-scene and the average of scene features.

In table 1, it can be seen that the architecture of joint learning has advantages in predicting relationships and interactions between characters. Also, the performance of interaction prediction is improved when the average features of scene represent relationship feature, and relationship prediction is enhanced with the features from the complete super-scene. We think that the feature of the complete super-scene brings some unwanted noise to a single sence.

Table 1 also shows that the sentiment branch performs better when it is trained alone. In the joint architecture, we added union feature to the branch of sentiment and found that the performance decreased, which indicates that sentiment is based on the whole scene, and the effect of character pair is not so desirable. Moreover, since there are some scenes that only involve sentiment, if we just train sentiment branch with the scenes that include interactions, we will miss many samples.

Finally we use the joint learning model to answer questions about relationships and interactions between characters and use the sentiment model which is trained alone to answer questions about sentiment.

5 CONCLUSIONS

In this paper, we proposed a joint learning method using multi modal features to extract knowledge graphs of movies for deep video understanding on the basis of the analysis of the relationships and interactions among entities in movies. We evaluated our method on the HLVU dataset, and the experimental results validated the effectiveness of our method.

ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China (62072232), Natural Science Foundation of Jiangsu Province (BK20191248), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European Conference on Computer Vision*. 404–417.
[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. 2019. Tracking without bells and whistles. In *IEEE International Conference on Computer Vision*. 941–951.
[3] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. 2021. Shot Contrastive Self-Supervised Learning for Scene Boundary Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9796–9805.
[4] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *International Conference on Multimedia Retrieval*. 355–361.
[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5203–5212.
[6] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.
[7] Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. 2009. Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 10 (2009), 1775–1789.

[8] Allan D. Jepson, David J. Fleet, and Thomas F. El-Maraghi. 2003. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 10 (2003), 1296–1311.

[9] Adarsh Kowdle and Tsuhan Chen. 2012. Learning to segment a video to clips based on scene and camera motion. In *European Conference on Computer Vision*. Springer, 272–286.

[10] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning Interactions and Relationships between Movie Characters. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9849–9858.

[11] Ali Rahimi, Louis-Philippe Morency, and Trevor Darrell. 2008. Reducing drift in differential tracking. *Computer Vision and Image Understanding* 109, 2 (2008), 97–111.

[12] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10146–10155.

[13] Mrigank Rochan and Yang Wang. 2019. Video Summarization by Learning from Unpaired Data. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7894–7903.

[14] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM international conference on Multimedia*. 1300–1308.

[15] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. 2011. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM International Conference on Multimedia*. 423–432.

[16] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A domain based approach to social relation recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3481–3490.

[17] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*. 4489–4497.

[18] Jiangyue Xia, Anyi Rao, Qingqiu Huang, Linning Xu, Jiangtao Wen, and Dahua Lin. 2020. Online multi-modal person search in videos. In *European Conference on Computer Vision*. Springer, 174–190.

[19] Fan Yu, DanDan Wang, Beibei Zhang, and Tongwei Ren. 2020. Deep Relationship Analysis in Video with Multimodal Feature Fusion. In *ACM International Conference on Multimedia*. 4640–4644.

[20] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* (2016), 1499–1503.

[21] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *European Conference on Computer Vision*. 474–490.