

Recognizing Social Relationships in Long Videos via Multimodal Character Interaction

Abstract:

Social relationships between characters in multiple clips of a long video may necessitate multi-step reasoning. However, most existing long video understanding approaches fail to capture the relational dependencies between characters in different clips. To solve this problem, we propose a Multimodal and Multi-granularity Relation Recognition (MMRR) framework to extract social relationships from long videos. First, we design a novel Multimodal Heterogeneous Graph (MHG) that learns the relational interactions between characters by propagating information between multimodal character nodes and multimodal clip nodes. Second, to better incorporate the contextual information from multiple character feature representations, we build a Multi-granularity Character Representation Module (MCRM) that learns global character representations over the entire long video, as well as person-specific local character representations via attention mechanism. Experimental results on two real-world datasets demonstrate the superiority of our method.

Published in: [IEEE Signal Processing Letters](#) (Volume: 30)

Page(s): 573 - 577

Date of Publication: 11 May 2023

ISSN Information:

DOI: [10.1109/LSP.2023.3275429](#)

Publisher: IEEE

Funding Agency:

SECTION I.

Introduction

Automatic recognition of social relationships from videos plays an important role in social network construction and video content understanding [1]. Existing studies mainly focus on discovering social relationships in short videos (e.g. a few seconds) [2]. However, many social relationships are expressed through long videos that last a few minutes or longer. Compared to short video-based social relation recognition, long video-based scenarios demand the application of long-term semantic knowledge and more complex reasoning abilities [3].

Recognizing social relations from long videos faces unique challenges [4], [5]. First, in a long video, the social relationship between two characters appearing in different video clips may require multi-step relational reasoning [6]. As shown in Fig. 1, assume we want to extract the social relation between Katherine in the first clip and Al in the second clip. We discover the social relation between Katherine and Sam is colleague from the first clip, and the social relation between Al and Sam are leader-sub from second video clip. This chain of interactions helps to infer the social relation leader-sub between Katherine and Al across the video clips. Most existing short video-based social relationship recognition methods use video clips from the same scene, allowing them to directly exploit interactions between co-occurring characters or exploit interactions between characters and the same scene to identify social relationships. However, in long videos, two characters with social relationships may appear in different scenes. Second, to identify social relationships between characters, the model must be able to synthesize semantic information from multiple character feature representations (i.e., face, body, speech) [7], [8], [9]. Existing methods mainly use the global character representation which aggregates all feature representations of a character indiscriminately. However, each feature representation of a character plays a different role in predicting his social relationship with others in different context. Therefore, it is necessary to learn a local character representation that considers the significance of each feature representation for a certain person.

Fig. 1.

Example of recognizing social relationships in long videos through multi-step reasoning.

Show All

To solve the above challenges, we propose a Multimodal and Multi-granularity Relation Recognition (MMRR) framework to identify social relations from long videos. First, we construct a Multimodal Heterogeneous Graph (MHG) to simulate the relational interactions of characters between video clips. Specifically, we model the information propagation process between multimodal character nodes and multimodal clip nodes. Second, we build a Multi-granularity Character Representation Module (MCRM) to learn global and person-specific local character representations from multimodal character feature representations.

In summary, our contributions are three-fold: (1) We propose MMRR framework to extract social relationships from long videos. (2) We build an MHG to model the relational interactions between characters in multiple clips. (3) We design a MCRM to learn global and local character representations.

SECTION II.

Methods

The overall architecture of the MMRR framework is shown in Fig. 2. The framework takes one long video and its audio as input and outputs the probability of the existence of each social relation between each character pair. First, MMRR uses Convolutional Neural Networks (CNN) to extract visual and audio character and scene features. Second, MMRR performs character interaction with the MHG. This module initializes the node feature from multimodal character and scene features, and iteratively propagates and updates node information by a stacked R-GCN (Relational Graph Convolutional Network) [10]. In addition, MMRR employs MCRM to learn global and local character representations via a self-attention layer and a multi-head attention layer, respectively [11], [12], [13]. At last, we concatenate the global and local character representations and feed them into the classifier to predict social relationships. Next, we present the details of the MMRR framework.

Fig. 2.

Overall architecture of the MMRR framework. In MHG, round and square nodes represent character and shot nodes, respectively. The red, yellow, and purple edges in MHG indicate character-character edges, character-shot edges and shot-shot edges, respectively.

Show All

A. Multimodal Heterogeneous Graph

Let $M=[s_1, s_2, \dots, s_K]$ be an input long video, where s_i is the i th video shot. We denote the set of all characters in M as $P=[p_1, p_2, \dots, p_N]$. Since a video is composed of both visual and audio modalities, each shot s_i has two modal inputs: visual input s_{vi} and audio input s_{ai} . Similarly, each person contains inputs in both visual and audio modalities. We use p_{fj} , p_{bj} and p_{sj} to represent face, body and speech of character p_j , respectively. Next, we use pre-trained CNN to extract face, body, visual scene feature

representations hf_j, hb_j, fvi from face, body, visual shot inputs pf_j, pb_j, svi , and extract speech, audio scene features hs_j, fai from speech, audio shot inputs ps_j, sai .

In order to capture character interactions across video clips, we build a multimodal heterogeneous graph for each long video. The graph has two types of nodes: character nodes (face nodes npf , body nodes npb , and speech nodes nps) and shot nodes (visual shot nodes nsv and audio shot nodes nsa). The three character nodes npf , npb , and nps are initialized by character face, body and speech features hf , hb , and hs , respectively. The two shot nodes nsv and nsa are separately initialized by visual and audio scene features fv and fa . In addition, we use the following edges to connect nodes, as shown in different colors in Fig. 2.

1) Character-Character Edge

As coreference is an important indicator of local and non-local dependencies [14], we connect the co-referring visual and audio character nodes in the graph. Furthermore, the co-occurrence often reveals the existence of social relationships between two characters, we thus connect two character nodes if they appear in the same shot.

2) Character-Shot Edge


Inspired by the syntactic structure dependencies of words in documents [15], we connect character nodes to shot nodes if characters appear in the shots, which helps to reveal the relationship recognition within the shots.

3) Shot-Shot Edge

We connect all shot nodes to model the non-sequential dependencies.

Since R-GCN is specifically designed to handle multi-relational data [10], [16], we compute the node representation through a stacked R-GCN. At each layer $l+1$, each node $n_{l+1,i}$ is updated as:


$$n_{l+1,i} = \sigma \left(\sum_{y \in Y} \sum_{j \in N_{y,i}} \left(\frac{1}{|N_{y,i}|} \right) W_{ly} n_{l,j} + W_{l0} n_{l,i} \right) \quad (1)$$

View Source  where σ is the activation function, $N_{y,i}$ denotes the set of neighbor indices of node i under edge type $y \in Y$, and W_{ly} and W_{l0} are weight matrices.

B. Multi-granularity Character Representation Module


Self-attention can calculate a sequence representation by relating distinct positions of a sequence. Its computation requires a query Q , key K and value V from the same source:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{d_k} \right) V \quad (2)$$

[View Source](#) 

Since the global character representation reflects the semantics of a person throughout the long video, we use self-attention mechanism to aggregate all visual and audio character feature representations of a person to learn the global character representation:

$$\text{eglo} = \text{LN}(\text{Attention}(\{\text{np}\}, \{\text{np}\}, \{\text{np}\}))(3)$$

[View Source](#)  where $\text{LN}(\cdot)$ denotes layer normalization operation, and $p \in \{\text{pb}, \text{pf}, \text{ps}\}$.

Multi-head attention allows the model to jointly attend to information from different representation subspaces. Multi-head attention can be formulated as:


$$\text{MultiHead}(Q, K, V) \text{head}_i = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_o = \text{softmax}(QW_{Qi}KW_{Ki}d_k - \sqrt{d_k})VW_{Vi}(4)(5)$$

[View Source](#) 

where $W_{Qi} \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_{Ki} \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_{Vi} \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W_o \in \mathbb{R}^{h d_k \times d_{\text{model}}}$.

For character $p_x \in P$, it targets the local representation of character $p_z \in P$ as:


$$\text{eloc}_{x \rightarrow z} = \text{LN}(\text{MultiHead}(\text{eglo}_z, \{\text{ns}_x\}, \{\text{np}_x\}))(6)$$

[View Source](#)  where eglo_z denotes the global representation of p_z . np_x indicates the face, body and speech feature representations of character p_x extracted from MHG. The audio or visual feature representations extracted from MHG of the shot in which p_x occurs is represented by ns_x . In general, if a shot containing p_x has more semantic similarity to eglo_z , the feature representations of p_x in this shot should contribute more to $\text{eloc}_{x \rightarrow z}$.

C. Social Relation Prediction

We concatenate the global and local character representations to predict the social relationships between two characters. Assume that the set of all social relations is R , the probability of the existence of the social relation $r \in R$ between characters p_x and p_z is:

$$o_r = \text{sigmoid}(\text{FNN}([\text{eglo}_x; \text{eloc}_{x \rightarrow z}; \text{eglo}_z; \text{eloc}_{z \rightarrow x}]))(7)$$


[View Source](#)  where FNN is a feedforward neural network,

and eglo_x , $\text{eloc}_{x \rightarrow z}$, eglo_z , $\text{eloc}_{z \rightarrow x}$ are the global and local representations of

characters p_x and p_z , respectively.

We employ the cross-entropy loss to train our model:

$$L = - \sum_{r \in R} (o_r \log(o_r) + (1 - o_r) \log(1 - o_r)) \quad (8)$$

View Source  where $o_r \in \{0, 1\}$ denotes the true label of o_r .

SECTION III.

Experiments

A. Implementation Details

Existing video social relationship datasets are mainly based on short videos [17]. To facilitate the social relation recognition in long videos, we built a Long Video Relation Recognition (LVRR) dataset based on the existing VPCD (Video Person-Clustering Dataset) dataset [18]. We began by collecting over 800 videos lasting more than one minute from six movies and TV shows in the VPCD dataset, each with at least two socially related characters. Then, we choose labels from eight types of social relationships to annotate each video. Given the duration of each movie and TV series in the VPCD dataset, we divide the dataset using one from The Big Bang Theory, About Last Night, and Hidden Figures in turn as the validation set, one from Buffy, Sherlock as the test set, and the rest as the training set. In the end, we average the results of the six experiments as the final result. The LVU dataset contains $\sim 30K$ 1-2 minute long videos from $\sim 3K$ movies to cover nine long video understanding tasks. We exclusively use the LVU dataset for the social relationship identification task which contains three social relationship categories. We divide LVU into 70% for training, 15% for validation, and 15% for testing.

We use PySceneDetect to split each long video into shots. For the LVRR dataset, we use the face, body and speech features given in the VPCD dataset. We uniformly partition an input shot into 20 segments, and then randomly sample one frame for each segment to obtain 20 frames for one shot. We feed the frames into the pretrained ResNet50 [19], [20], and then aggregate frame-level features into a shot-level feature by average pooling. For audio shot features, we use a fixed length 2-second temporal segment, extracted randomly from each shot. Spectrograms are extracted with a hamming window of width 25ms and step 10ms. We input the raw spectrograms to the pretrained thin-ResNet-34 [21], [22], followed by average pooling to aggregate the features into the shot level. For the LVU dataset, we use a Faster R-CNN-based person detector pre-trained on the AVA dataset to detect characters [23], [24]. If a person appears on the video screen, the audio for that time period is used to provide speech

for that person. We use Multi-MuHPC method for multimodal person clustering in each long video [18], [25]. For each long video, we select at most 20 character features at random. We apply the same pretrained network that we use on the LVRR dataset to extract the character and shot features from the LVU dataset. The number of R-GCN layers is set to 2, and the node hidden state size is 256. The multi-head attention layer in MHG has 2 attention heads. We train our model by Adam optimization with a learning rate of 10^{-4} [26]. The model is trained for 50 epochs with a batch size of 16.

B. Comparison With State-of-the-Art Methods

We present top-1 accuracy of all kinds of social relationships or corresponding overall accuracy. Table I shows the recognition results of MMRR and three long video understanding methods in the LVU dataset [27], [28]. As can be seen, MMRR achieves the highest accuracy. This is due to the fact that MMRR can better learn character representations to predict social relationships between characters in different video clips.

TABLE I Comparison With State-of-the-Art Methods on LVU

Table II shows the results on the LVRR dataset. Two image-based social relation recognition methods GRM and GR 2 N obtain poor results [29], [30]. The reason is that image-based methods need to detect character co-occurrence in images, however in videos, characters with social relations often appear in different frames. The short video social relationship recognition method, MSTR, achieves improvements because MSTR makes better use of the dynamic nature of video than the image-based approaches [31]. However, because MSTR only uses visual information, it cannot capture multimodal cues for social relation prediction. The overall accuracy of MMRR is 2.9%, 2.5%, and 2.7% greater than the three long video understanding methods obj-Trans, ViS4mer, and Long-Trans, respectively, since MMRR can better capture the interaction between characters throughout video clips. In addition, we find that MMRR achieves the best results on friend, service, and sibling. It could be that while the understanding of working relationships, such as colleague, leader-sub, pays more attention on reasoning in conjunction with the scenes, our approach places more emphasis on the interactions between characters.

TABLE II Comparison With State-of-the-Art Methods on LVRR Dataset

C. Ablation Study

We first evaluate the impact of the multimodal information. According to Table III, we find that MMRR w/o visual achieves a low accuracy, the reason may be that MMRR w/o visual merely uses the audio features to represent the tone, timbre or emotion of the characters, which may cause the model make ambiguous judgments about the social relations between the characters. The overall accuracy of MMRR w/o audio is higher than that of MMRR w/o visual, implying that visual cues can provide more discriminative information than audio cues [32]. Additionally, we notice that these three models have various effects on different types of social relationships. For example, the MMRR framework results in significant improvements in leader-sub and service. This highlights the significance of combining multimodal information for social relationship recognition in long videos [33], [34].

TABLE III The Influence of Multimodal Information and Each Module on the LVRR Dataset

Next, we validate the influence of each module in MMRR. As shown in Table III, we can find that our accuracy is improved by adding MHG and MCRM. This validates the importance of modeling character relational interactions across video clips and learning multi-granularity character representations for social relationship recognition in long videos.

D. Quality results

We present two case studies to discuss how MMRR framework recognizes social relations. Fig. 3 shows the correct social relations generated through MMRR. In this case, each person appears in a separate video clip, and there is no interaction between John and Mrs. Hudson. Our model can deduce from the social relationships of John and Sherlock, and Sherlock and Mrs. Hudson that the social relationship between John and Mrs. Hudson is service. Fig. 4 shows a wrong case in which MMRR recognizes the social relationship between John and others as friend but labels them as service. Although we can identify the relationship between John and others as service by reading the character dialogues, MMRR is unable to process the textual information. This

phenomenon reminds us that the use of more modal information may be more beneficial for understanding social relationships in long videos.

Fig. 3.

Correct case for social relation recognition in long videos. The blue, green and orange bounding boxes indicate John, Sherlock, and Mrs. Hudson, respectively.

Show All

Fig. 4.

Wrong case for social relation recognition in long videos, where the blue bounding box indicates John.

Show All

SECTION IV.

Conclusion

This paper presents MMRR framework for long video social relationship recognition. MMRR uses MHG to represents a long video as a graph to address the relational interaction between characters across video clips, and captures global and local character representations through MCRM. Extensive experiments on two benchmarks demonstrate that MMRR achieves state-of-the-art performances.