Linking the Characters: Video-oriented Social Graph Generation via Hierarchical-cumulative GCN

ShiweiWu[1], Joya Chen[2], Tong Xu[1,2]*, Liyi Chen[1], Lingfei Wu[3], Yao Hu[4], Enhong Chen[1,2]
[1]School of Data Science, University of Science and Technology of China,
[2]School of Computer Science and Technology, University of Science and Technology of China,
[3]JD.COM Silicon Valley Research Center,
[4]Alibaba Youku Cognitive and Intelligent Lab
{dwustc,chenjoya}@mail.ustc.edu.cn,{tongxu,cheneh}@ustc.edu.cn,liyichencly@gmail.com,
lwu@email.wm.edu,yaoohu@alibaba-inc.com

ABSTRACT

Recent years have witnessed the booming of online video platforms. Along this line, a graph to illustrate social relation among characters has been long expected to not only benefit the audiences for better understanding the story, but also support the fine-grained video analysis task in a semantic way. Unfortunately, though we humans could easily infer the social relations among characters, it is still an extremely challenging task for intelligent systems to automatically capture the social relation by absorbing multi-modal cues. Besides, they fail to describe the relations among multiple characters in a graph-generation perspective. To that end, inspired by the human inference ability on social relationship, we propose a novel **H**ierarchical-**C**umulative **G**raph **C**onvolutional **N**etwork (**HC-GCN**) to generate the social relation graph for multiple characters in the video. Specifically, we first integrate the short-term multi-modal cues, including visual, textual and audio information, to generate the frame-level graphs for part of characters via multi-modal graph convolution technique. While dealing with the video-level aggregation task, we design an end-to-end framework to aggregate all frame-level subgraphs along the temporal trajectory, which results in a global video-level social graph with various social relationships among multiple characters. Extensive validations on two real-world large-scale datasets demonstrate the effectiveness of our proposed method compared with SOTA baselines.

CCS CONCEPTS

• **Information systems** → **Multimedia streaming**; • **Computing methodologies** → *Activity recognition and understanding*.

KEYWORDS

Social relationship, Video understanding, Multimodal analysis, Graph convolutional network

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475684

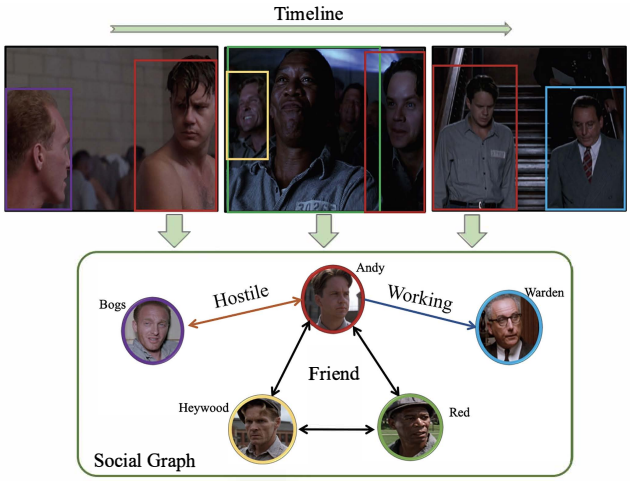

Figure 1: An Example of Social Graph in Video

ACM Reference Format:
Shiwei Wu, Joya Chen, Tong Xu, Liyi Chen, Lingfei Wu, Yao Hu, Enhong Chen. 2021. Linking the Characters: Video-oriented Social Graph Generation via Hierarchical-cumulative GCN. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475684

1 INTRODUCTION

With the prosperity of online social media platforms, large audiences have been attracted to view abundant video content with affiliated intelligent services like retrieval, recommendation and summarization tasks. In this case, the *social relation* among characters has long been treated as a crucial factor to support semantic-related services [34]. On the one hand, audiences will be guided when enjoying masterpieces like "*Game of Thrones*" and "*Harry Potters*" with dozens of characters, which leads to a better experience. On the other hand, thanks to the social information as prior knowledge, some fine-grained semantic analysis is now available, e.g., a video clip which indicates dining of lovers could be probably labeled as a "dating" rather than an ordinary "gathering". Therefore, it is valuable and necessary to capture the social relation among characters for better video understanding.

Unfortunately, though we humans could easily infer the social relations among characters via some comprehensive cues, e.g., their interactions, conversations, clothing styles and the background, however, it is still an extremely challenging task for intelligent systems to automatically capture the social relation by absorbing multi-modal cues, which severely limit the application of social relation factors. To that end, large efforts have been made on this issue based on various technical tools. But, in general most of them still mainly focus on the visual cues, e.g., spatial relations or interactions [5, 6, 18, 23], while few of them successfully integrate textual or audio cues to enhance the performance. Therefore, the multi-modal techniques for this task are still urgently required.

Moreover, current works mainly focus on only general or pairwise relationships, which impair the benefit of social relation factors for video understanding. For instance, some works target at labeling the general relation on the whole video clips [18, 19], i.e., they treat the pairwise relationship in one clip as the same. In this way, the application could be extremely limited as they could not fit the complicated relations in most videos. At the same time, some other prior arts attempt to capture the relations among character pairs [13, 34], i.e., they treat all the character pairs as a mutual independent. Though these attempts may distinguish different relations between characters, however, they may fail to achieve competitive performance as rich side information from other pairs has been ignored. Let's take the social graph in Figure 1 as an example. In the classic movie "*The Shawshank Redemption*", we all know that Andy and Red are the best friends according to the story. Therefore, if we could reveal the friendship between Red and Heywood, our judgment on the friendship between Andy and Heywood could be enhanced, as they form the "triadic closure" of friendship in a social network. Similarly, considering that Bogs and Andy are mutually hostile, and we realize that the Warden punished Bogs due to Andy's encounter, it is probably that Andy is beneficial to the Warden, e.g., working for him. In summary, it is valuable to model the social factors in the perspective of a social graph to achieve a more accurate summarization of social relations among characters.

Inspired by these observations, in this paper, we propose a novel **H**ierarchical-**C**umulative **G**raph **C**onvolutional **N**etwork (**HC-GCN**) to generate the social relation graph for multiple characters in the video. Specifically, we first integrate the short-term multi-modal cues, including visual and textual information, to generate the frame-level subgraphs for part of characters via multimodal graph convolution technique. In this step, a subgraph will be generated for each frame to describe its social state. Along this line, to deal with the clip-level aggregation task, we design an end-to-end framework to aggregate all frame-level subgraphs along the temporal trajectory. Specifically, we update characters and character-pairs representation to portrait the variations using LSTMs, and further fuse the global multimodal features as supplementary information to provides global guidance. Moreover, we enhance the representation of the character-pairs by passing messages using a graph convolutional network (GCN), and then put each character-pairs into the classifier in a weakly supervised way. In this way, character pairs even without co-occurrence frames could be captured via message passing in a global social graph. We summarize our technical contribution as follows:

- We formulate the social relation recognition task for video characters in a novel perspective of social graph generation, to achieve more accurate social relation summarization.
- We propose a novel Hierarchical-Cumulative GCN structure to generate the social graph for characters, in which the multi-modal cues have been comprehensively integrated.
- Extensive validations on two real-world large-scale datasets demonstrate the effectiveness of our proposed technical framework compared with SOTA baselines.

2 RELATED WORK

In general, the related works could be roughly grouped into three categories, namely *social relation recognition for visual content*, *scene graph generation* and *GCN application in computer vision (CV) field*.

2.1 Social Relation Recognition in CV

Among all the prior arts, the most related part is the task of social relation recognition in visual content, e.g., images, videos and so on. Indeed, in recent years, this task has attracted wide attention in both academic and industry [15, 23, 27, 38]. Usually, most of the existing studies mainly focus on the recognition task on still images [5, 6, 15, 22, 23]. Along this line, two large-scale datasets, namely The People in Photo Albums (PIPA) [37] and the People in Social Context (PISC) [15] are published for this task. To deal with this task, large efforts have been made, e.g., Li et al. [15] proposed a dual-glance model for social relationship recognition, where the first glance focused on persons of interest and the second glance applied attention mechanism to discover contextual cues. Also, Sun et al. [23] adopted a CNN to recognize social relations from a group of semantic attributes, and Wang et al. [27] proposed to represent the persons and objects in an image as a graph, with social relation reasoning by a Gated Graph Neural Network. However, as still images could not provide temporal dynamics between characters, this task could be significantly different with our problem.

At the same time, some other researchers attempted to capture social relation in videos [13, 18, 19, 34]. Traditionally, most studies [18, 19] simply considered the social relation recognition as a classification task, e.g., Lv et al. [19] built the first video dataset for social relation recognition named Social Relation In Video (SRIV), which contained about 3,000 video clips with multi-label annotation, and then exploited the Temporal Segment Networks [25] to classify a video using the RGB frames, optical flows, and audio of the video. Along this line, Liu et al. [18] proposed a large-scale and high-quality Video dataset called as ViSR, and proposed a graph network to capture long-term and short-term temporal cues in the video. Besides, Kukleva et al. [13] proposed neural models to jointly predict interactions, social relations and the pair of characters that are involved with visual and dialog cues, and Xu et al. [34] recognized the social relation between character-pairs with integrating visual-textual cues. However, none of them formulate the social recognition task in a social graph generation task with utilizing the hierarchical GCN technique, which is different from our solution.

2.2 Scene Graph Generation

Another related topic is the generation of the scene graph, which is widely studied in the CV field to describe the spatial or structural

relationships between objects. Indeed, the idea of using graph-based context to improve scene understanding has been investigated by numerous studies in the last decades [2, 14, 17, 31, 33]. For instance, Johnson et al. [10] firstly introduced the problem of modeling the objects and their relationships using scene graphs, which aims to simultaneously detect objects and their pairwise relationships. Afterward, Zellers et al. [36] proposed to capture higher-order repeated structures of scene graph for better performances. Also, Yang et al. [35] developed an attention-aware GCN framework to update node and relationship representations by propagating context between nodes in candidate scene graphs, and Xu et al. [32] used RNNs to jointly refine the object and the relationship features in an iterative way to construct the scene graph. Inspired by the structural representation of the scene graph, we formulate the social relation recognition task in the video from the perspective of generating the social graph, which could further enhance the basic scene graph with rich semantic information for more applications.

2.3 GCN Application in CV

Finally, we will summarize the application of the GCN technique in the CV field. The GCN tool was first proposed in [12] in the context of semi-supervised learning. In detail, GCN is designed to decompose complicated computation over graph data into a series of localized operations (typically only involving neighboring nodes) for each node at each time step, which was used in various fields [3, 4, 29, 30]. Most recently, GCN has been adopted to computer vision tasks, e.g., Wang and Gupta [26] proposed to represent a video as a space-time region graph by the persons and objects in videos, and adopted a GCN to learn video level features for action recognition. Also, Liu et al. [18] proposed to represent the actions and interactions of persons and objects in videos as graphs, with reasoning the objects by pyramid GCN for social relation recognition. Inspired by the above studies, we propose to generate the social graph via performing GCN from frame graph to video graph in a hierarchical way.

3 PROBLEM STATEMENT

Table 1: Mathematical Notations

Notation	Description
M	Videos
V	Video clips
SG	Social graph
G_t	Global textual feature
G_v	Global video feature
G_a	Global audio feature
F_c	Frame character feature
F_p	Frame character-pair feature
F_t	Frame textual feature
F_b	Frame background feature
C_c	Cumulative character feature
C_p	Cumulative character-pair feature
R	The set of social relations between characters

In this section, we will introduce the preliminaries of our social graph generation task for video characters, and then formally define the task with mathematical formulation. For facilitating illustration, we summarize related mathematical notations in Table 1.

In this paper, given the input raw video set M, we target at generating a social graph SG from each video. Specifically, to integrate the multi-modal cues, we extract textual and background audio cues along with the videos, and then summarize the global video feature G_v, the global text feature G_t and the global audio feature G_a to enhance the modeling.

Correspondingly, we have a pre-defined social relation set R, and each character pair $<c_i, c_j>$ is labeled as a relation $r_{ij} \in R$. Finally, we formulate the social graph generation task for videos M, which is formally defined as follow:

Definition 1. *Given the video set M along with textual and background audio information, as well as the pre-defined social relation label set R, we target at generating the social graph SG for target characters in video.*

4 METHODOLOGY

In this section, we first present the overview of our proposed technical framework towards generating the social graph for video characters. Then, we introduce the technical details of modules in our framework step-by-step.

4.1 Framework Overview

To deal with the aforementioned problems, we propose a novel GCN-based framework containing two main approaches, i.e. Hierarchical-cumulative GCN, with the weakly supervised training method, which is illustrated in Figure 2. The functions of these two modules are briefly introduced as follows.

Hierarchical-cumulative GCN Module. In this module, we aim to get the comprehensive and cumulative representation of all the character pairs. First, to enhance the representation of characters and character-pairs in each frame, we build a *frame-level graph* to aggregate the multi-view and multi-modal information through GCN framework, which indicates the social relation among characters in the current frame. Along this line, we design the *Multi-way Temporal Cumulation* tool to update and aggregate the representation of characters and character-pairs along the temporal dimension, respectively. Moreover, We treat the output of the last step in each LSTM as the input node of *clip-level graph* for a video clip, and then fuse the global multi-modal features into these nodes. Afterward, GCN is adopted to propagate messages in different modal among these node features.

Weakly Supervised Training and Inference. As the fine-grained annotating in the video are nearly impossible, we train the model in a weakly supervised way. Specifically, we perform a cross softmax operation to generate the confidence score matrix, thus the contribution of each character pair to each type of social relationship could be evaluated, and then accumulated by cross-entropy criterion to compute the loss. Finally, we merge all the clip-level subgraphs to generate the final social graph for the whole video.

4.2 Data Pre-processing

Then, we turn to introduce the details for data pre-processing. As it is extremely difficult to directly analyze the untrimmed long

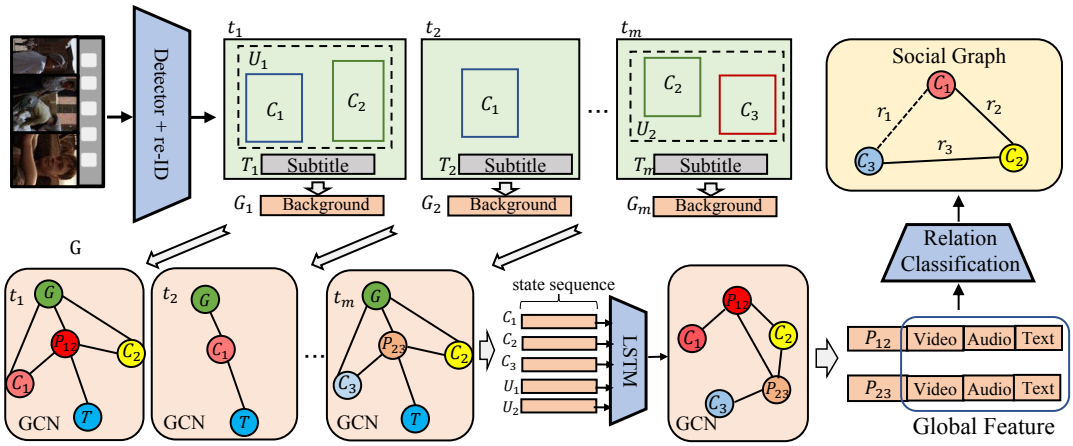

Figure 2: The Overall Framework of Hierarchical-Cumulative GCN Framework

videos, we split them into short clips by pre-defined sliding windows. Therefore, all videos are in the same length and it could be trained in batches.

At the same time, we localize and re-identify each character in an unsupervised way to provide detailed annotations. In detail, we first adopt Faster R-CNN detector [21] to indiscriminately locate each character in a video clip v, which could produce character person boxes frame by frame due to its strong capability of detecting varying sized objects in unconstrained scenes. Afterwards, we adopt the ResNet-50 [7] backbone network, which is pre-trained on the CSM movie dataset[1] via PPCC method [9], to extract character features F_c on each frame. Since the character feature provides a discriminative visual representation of a certain character, the distance between two different character features will be larger than the same character features. Along this way, we can re-identify the same character in v by comparing the similarity distance among all characters' features F_c occurred in the video clip.

4.3 Hierarchical-cumulative GCN Module

We now turn to explain the technical details for the Hierarchical-cumulative GCN module. Specifically, given a video clip v, we utilize three components, namely *Frame-level GCN, Multi-way Temporal Cumulation* and *Clip-level GCN* in HC-GCN to generate a social graph SG_v, then we merge all the clip-level subgraphs into the global social graph SG_m for each whole video $m \in M$.

4.3.1 Frame-level GCN. In this module, we aim to generate a frame-level subgraph to provide a structural representation of the social state on the current frame. In this way, we can enrich the representation of characters and character-pairs with multi-view and multi-modal information through GCN on each frame.

In detail, as all the characters in the video are localized and re-identified, we use the visual feature of the union box over the proposal boxes as the representation of the character-pairs. We adopt the ResNet-50 backbone to extract the union box feature F_p. It is worth noting that this visual feature extractor targets at capturing the interaction between the character pairs from visual perspective. Also, we use the ResNet-50 person backbone to extract character representation as previously mentioned [9]. To get the global background feature F_b, we put the current frame image into the ResNet-50 place backbone pre-trained on the Places365 dataset[2], which contains rich scene information. Moreover, we use the Sentence Transformers [20] to extract textual features F_t from the textual information aligned with the current frame. What should be noted is that the character-pairs are processed as a kind of special nodes rather than edges for convenience. Therefore, four different nodes are extracted into the frame graph, including the text cues F_t and the multi-view information, i.e., characters F_c, character-pairs F_p and global context F_b.

To combine the comprehensive information of the frame-level graph, we use GCN to mutually propagate node information. Different from traditional Convolutional Neural Networks (CNN) which usually apply 2-D or 3-D filters on images or videos to abstract visual features from low-level space to high-level space [7], GCN performs message propagation from nodes to its neighbors in the graph. Therefore, we can apply GCN on the frame-level graph to enhance the representation of the characters and character-pairs.

To be specific, as in [12], given a graph with N nodes in which each node has a d-length feature vector, the operation of one graph convolution layer can be formulated as:

$$X^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}) \tag{1}$$

where $\tilde{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix of the graph, $\tilde{D} \in \mathbb{R}^{N\times N}$ is the degree matrix of $\tilde{A}$, $X^l \in \mathbb{R}^{N\times d}$ is the output of the $l-1$-th layer, $W^l \in \mathbb{R}^{d\times d'}$ is the learned parameters, and $\sigma(\cdot)$ is a non-linear activation function like ReLU. Particularly, in our frame-level graph, the adjacency matrix is defined as below, where F_c^i and F_p^{ij} denotes the feature for i-th character and character-pair between i and j, respectively.

$$A(N_1,N_2) = \begin{cases} 1 & (N_1,N_2) \in \{(F_c^i,F_p^{ij}),(F_c^j,F_p^{ij}),(F_p^i,F_p^j) \\ & \quad (F_c,F_b),(F_p,F_b),(F_p,F_t)\}, \\ 0 & otherwise, \end{cases} \tag{2}$$

[1] http://qqhuang.cn/projects/eccv18-person-search/

[2] http://places2.csail.mit.edu/download.html

Finally, we can aggregate the multi-view visual features F_c, F_p, F_b and the textual features F_t to obtain the enhanced representation of characters and character-pairs for current frame via GCN.

$$F_c, F_p = GCN(F_c, F_p, F_b, F_t, A) \tag{3}$$

4.3.2 Multi-way Temporal Cumulation. After obtaining the enriched representation of the characters F_c and character-pairs F_p on each frame, we now focus on modeling in the temporal domain to capture the feature variation along the timeline. Considering that the LSTM [8], as a variant of the recurrent neural network (RNN) but without the shortages like gradient disappearance and gradient explosion, is widely used for processing sequence data. Therefore, we choose the LSTM network to update the representation of each node along the timeline. Specifically, LSTM obtains the representation status at time t by absorbing the hidden status at time $t-1$ and the current representation of characters or character-pairs as inputs. While capturing the representation for the current frame, the model still retains the changing trend of the previous representation and can capture temporal dependence.

Therefore, for each character feature F_c^i and character-pair feature F_p^{ij}, we use LSTM to capture temporal dynamics of them on all frames along the timeline. Specifically, we adopt two independent LSTM to propagate the series of features due to the difference in representation between character feature and character-pair feature. In detail, the LSTM for character aims to depict the variation of the person, such as the changing clothes and posture of the certain character. Meanwhile, the LSTM for character-pair targets at describing the variation of interaction between two people. As we put the character and character-pair feature into the multi-way LSTM individually, we obtain the cumulative feature C_c and C_p from all frames in the video. We take the output of the last step in each LSTM as the input node of the clip-level graph.

4.3.3 Clip-level GCN. Further, to depict all the characters in clips V, and generate a clip-level social graph SG_v, we merge all the subframe graphs through *Multi-way Temporal Cumulation* module, with taking the cumulative character feature C_c and the cumulative character-pair feature C_p as the input of the clip-level social graph SG_v. It is worth noting that the cumulative representation of the character C_c and the character-pairs C_p in the social graph SG_v not only contains multi-view and multi-modal information due to the frame-level GCN, but also captures the variations on temporal dimension thanks to the *Multi-way Temporal Cumulation* module.

$$C_p = GCN(C_c, C_p, A), \tag{4}$$

where the adjacency matrix is defined as

$$A(N_1, N_2) = \begin{cases} 1 & (N_1, N_2) \in \{(C_c^i, C_p^{ij}), (C_c^j, C_p^{ij}), (C_p^i, C_p^j)\} \\ 0 & otherwise, \end{cases} \tag{5}$$

Same as prior work [18], to enrich the cumulative representation of character-pairs C_p with global information, C_p is concatenated with the global video feature G_v, global textual feature G_t and global audio feature G_a to form a joint representation. Then the joint representation is fed into the relation classification module.

4.4 Weakly Supervised Training and Inference

Finally, we turn to introduce the details for training and inference. For a video, it is difficult to obtain the character-level annotations on each frame. In this case, we should only leverage the clip-level social relationship annotations to predict the character-pair social relationship. To that end, we propose a weakly supervised loss function to address the challenging task. Suppose that there are R kinds of social relationships as well as P pairs of characters, then we add an R-way classifier to generate a predicted logits matrix Q of $R \times P$ shape. We want to adaptively learn which pairs are corresponding to the clip-level social relationship labels. Inspired by the weakly supervised object detection [1], we perform a cross softmax operation to Q to generate the confidence score matrix:

$$\mathrm{S} = softmax_r(Q) \odot softmax_p(Q), \tag{6}$$

where $softmax_r$ and $softmax_p$ denote the softmax operation over all character pairs and all kinds of social relationships, respectively. We believe that Equation 6 could evaluate the contribution of each character pair to each type of social relationship. Then, we can accumulate the confidence score of each character pair, with cross-entropy criterion to compute the loss:

$$L = -\sum_{r}^{R} \log |y_r - \sum_{p}^{P} s_r^p|, \tag{7}$$

where $y_r = 1$ denotes the video has the social relationship of type r, otherwise $y_r = 0$. $\sum_{i,j} s_r^{i,j}$ means that we accumulate all character pair normalized scores on the relationship of type r, which can represent the clip-level confidence score in that type. By this loss function, we can train the network with only clip-level ground-truth labels, and the related character-pairs will gradually be highlighted by cross softmax activation during training.

As illustrated above, we apply the weakly supervised loss strategy for training the clip-level social graph generation. During the inference process, to obtain the global social graph SG_m, we merge all clip-level social subgraphs $\{SG_{v_1}, SG_{v_2}, ..., SG_{v_n}\}$ from the same video to generate SG_m. Our merging strategy is based on the character feature similarity. For two characters from the different clip-level subgraph, if their feature cosine similarity is larger than 0.7, which is pre-set according to [9], we consider these two characters as the same. After this procedure, we can merge relations in two subgraphs as a larger social graph. By iteratively merging every two graphs, we can finally obtain the social graph for the whole video.

5 EXPERIMENTS

In this section, we will introduce the details for extensive experiments on two real-world datasets to validate our framework. Specifically, we first describe both datasets, and then introduce the overall comparison as well as the ablation study results. After the validation part, we also have discussions on the challenges of the task and further reveal some interesting rules of multimodal cues.

5.1 Datasets

We conduct extensive experiments on two large video-based social relation datasets, namely the ViSR dataset [18], as well as our self-constructed dataset. We will publish this dataset after acceptance.

Table 2: The Categorization Scheme of Social Relations in Two Datasets

Relation on Bilibili	Relation on ViSR	Examples
Working	Colleague	Co-worker, schoolmate
	Leader-sub.	Teacher-student, leader-member
	Service	Passenger-driver, customer-waiter
Kinship	Parent-offs.	Parent-child, grandparent-grandchild
	Sibling	Brothers, sisters
Hostile	Opponent	Enemies
Friend	Friend	Friends in general scenes
Couple	Couple	Husband-wife, boyfriend-girlfriend

Specifically, the ViSR dataset contains more than 8,000 valid videos, which are collected from more than 200 movies. All videos are classified into 8 categories as shown in Table 2, and each video only has one label which reveal the main social relation in the video. The length of each video is limited in $10 \sim 30$ seconds. At least two persons that have interactions must exist in one clip. We follow the experiment setting in [18] to validate our method.

At the same time, our self-constructed dataset is collected on Bilibili[3], which is one of the biggest social media platforms in China. In detail, the data set contains 70 untrimmed movies with average length of 1.9 hours. Along this line, a total of 376 main characters in these movies are selected and two types of annotation are conducted, i.e., we first annotated the timestamp when the certain character appears in the movie, and then labeled the social relations among characters based on the movie content and material from *Baidu Encyclopedia*. All social relations are grouped into 5 categories as shown in Table 2 referred to [18].

Different from the relation definition in Liu et al. [18], we aggregate "Parent-offspring" and "Sibling" relations into "Kinship" relation, and aggregate "Leader-subordinate", "Colleague" and "Service" relations into "Working" relation due to the data sparsity. To deal with our social graph generation task, we first split the untrimmed movie into several video clips using PySceneDetect[4], and then picked out all the video clips which contain more than two characters in the segment. Considering the short video can hardly depict the social relation explicitly, we select video clips longer than 8 seconds. Afterward, we randomly divide our self-constructed dataset by the ratio $7:1:2$ for training, validation and testing set respectively.

5.2 Experimental Settings

Preprocessing. We split all the videos M into video clips $V_i \in [1, n]$ by sliding window with stride $\tau = 1$. Then, we sample the video clips by the frequency as 2 frames/second, and each video lasts 16 frames. For character detection, we use the Faster R-CNN detector [21] with ResNet-50 [7] backbone to detect all persons in videos. We select the person boxes with a confidence score above 0.8 and the NMS threshold at 0.4. To eliminate the interference of background characters, we remove the small person boxes whose length and width are both below 3% of the current frame. Due to the limited GPU memories, we only select no more than 5 detected person boxes with top confidence scores during training, while the limitation is removed during inference. Then, in the re-ID part, we extract the character features based on [9], then we treat these two characters are the same person if the similarity score is above the threshold $\eta = 0.7$ empirically according to [9].

Implementation Details for HC-GCN. For each frame in video clips V, we cropped the characters and character-pairs images according to the box coordinates which are pre-computed. Then, we resize the cropped character, character-pair and the current frame images into 256×128, 256×256 and 256×256, respectively. Then, character images and character-pair images were sent into the ResNet-50 person extractor to get 2048-dim character features and 2048-dim character-pair features, while the current frame images are sent into the ResNet-50 place extractor to get 2048-dim context features. Also, the Multi-way Temporal Cumulation which we used to aggregate the information are two independent 2-layer LSTM. Besides, we used Sentence Transformers [20] to extract 768-dim textual dialog features and use the 257×90 dim Short-time Fourier transform (STFT) features of the background audio as the audio feature. Besides, for the global video feature, we adopted the ResNet(2+1)D-18 network [24] pretrained on the Kinetics-400 [11] dataset to extract 512-dim features from videos.

5.3 Baseline Methods

To evaluate the performance of our HC-GCN framework, we compare it with the following state-of-the-art methods, which could be roughly grouped into the models based on still images and videos.

For our self-constructed Bilibili dataset, as we have the social relation label between each character-pairs and the timestamp when the character appears in the movie, we can generate the social graph in the weakly supervised training method. Thus, we verify our HC-GCN framework compared with several baselines as follows:

- Double-stream CaffeNet (DSC) [23], which is a still image-based method using two CaffeNet to extract body feature after *fc7* layer, and then classify the feature via a linear SVM for social relation classification.
- Multi-stream Fusion Model (MSFM) [19], which jointly models the multi-modal information such as visual, optical flow and audio features and then fuses them with logical regression to predict the label of the social relation.
- Textual-enhanced Fusion Model (TEFM) [34], which jointly embeds the visual and textual information via attention mechanism, then classify social relations for character-pairs.

For the ViSR dataset, since each video only has one main social relation label, we can only predict the main relation of the whole video via taking the average feature of all character-pairs into the classifier. Specifically, we validated the performance of our HC-GCN framework with the several state-of-the-art methods on the ViSR dataset as follows, and then use the top-1 accuracy on each social relation class, as well as the mean Average Precision (mAP) over all classes to evaluate the performance.

- Graph Reasoning Model (GRM) [28], which is a still image-based method to classify the social relation in still images via GGNN [16]. For each frame in the video, it performs GRM and

[3] http://www.bilibili.com/
[4] https://pyscenedetect.readthedocs.io/

Table 3: Comparison with the State-of-the-art Methods on ViSR Dataset

Methods	Top-1 Accuracy								
	Leader-Sub.	Colleague	Service	Parent-offs.	Sibling	Couple	Friend	Opponent	mAP
GRM [28]	48.67		6.67	0.00		4.17	0.67	30.13	16.69
TSN-ST [25]	41.05	33.33	30.00	32.83	45.78	29.17	63.76	32.87	43.23
MSTR [18]	57.53	51.09	30.00	45.60	39.33	38.71	53.23	47.41	47.75
HC-GCN	49.32	**54.21**	**35.62**	**49.60**	40.54	36.52	62.27	40.78	**48.74**

Table 4: Comparison on self-constructed Bilibili Dataset

Methods	5-category Classification		
	R(%)	P(%)	F1(%)
DSC [23]	18.4	24.0	16.8
MSFM [19]	32.2	30.2	26.4
TEFM [34]	47.7	35.8	32.5
HC-GCN	**68.9**	**58.1**	**62.2**

then aggregates the results from all frames for video-based social relation recognition.

- Temporal Segment Network using Spatial-Temporal features (TSN-ST) [19], which uses TSN [25] to learn the spatial and temporal features for social relation recognition. We follow the training strategy in [19] and change the multi-label classification task to the single-label classification task.
- Multi-scale Spatial-Temporal Reasoning (MSTR) [18], which adopts the pyramid GCN to learn multi-scale dynamics of persons from the Triple Graphs and TSN to learn the global spatial features, and then predicts the social relation in videos after the weighted fusion of the PGCN and TSN.

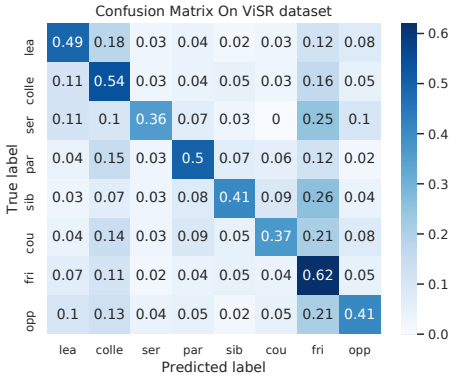

Figure 3: The Normalized Confusion Matrix of HC-GCN framework on ViSR Dataset

5.4 Experimental Results

The overall performance on two datasets are summarized in Table 3 and 4, respectively. For ViSR datasets, we follow [18] to show the top-1 accuracy of all social relationships, as well as corresponding overall mAP. According to Table 3, it can be seen that the pure image-based method (e.g., GRM) usually performs poorly, with only 16.69 for mAP metric. Conversely, the video classifier TSN-ST could achieve 43.23 for mAP, much better than pure image-based methods. This phenomenon teaches us that the recognition of social relationships in videos is significantly dependent on temporal information. Also, the TSN module in MSTR [18] method is combined with spatial and temporal cues of characters and objects, which results in the competitive mAP value of 47.75, which indicates that the fine-grained visual cues in the video are also beneficial.

However, MSTR only relies on visual cues to recognize social relationships, which may fail to integrate multimodal cues, like character dialogue, to reveal the social relationship between characters. Correspondingly, our HC-GCN framework comprehensively utilizes the multimodal cues for social relationship recognition, which achieves the highest mAP metric as 48.74 on the ViSR dataset. On top-1 accuracy of each category, our HC-GCN framework also obtains the highest accuracy on categories of colleague, service, and parent-offs. Interestingly, we find that it has no advantage in hostile relations (leader-sub, opponent). This may be because characters in a hostile relationship do not interact and communicate much in the movie, which impairs the performance. On the contrary, our method performs well in most intimate relationships, e.g., colleague, service, parent-offs, sibling, couple, and friend. Besides, the confusion matrix in Figure 3 of our HC-GCN framework on ViSR dataset shows that the model may make mistake on the intimate relationships such as friend, sibling and couple. This is acceptable as characters who belong to these social relationships often have similarities in their activities and dialogues.

At the same time, as shown in Table 4, for our self-constructed Bilibili dataset, our HC-GCN framework achieves the best performance, which outperforms the other methods by a large margin. This is due to that DSC, MSFM and TEFM are image-based methods for social relationship recognition. Therefore, though they use the multimodal cues during training and inference, they could not take advantages of temporal information, which limits the performance.

Table 5: The Ablation Studies for HC-GCN Framework

Combinations	5-category Classification		
	R(%)	P(%)	F1(%)
Global Video	70.1	56.2	61.8
Global (Video+Text+Audio)	63.3	62.8	61.7
GCN + Global (Video+Text+Audio)	68.9	58.1	62.2

5.5 Ablation Study

To verify the contribution of each module in our HC-GCN framework, we further design several variants to conduct the ablation study, as shown in Table 5. Indeed, two comparisons were designed. For the first comparison, we realize that the global features enhanced by multi-modal information, i.e., textual and audio cues, perform better in precision metric than basic global features, which validates the potential to integrate multi-modal cues, as well as the bias introduced by other modalities. Therefore, it is necessary to refine the multi-modal cues for better performance.

For the second comparison, we further enhance the technical framework with GCN. We believe that although the improvement

(a) A correct case of social graph generation for multiple characters in a video. The model successfully recognizes the "Working" social relationship between "C0", "C1", "C3". The "Opponent" social relationship between "C2" character and other characters is also identified.

(b) Another correct case of social graph generation. Note that the visual region of "C1" character here is very vague, the model may infer the "Kinship" social relationship according to the textural dialogue "Say, brother".

(c) As shown in this case, our method recognizes the social relation between "C0" and "C1" as "Opponent" social relationship while the ground truth is a couple, though the relation at the current video is hostile since "C1" is threatening "C0" with weapon. The behind reason is the social relationships may change along the timeline while the annotation in the movie is unchanged.

Figure 4: The visualization results on our self-constructed dataset. The proposed HC-GCN could generate a social graph for a video, where the nodes and edges in the social graph denote characters and pair-wise social relationships

might not be significant, it can still benefit the discovery of character-pair relationships, which validates our motivation to formulate the social relation recognition task via social graph generation.

5.6 Case Study

Finally, we turn to discuss some interesting rules based on several case studies, which are summarized in Figure 4. Specifically, Figure 4(a) represents a correct social graph generated through HC-GCN which depicts comprehensive social relation in this video, as relations within a dense subgraph could be mutually enhanced. Also, Figure 4(b) shows the benefits of textual information, as the keyword "brother" will lead to the social relation as kinship, when it is confused through the ambiguous visual content. Moreover, Figure 4(c) provides a case that HC-GCN recognizes the social relation between C_0 and C_1 as an opponent, but it may be "wrong" as the ground truth is a couple. It is an interesting example, as the relation of these two characters is finally a couple in the end, but currently, in this frame, they are still hostile as C_1 is threatening C_0 with a sword. This phenomenon reminds us that the social relationships may change along the timeline, thus a fixed annotation might not be proper in some cases.

6 CONCLUSION

In this paper, we proposed a novel Hierarchical-Cumulative GCN structure to capture the social relation among characters in a social graph generation perspective. Specifically, we first integrated the short-term multi-modal cues to generate the frame-level graphs for part of characters via multimodal graph convolution technique. Then, to deal with the video-level aggregation task, we designed an end-to-end framework to aggregate all frame-level subgraphs along the temporal trajectory, which results in a global video-level social graph with various social relationships among multiple characters. Extensive validations on two real-world large-scale datasets demonstrate the effectiveness of our proposed technical framework compared with SOTA baselines, and further revealed some interesting rules of multimodal cues, especially for the advantage that character pairs without co-occurrence could be captured via global social graph, which validated the effectiveness of social graph in revealing social relation among characters.

7 ACKNOWLEDGMENTS

This work was partially supported by the grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (No.61727809, 62072423).

REFERENCES

[1] Hakan Bilen and Andrea Vedaldi. 2016. Weakly Supervised Deep Detection Networks. In *CVPR, 2016*. 2846–2854.

[2] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. 2019. Counterfactual Critic Multi-Agent Training for Scene Graph Generation. In *ICCV, 2019*. 4612–4622.

[3] Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Reinforcement learning based graph-to-sequence model for natural question generation. In *ICLR 2020*.

[4] Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2020. Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings. In *NeurIPS, 2020*.

[5] Andrew C. Gallagher and Tsuhan Chen. 2009. Understanding images of groups of people. In *CVPR, 2009*. 256–263.

[6] Arushi Goel, Keng Teck Ma, and Cheston Tan. 2019. An End-To-End Network for Generating Social Relationship Graphs. In *CVPR, 2019*. 11186–11195.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR, 2016*. 770–778.

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[9] Qingqiu Huang, Wentao Liu, and Dahua Lin. 2018. Person Search in Videos with One Portrait Through Visual and Temporal Links. In *ECCV, 2018*. 437–454.

[10] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2015. Image retrieval using scene graphs. In *CVPR, 2015*. 3668–3678.

[11] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017).

[12] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR, 2017*.

[13] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning Interactions and Relationships Between Movie Characters. In *CVPR, 2020*. 9846–9855.

[14] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. 2019. Heterogeneous Domain Adaptation Through Progressive Alignment. *IEEE Trans. Neural Networks Learn. Syst.* 30, 5 (2019), 1381–1391.

[15] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2017. Dual-Glance Model for Deciphering Social Relationships. In *ICCV, 2017*. 2669–2678.

[16] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated Graph Sequence Neural Networks. In *ICLR, 2016*.

[17] Anan Liu, Yuting Su, Weizhi Nie, and Mohan S. Kankanhalli. 2017. Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1 (2017), 102–114.

[18] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In *CVPR, 2019*. 3566–3574.

[19] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. 2018. Multi-stream Fusion Model for Social Relation Recognition from Videos. In *MMM, 2018*. 355–368.

[20] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *EMNLP, 2020*.

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. [n.d.]. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS, 2015*.

[22] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *NeurIPS, 2013*. 926–934.

[23] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A Domain Based Approach to Social Relation Recognition. In *CVPR, 2017*. 435–444.

[24] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR, 20188*. 6450–6459.

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV, 2016*. 20–36.

[26] Xiaolong Wang and Abhinav Gupta. 2018. Videos as Space-Time Region Graphs. In *ECCV, 2018*. 413–431.

[27] Zhouxia Wang, Tianshui Chen, Jimmy S. J. Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep Reasoning with Knowledge Graph for Social Relationship Understanding. In *IJCAI, 2018*. 1021–1028.

[28] Zhouxia Wang, Tianshui Chen, Jimmy S. J. Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep Reasoning with Knowledge Graph for Social Relationship Understanding. In *IJCAI, 2018*. 1021–1028.

[29] Likang Wu, Zhi Li, Hongke Zhao, Qi Liu, and Enhong Chen. 2021. Estimating Fund-Raising Performance for Start-up Projects from a Market Graph Perspective. *Pattern Recognition* (2021).

[30] Likang Wu, Zhi Li, Hongke Zhao, Qi Liu, Jun Wang, Mengdi Zhang, and Enhong Chen. 2021. Learning the Implicit Semantic Representation on Graph-Structured Data. *DASFAA* (2021).

[31] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang. 2019. Convolutional Attention Networks for Scene Text Recognition. *ACM Trans. Multim. Comput. Commun. Appl.* 15, 1s (2019), 3:1–3:17.

[32] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene Graph Generation by Iterative Message Passing. In *CVPR, 2017*. 3097–3106.

[33] Ning Xu, An-An Liu, Yongkang Wong, Weizhi Nie, Yuting Su, and Mohan S. Kankanhalli. 2021. Scene Graph Inference via Multi-Scale Context Modeling. *IEEE Trans. Circuits Syst. Video Technol.* 31, 3 (2021), 1031–1041.

[34] Tong Xu, Peilun Zhou, Linkang Hu, Xiangnan He, Yao Hu, and Enhong Chen. 2021. Socializing the Videos: A Multimodal Approach for Social Relation Recognition. In *ACM Transactions on Multimedia Computing, Communications, and Applications*.

[35] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph R-CNN for Scene Graph Generation. In *ECCV, 2018*. 690–706.

[36] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *CVPR, 2018*. 5831–5840.

[37] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir D. Bourdev. 2015. Beyond frontal faces: Improving Person Recognition using multiple cues. In *CVPR, 2015*. 4804–4813.

[38] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. Learning Social Relation Traits from Face Images. In *ICCV, 2015*. 3631–3639.