Social Relation Recognition from Videos via Multi-scale Spatial-Temporal Reasoning

Xinchen Liu[†], Wu Liu[†], Meng Zhang[†], Jingwen Chen[§], Lianli Gao[¶], Chenggang Yan[§], and Tao Mei[†]
[†]JD AI Research, Beijing, China
[§]Hangzhou Dianzi University, Hangzhou, China
[¶]University of Electronic Science and Technology of China, Chengdu, China
{liuxinchen1, zhangmeng1208}@jd.com, liuwu@live.cn
{jingwenchen, cgyan}@hdu.edu.cn, lianli.gao@uestc.edu.cn, tmei@live.com

Abstract

Discovering social relations, e.g., kinship, friendship, etc., from visual contents can make machines better interpret the behaviors and emotions of human beings. Existing studies mainly focus on recognizing social relations from still images while neglecting another important media—video. On the one hand, the actions and storylines in videos provide more important cues for social relation recognition. On the other hand, the key persons may appear at arbitrary spatial-temporal locations, even not in one same image from beginning to the end. To overcome these challenges, we propose a Multi-scale Spatial-Temporal Reasoning (MSTR) framework to recognize social relations from videos. For the spatial representation, we not only adopt a temporal segment network to learn global action and scene information, but also design a Triple Graphs model to capture visual relations between persons and objects. For the temporal domain, we propose a Pyramid Graph Convolutional Network to perform temporal reasoning with multi-scale receptive fields, which can obtain both long-term and short-term storylines in videos. By this means, MSTR can comprehensively explore the multi-scale actions and storylines in spatial-temporal dimensions for social relation reasoning in videos. Extensive experiments on a new large-scale Video Social Relation dataset demonstrate the effectiveness of the proposed framework. Our dataset is available on https://lxc86739795.github.io.

1. Introduction

Social relation is the close association between multiple individual persons and forms the basic structure of our society. Recognizing social relations from images or videos can empower machines to better understand the behaviors or emotions of human beings. However, compared to image-based social relation recognition [32], the video-based scenario is an important but frontier topic which is often neglected by the community. It has many potential applications such as family video search on mobile phones [27] and product recommendation to groups of customers in stores [21].

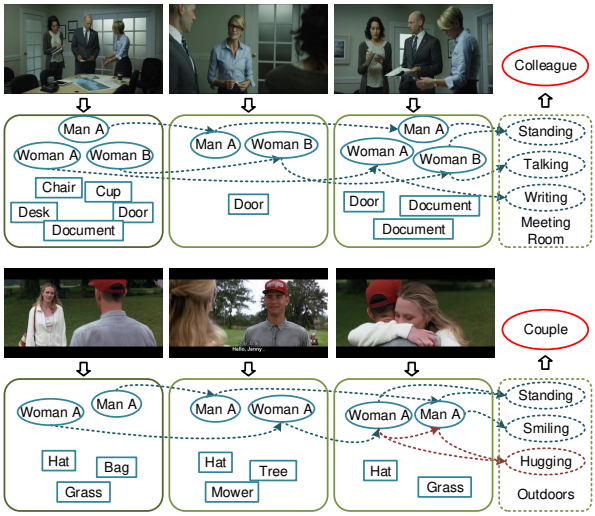

Figure 1. How do we recognize colleagues or couples from a video? The appearance of persons, interactions between persons, and the scenes with contextual objects are key cues for social relation recognition.

Sociological analytics from visual content has been a popular area during the last decade [27, 31]. Existing social recognition studies mainly focus on image-based condition, in which algorithms recognize social relations between persons in a single image. Appearance and face attributes of persons and contextual objects are explored to distinguish different social relations [26, 30, 33]. Although discovering social networks [5], communities [6], roles [23, 24], and

group activity [1, 2] in videos or movies has been widely studied, explicit recognition of social relations from video clips attract far less attention. Recent methods only consider video-based social relation recognition as a general video classification task [8], which take RGB frames, optical flows, or audio of a video as the input and categorize video clips into pre-defined types [20]. However, such a general model is obviously over-simplified, which neglects the appearance of persons, interactions between persons, and the scenes with contextual objects as shown in Figure 1.

Social relation recognition from videos faces unique challenges. Firstly, compared to social community discovery, social relations are more fine-grained and ambiguous in different scenes. The models must discriminate very similar social relations such as friends and colleagues through visual contents, which might be very difficult even for human beings. Moreover, in contrast to image-based social relation recognition, persons and objects could appear in arbitrary frames or even separate frames. This makes persons and objects extremely varied in sequential frames. Therefore, image-based methods cannot be directly adopted for video-based scenarios. Furthermore, videos provide dynamics of persons or objects in the temporal domain than still images. However, the location and duration of a key action for discriminating a social relation are uncertain in a video. Modeling the latent correlation between the varied dynamics of persons and social relations remains great challenging.

To this end, we propose a Multi-scale Spatial-Temporal Reasoning (MSTR) framework for social relation recognition from videos. The multi-scale reasoning is two-fold. **In the spatial domain**, we consider both global cues and semantic regions like persons and contextual objects. In particular, we adopt Temporal Segment Network (TSN) [28] to learn global features from scenes and backgrounds in the full frames. Moreover, we also design a Triple Graphs model to represent the visual relations between persons and objects. The multi-scale spatial features can provide complementary visual information for social relation recognition. **For the temporal domain**, we propose a Pyramid Graph Convolution Network (PGCN) to perform temporal reasoning on the Triple Graphs. Specifically, we apply multi-scale receptive fields in the graph convolution block, which can capture temporal features from both long-term and short-term dynamics in videos. Finally, our MSTR framework achieves social relation recognition by spatial-temporal reasoning from comprehensive information in videos.

In summary, the contributions of this paper include:

- We propose a Multi-scale Spatial-Temporal Reasoning framework to recognize social relations from videos with global and local information in the spatial-temporal domain.
- We design a novel Triple Graphs model to represent visual relations of persons and objects. By combining with TSN for global features, our framework can learn multi-scale spatial features from video frames.
- To effectively capture both the long-term and short-term temporal cues in videos, we propose a PGCN which performs relation reasoning with multi-scale temporal receptive fields.

In addition, to validate our framework and facilitate research, we build a large-scale Video Social Relation dataset, named ViSR. It not only contains over 8,000 video clips labeled with eight common social relations in daily life, but also has diverse scenes, environments and backgrounds. Extensive experiments on the ViSR dataset show the effectiveness of the proposed framework.

2. Related Work

Social relation discovery in visual content. The interdisciplinary study of sociology and computer vision has been a popular area in the last decade [5, 25, 26, 27]. The main topics include social networks discovery [6, 31], key actors detection [23, 24], multi-person tracking [1], and group activity recognition [2].

In recent years, explicit recognizing social recognition from visual content has attracted attention from researchers [12, 26, 30, 32]. Existing methods are mainly focused on still images. For example, Zhang *et al.* proposed to learn social relation traits from face images by a Convolutional Neural Network (CNN) [32]. Sun *et al.* proposed a social relation dataset based on the social domain theory [3] and adopted a CNN to recognize social relations from a group of semantic attributes [26]. Li *et al.* proposed to a dual-glance model for social relationship recognition, where the first glance focused persons of interest and the second glance applied attention mechanism to discover contextual cues. [12]. Wang *et al.* proposed to represent the persons and objects in an image as a graph and perform social relation reasoning by a Gated Graph Neural Network [30]. For the video-based condition, social relation recognition is only considered as a video classification task. For example, Lv *et al.* exploited the Temporal Segment Networks [28] to classify a video using the RGB frames, optical flows, and audio of the video [20]. They also built a Social Relation In Video (SRIV) dataset which contained about 3,000 video clips with multi-label annotation. However, this method only considered global and coarse features while neglecting the persons, objects, and scenes in videos. Therefore, we propose to embed the spatial and temporal features of persons and objects into a Triple Graphs model, on which social relation reasoning is performed.

Graph model in computer vision. In the computer vision field, pixels, regions, concepts, and prior knowledge can be represented as graphs to model their relations for

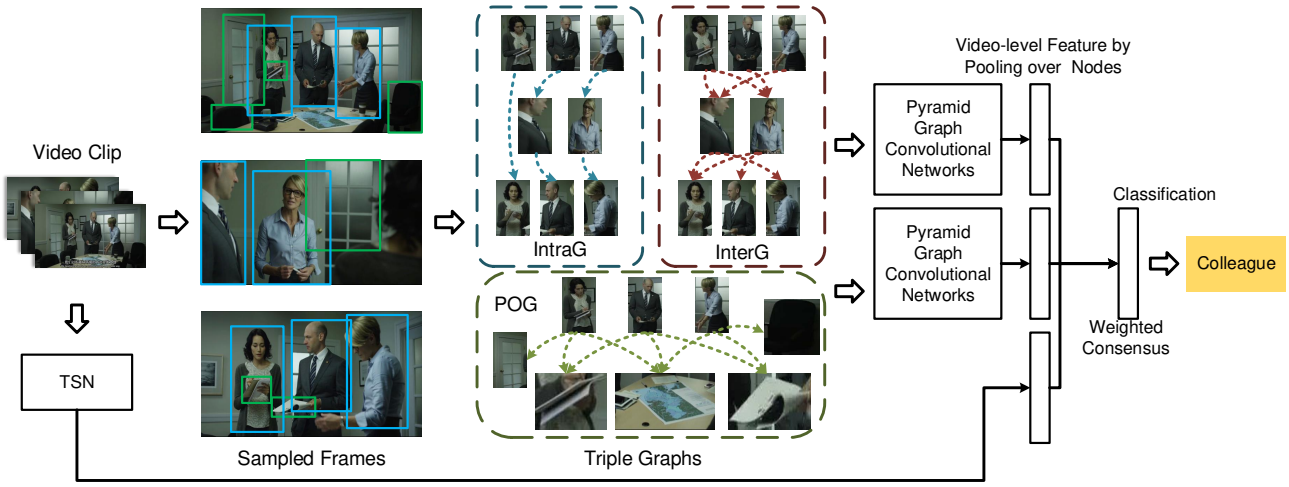

Figure 2. The overall framework of the Multi-scale Spatial-Temporal Reasoning framework.

different tasks such as object shape detection [14], image segmentation [7], image retrieval [17], vehicle search [19], etc. In recent years, researchers of machine learning have studied message propagation in graphs by end-to-end trainable networks, such as Graph Convolutional Networks (GCN) [4, 11]. Most recently, these models have been adopted to computer vision tasks [13, 22, 29, 30]. For example, Liang *et al.* proposed a Graph Long Short-Term Memory to propagate information in graphs built on super-pixels for semantic object parsing [13]. Qi *et al.* proposed a 3D Graph Neural Network to build a k-nearest neighbor graph on 3D point cloud and predict the semantic class of each pixel for RGBD data [22]. Wang *et al.* 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 [29]. Inspired by above studies, we propose to represent the actions and interactions of persons and objects in videos as graphs, on which reasoning is performed by a novel pyramid graph convolutional network for social relation recognition.

3. The Proposed Framework

3.1. Overview

The overall architecture of the Multi-scale Spatial-Temporal framework mainly contains two parts as shown in Figure 2. The first part is the construction of the Triple Graphs structure. The framework takes as input one video clip which is sampled into F frames for efficiency. To capture local details from regions of interest, the persons and objects are first cropped from frames with Mask R-CNN [9] pre-trained on MS-COCO dataset [15]. To model the spatial and temporal representation of persons and objects, we build an Intra-Person Graph (IntraG) for the same person, an Inter-Person Graph (InterG) for different persons, and a Person-Object Graph (POG) to capture the co-existence of persons and contextual objects. The ResNet [10] is adopted to extract the spatial features of persons and objects. The second module adopts the PGCN to perform relation reasoning by message propagation in each graph. In PGCN, multi-scale temporal receptive fields are explored to learn dynamics in varied temporal range. The node-level features are fused into a normalized graph-level representation for each graph. Besides, a global video classification network, such as TSN [28] or T-C3D [16], is exploited to learn global features by taking as input of full frames. Finally, the social relation in a video is predicted by integrating the global feature from TSN and the reasoning feature from the PGCN. Next, we will present the details of the construction of the Triple Graphs model and relation reasoning by PGCN.

3.2. Triple Graphs Model

We can recognize the social relationships in videos by observing the actions of persons, the interactions between persons, and the co-existence of persons and contextual objects in the scenes, as shown in Figure 1. Graph model has been found effective to represent the spatial, temporal, conceptual, or similarity relations of objects in visual content [17, 18, 29]. Therefore, we design a Triple Graphs model, which includes three types of graphs, to model the visual relations of persons and objects for social relation reasoning in videos, as shown in Figure 2. To build the Triple Graphs, we first detect bounding boxes of persons and objects as $P = \{p_1, p_2, ..., p_N\}$ and $O = \{o_1, o_2, ..., o_M\}$ from the sampled F frames by Mask R-CNN [9]. To balance the accuracy and efficiency, we remain fixed N persons and M objects for each video by confidence scores. The feature of each bounding box is extract-

ed by the backbone network $f(\cdot)$, i.e., ResNet [10]. These bounding boxes are adopted as the nodes to build the graphs, while the features of each node will be utilized in graph convolution for social relation reasoning.

Intra-Person Graph. We model the appearance variance of the same person through the video by the Intra-Person Graph (IntraG). The IntraG is represented by an adjacent matrix $A_s \in \mathbb{R}^{N_p \times N_P}$, in which the indexes of rows and columns correspond to the temporal order of bounding boxes in the video. To match the same person in different frames, we measure the visual similarity between each pair of persons, (p_i, p_j), in two neighboring frames. Therefore, the adjacent matrix A_s is filled by:

$$A_s(p_i, p_j) = \begin{cases} 1 & dist(p_i, p_j) < \tau, \\ 0 & otherwise, \end{cases} \tag{1}$$

where $dist(p_i, p_j) = 1 - \frac{f(p_i)^T f(p_j)}{||f(p_i)|| \cdot ||f(p_j)||}$ is the cosine distance of p_i and p_j, τ is a hyper-parameter.

Inter-Person Graph. To capture the interactions between different persons in videos, we build the Inter-Person Graph (InterG) by estimating the distances of persons in one frame and its neighboring frame. For the adjacent matrix of InterG, $A_d \in \mathbb{R}^{N_p \times N_P}$, we directly set $A_d(p_i, p_j) = 1$, if p_i and p_j are two persons in one frame. For p_i and p_j in neighboring frames, we set

$$A_d(p_i, p_j) = \begin{cases} 1 & dist(p_i, p_j) >= \tau, \\ 0 & othersise, \end{cases} \tag{2}$$

where $dist(p_i, p_j)$ is also the cosine distance of p_i and p_j.

Person-Object Graph. The contextual objects in the scene are vital information for social relation recognition. However, due to the shot changes, the persons and contextual objects may become varied in different frames, which makes it difficult to capture the interactions between persons and contextual objects through one video. Therefore, different from IntraG and InterG for persons, the Person-Object Graph (POG) is designed to model the co-existence of persons and contextual objects. The adjacent matrix of POG, $A_o \in \mathbb{R}^{(N_p+N_o)\times(N_P+N_o)}$, represents the relation between each person and the objects that exist in the one frame. Therefore, we set $A_o(p_k, o_l) = 1$, if p_k and o_l are from the same frame, and $A_o(p_k, o_l) = 0$, otherwise.

To this end, the Triple Graphs are built to represent the visual relations of persons and objects in videos, i.e., the appearance and actions of each person, the interactions between different person, and the co-existence of persons and objects. In particular, the indexes of adjacent matrixes correspond to the temporal order of bounding boxes in the video, by which the temporal information is implicitly embedded in the graphs. Next we present how to perform social relation reasoning from visual features embedded in the graphs by the Pyramid Graph Convolutional Network.

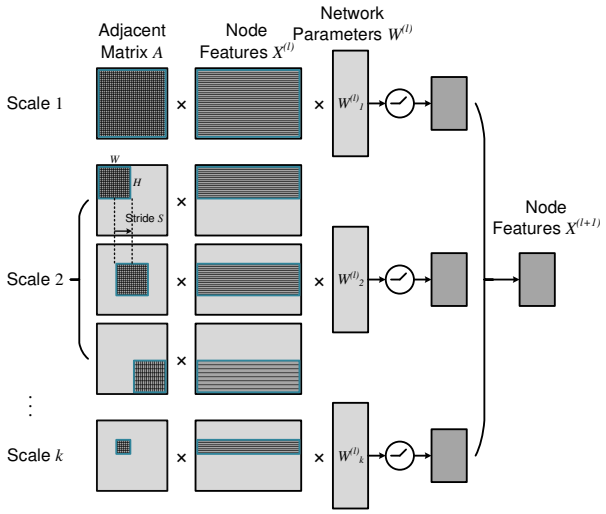

Figure 3. Pyramid Graph Convolution Block with multi-scale receptive fields in the temporal domain. Here we use A to represent the normalized adjacent matrix $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ for simplicity.

3.3. Reasoning by Pyramid GCN

Graph Convolutional Network. Traditional Convolutional Neural Networks usually apply 2-D or 3-D filters on images or videos to abstract visual features from low-level space to high-level space [10]. In contrast, Graph Convolutional Network (GCN) performs relational reasoning by performing message propagation from nodes to its neighbors in the graphs [11]. Therefore, we can apply GCNs on the Triple Graphs to achieve social relation reasoning.

As in [11], 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:

$$X^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}), \tag{3}$$

where $\tilde{A} \in \mathbb{R}^{N\times N}$ is the adjacent 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. In particular, in our social relation reasoning framework, the adjacent matrixes of the Triple Graphs are A_s, A_d, and A_o as defined in Section 3.2. The indexes of adjacent matrixes are arranged by the temporal order of the nodes in a video, by which the temporal information is implicitly embedded in the built graphs. $X^{(0)} = [f(x_1), f(x_2), ..., f(x_N)]^T$ is the initial feature matrix, where $f(x_i)$ is the column vector extracted from the nodes $\{x_i\}^N$ like persons or objects in videos. The final outputs of the GCNs are updated features of nodes, $X^{(L)}$, in the graphs, which can be aggregated into a video level feature vector for social relation prediction.

Pyramid Graph Convolutional Network. GCN per-

Table 1. The descriptions of social relations in the ViSR dataset based on the domain theory [3].

Domain	Relation	Examples
Attachment	Parent-offspring	Parent-child, Grandparent-grandchild
Mating	Couple	Husband-wife, boyfriend-girlfriend
Hierarchical power	Leader-subordinate	Teacher-student, team leader-member
	Service	Passenger-driver, Customer-waiter
Reciprocity	Sibling	Brothers, sisters
	Friend	Friends in general scenes
Coalitional groups	Colleague	Co-worker, school mate, teammate
	Opponent	Enemy, competitor, disputant

forms operations on all nodes in one graph together as well as the full temporal range of a video, which means GCN can capture a global view in the temporal domain. However, the key factor for social relation recognition such as a specific action of a person may appear in local temporal position which may be overwhelmed by unimportant information. Therefore, we design a Pyramid Graph Convolutional Network (PGCN) to learn both long-term and short-term information by a pyramid of temporal receptive fields.

Figure 3 illustrates the structure of one pyramid graph convolution block in PGCN. Each block contains multiple parallel branches with different receptive fields. Scale 1 is the standard GCN which performs graph convolution on the whole adjacent matrix and covers all nodes in the graph. Scale 2 gives an example of graph convolution with a smaller temporal receptive field, while Scale K is a more general illustration. For each scale, the activations of all sliding windows are aggregated into one feature matrix which has the same shape with the output of the standard GCN. By sliding the receptive field along the diagonal of the adjacent matrix, the model can learn the relatively short-term features from the start to the end of a video. At last, the outputs of multiple scales are merged by average pooling to generate the feature matrix, $X^{(l+1)}$, for the next PGCN layer. The pyramid graph convolution block is end-to-end differentiable and can be inserted into other video-based GCN models for action recognition or video classification [29].

In our implementation, we stack two pyramid graph convolution layers of which the scales of the parameter matrix $W^{(l)}$ are 2048×512 and 512×128. In each pyramid graph convolution block, we adopt two scales of filters. The first scale has $N \times N$ filter, while the second scale has $\frac{N}{2} \times \frac{N}{2}$ filter and stride $S = \frac{N}{4}$. After forward propagation of PGCN, the final feature matrix $X^{(L)} \in \mathbb{R}^{N \times 128}$ is aggregated into a 128-D video-level feature vector. The video-level feature is fed into a fully connected layer to classify the video into one social relation class. In our framework, the pyramid temporal reasoning is performed by PGCNs on IntraG, InterG, and POG separately. The three branches generate a weighted consensus after the softmax layers in each branch. Moreover, to learn more global visual information about the scenes, environments, and backgrounds, we adopt the TSN [28] to directly take as input all sampled frames from a video. At last, the scores of PGCN and TSN are combined by weighted fusion for the final prediction.

4. Experiments

4.1. The ViSR Dataset

Existing datasets for social relation recognition are mainly based on still images [12, 26, 32]. The social relations of these datasets are defined by different psychological or sociological theories. For example, the social relation dataset in [32] is mainly focused on psychological or emotional traits. Therefore, the images in this dataset are annotated with attributes of faces like expressions. The People in Photo Album (PIPA) dataset [26] and People in Social Context (PISC) [12] dataset are both defined on sociological theories. The labels of PIPA are based on the social domain theory [3], in which social life is partitioned into five domains and 16 social relations. The PISC dataset contains several common social relations in daily life, which have a hierarchy of three coarse-level relationships and six fine-level relationships.

However, video-based dataset labeled with explicit social relations is rare. One of the largest is the Social Relation in Video (SRIV) dataset which contains about 3000 video clips collected from 69 movies [20]. It is annotated with eight subjective relations which are similar to the social relation traits in [32], and eight objective relations which are derived from the domain-based relations in [3]. There are three main limitations in SRIV: 1) the volume of the dataset is relatively small for the scalability of the models especially for CNN; 2) the videos are labeled by multiple labels, which makes the relation in a video ambiguous; 3) the social relations are very unbalanced especially for the objective relations.

To facilitate related research and validate our proposed framework, we build a large-scale and high-quality Video based Social Relation dataset, dubbed as ViSR. For our dataset, we define eight types of social relation derived from the domain-based theory [3], as listed in Table 1. The construction process contains three main steps: **1)** We first collect more than 200 movies which have a wide variety of types such as adventure, family, comedy, drama, crime, ro-

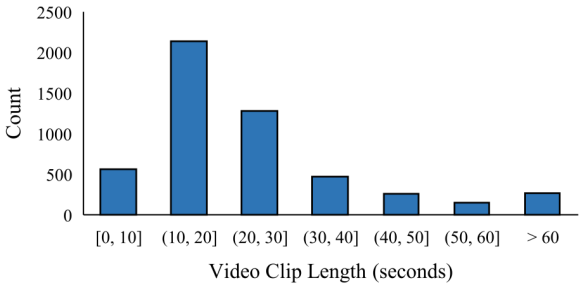

Figure 4. The statistics of video clip length in ViSR dataset.

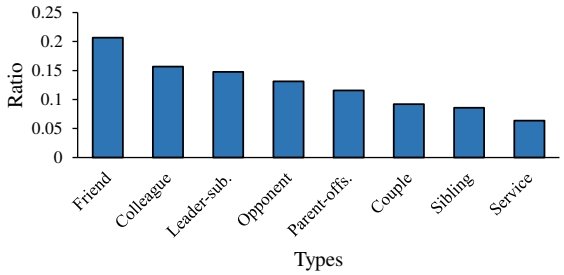

Figure 5. The distribution of social relations in the ViSR dataset.

mance, action, biography but exclude surreal types like fantasy and Sci-Fi. **2)** We then ask ten annotators to segment video clips from the movies. The length of each clip is limited in $10 \sim 30$ seconds. At least two persons that have interactions must exist in one clip. The scene in one clip should be fixed. By this means, we obtain about 10,000 candidate video clips for annotation. **3)** At last, each candidate video clip is labeled by at least five annotators by maximum voting to guarantee the quality. The clip will be discarded if all its labels are less than three votes.

Through elaborate annotation, the ViSR has several featured properties. First of all, the dataset contains more than 8,000 valid video clips, which can make the algorithms more scalable than existing datasets. Moreover, due to the variety of source movies, our dataset not only covers most common social relations in daily life with balanced class distribution as shown in Figure 4, but also contains various scenes, environments, and backgrounds, which makes ViSR a challenging dataset. Furthermore, as shown in Figure 5, the length of most clips is limited in 30 seconds to keep the stable scenes, which reduces the ambiguity of relations in videos. Figure 6 shows some examples of video clips in our dataset. In the experiments, we randomly split the dataset into training, validation, and testing subsets by the ratio $7:1:2$. The top-1 accuracy on each relation class and the mean Average Precision (mAP) over all classes are calculated to evaluate the performance of methods.

4.2. Implementation Details

This section presents the details on the construction of Triple Graphs and training strategy of the networks.

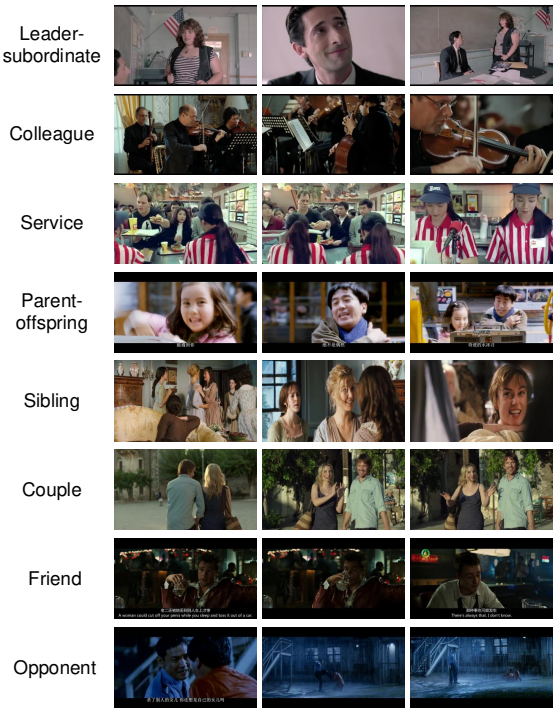

Figure 6. Some examples of videos in the ViSR dataset.

Triple Graphs Building. The Triple Graphs model is built as in Section 3.2. We uniformly partition an input video into 20 segments, in which one frame is randomly sampled to obtain 20 frames for one video. From the sampled frames, we adopt Mask R-CNN to obtain at most 40 bounding boxes of persons and 20 bounding boxes of objects. For construction of IntraG and InterG, the person similarity threshold τ in Equ. 1 and Equ. 2 are set to 0.2.

Networks Training. In our framework, the PGCN and TSN are trained separately. In each pyramid graph convolution After the construction of the Triple Graphs, three PGCNs of IntraG, InterG, and POG are pre-trained on the training set separately with the learning rate $lr = 0.01$. After 30 epochs, the three PGCNs are trained together for 120 epochs in which the learning rate starts from 0.001 and multiply 0.1 by every 30 epochs. The TSN is trained by the standard strategy as in [28]. The segment number is set to 20. The base learning rate is 0.001 and multiplies 0.1 by every 20 epochs until 80 epochs. For testing, the fusion weights for the results of PGCN and TSN are 0.6 and 0.4, respectively.

4.3. Comparison with the State-of-the-art Methods

To validate the effectiveness of the proposed Pyramid Temporal Reasoning framework, we compare it with several state-of-the-art methods on the ViSR dataset. The details of methods are as follows:

Table 2. Comparison to the state-of-the-art methods.

	Top-1 Accuracy								
	Leader-Sub.	Colleague	Service	Parent-offs.	Sibling	Couple	Friend	Opponent	mAP
GRM [30]	48.67		6.67	0.00		4.17	0.67	30.13	16.69
TSN-Spatial [20]	55.48	42.93	30.00	35.20	34.83	**39.78**	48.75	37.07	42.38
TSN-ST [20]	41.05	33.33	30.00	32.83	**45.78**	29.17	63.76	32.87	43.23
GCN	56.16	49.46	27.14	36.80	41.57	34.41	39.80	**50.00**	43.46
PGCN	54.11	**54.89**	25.71	40.80	34.83	33.33	45.27	48.28	44.73
MSRT	**57.53**	51.09	**30.00**	**45.60**	39.33	38.71	**53.23**	47.41	**47.75**

Table 3. Ablation study on the proposed framework.

	Module			Top-1 Accuracy								
	IntraG	InterG	POG	Leader-Sub.	Colleague	Service	Parent-offs.	Sibling	Couple	Friend	Opponent	mAP
GCN	✓			49.32	44.57	25.71	38.40	38.20	26.88	44.78	43.97	41.02
	✓	✓		52.74	48.91	25.71	38.40	**42.70**	29.03	42.29	44.83	42.48
	✓	✓	✓	**56.16**	49.46	27.14	36.80	41.57	34.41	39.80	**50.00**	43.46
PGCN	✓			52.74	52.17	25.71	**45.60**	40.45	34.41	38.81	40.52	43.07
	✓	✓		53.42	51.63	**27.14**	43.20	38.20	**39.78**	41.29	40.52	43.65
	✓	✓	✓	54.11	**54.89**	25.71	40.80	34.83	33.33	**45.27**	48.28	**44.73**

1) Temporal Segment Network using Spatial features (TSN-Spatial) [20]. This method uses only the RGB frames of videos as the input and adopts TSN to learn spatial features for social relation recognition. We use the parameters and training strategy as in [20]. We modify the original multi-label classification setting on their dataset to single-label classification task for our dataset.

2) Temporal Segment Network using Spatial-Temporal features (TSN-ST) [20]. This method uses the same framework with TSN-Spatial except that the optical flow is also taken as the input of TSN to learn both spatial and temporal features from videos. The implementation is the same as that in [20]. Because this paper is mainly focused on vision based methods, we do not use any audio information as in [20]. Therefore we consider this model as the state-of-the-art method on the SRIV dataset.

3) Graph Reasoning Model (GRM) [30]. This is the state-of-the model for image-based social relation recognition on two public datasets, i.e., PIPA [26] and PISC [12]. We apply GRM on each frame in a video. The results on all sampled frames are integrated by late fusion for video-based social relation prediction.

4) Graph Convolution Networks (GCN). In this model, we only adopt the standard GCN to perform reasoning on the Triple Graphs.

5) Pyramid Graph Convolution Networks (PGCN). In this model, we insert the temporal pyramid branches into each graph convolutional layer in GCNs.

6) Multi-scale Spatial-Temporal Reasoning (MSTR). This is the complete Pyramid Temporal Reasoning framework, which adopts PGCN to learn multi-scale dynamics of persons from the Triple Graphs and TSN to learn global spatial features. Finally, the social relation reasoning is achieved by weighted fusion of PGCN and TSN.

The results of these methods are listed in Table 2. We first find that the image-based method, GRM, obtains poor results on the video-based dataset. The reason is that image-based methods require the co-existence of two or more persons in one image, while in the video base condition there may be only one person in a frame. Therefore image-based method cannot be directly adopted to the video-based scenario. Moreover, by comparison of global feature based models, i.e., TSN-Spatial and TSN-ST, and graph-based methods, i.e., GCN and PGCN, we can find that global information and local regions are both effective for social relation recognition. Overall GCN and PGCN are better, because the detailed appearance and actions of persons and objects can provide more significant features for social relations. Furthermore, the combination of TSN and PGCN, i.e., MSTR obtain the best performance, which demonstrates the complementary effect of multi-scale spatial and temporal representation.

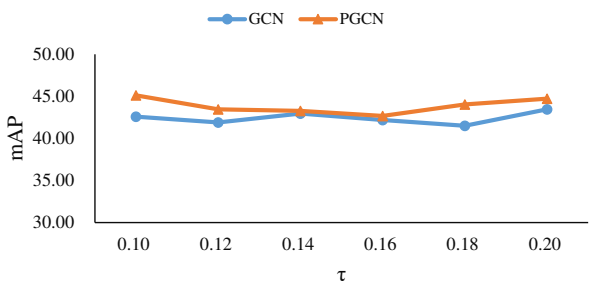

Figure 7. The results under different τ for building graphs.

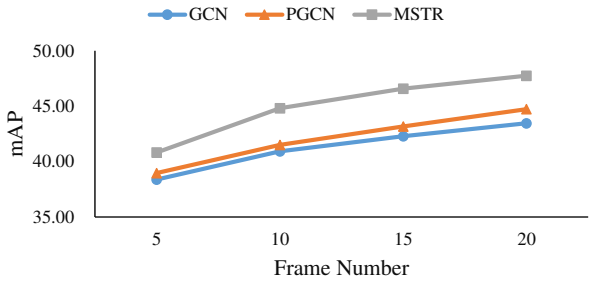

Figure 8. The results of different frame numbers for MSTR.

4.4. Ablation Study

Significance of Triple Graphs Here we explore the effect of each graph and the pyramid graph convolution block in the PGCN. Table 3 lists the results of models with different graph combinations for both GCN and PGCN. From the results, we can find that the overall accuracy of PGCN is higher than that of GCN, which demonstrates that the multi-scale receptive fields can capture useful features from long-term and short-term ranges. Moreover, for each network architecture, the mAP increases by incorporating IntraG, InterG, and POG. This validates the significance of actions, interactions between persons, and co-existence of persons and contextual objects for social relation recognition. We also observe that three graphs show different effects on different social relations. For examples, the POG brings significant boost on work relations, i.e., leader-subordinate and colleague. This reflects the importance of contextual objects in work scenes like office or meeting room.

Analysis on Hyper Parameters We explore the impact of two hyper parameters, i.e., the sampled frame number F and threshold τ in Section 3.2. We first set $\tau = 0.1$ to 0.2 for graph construction in both GCN and PGCN. The results are shown in Figure 7. The curves are stable under different τ, which shows the robustness of our Triple Graphs model. For sampled frame number, we compare the results of GCN, PGCN, and MSTR for $F = 5, 10, 15, 20$. Figure 8 shows that the mAP increases with the growth of input frames. This demonstrates that our graphs not only exploit more useful information from more frames, but also are robust to the noise from extra data.

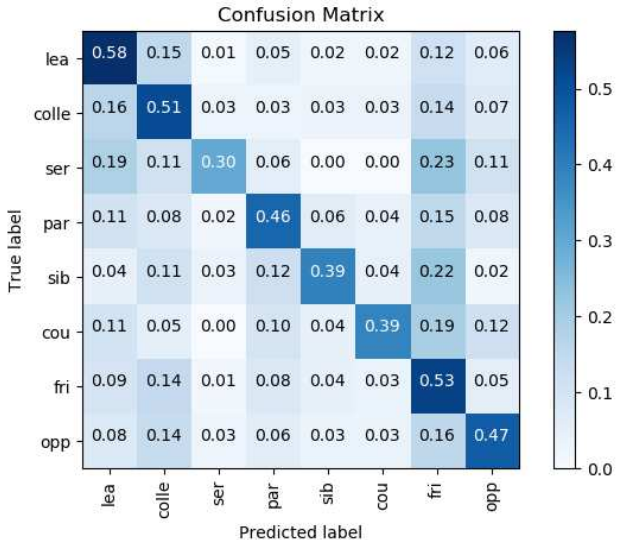

Figure 9. The normalized fusion matrix of the MSTR framework.

4.5. Discussion

From the experimental results, we can observe that explicit recognition of social relations from video clips is a challenging task. Figure 9 shows the confusion matrix of our MSTR framework. It is difficult to distinguish very similar relations only using visual content. For example, friend, sibling, and service may be very ambiguous if we only focus on the persons in the video. In this condition, the context like the scenes, backgrounds, objects may be more important for relation reasoning. Currently, we only simply adopt a standard TSN model to learn context cues. In future work, contextual information should be further mined for social relation recognition in videos.

5. Conclusion

In this paper, we propose a Multi-scale Spatial-Temporal Reasoning framework to recognize social relations from videos. The MSTR can learn robust representation which exploits multi-scale features in both spatial and temporal domains. To represent the appearance and actions of persons and objects, we propose a Triple Graphs model to capture the visual relations of nodes. By combining global features learned by TSN, our framework can learn multi-scale spatial features from video frames. To learn both long-term and short-term temporal cues in videos, we propose a Pyramid Graph Convolutional Network which performs relation reasoning with multi-scale temporal receptive fields. Extensive experiments on a large-scale and high-quality video social relation dataset demonstrate the effectiveness of the proposed framework.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Fei-Fei Li, and Silvio Savarese. Social LSTM: human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016. 2

[2] Timur M. Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, pages 3425–3434, 2017. 2

[3] Daphne Blunt Bugental. Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187, 2000. 2, 5

[4] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3837–3845, 2016. 3

[5] Lei Ding and Alper Yilmaz. Learning relations among movie characters: A social network perspective. In *ECCV*, pages 410–423, 2010. 1, 2

[6] Lei Ding and Alper Yilmaz. Inferring social relations from visual concepts. In *ICCV*, pages 699–706, 2011. 1, 2

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3

[8] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J. Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, pages 5589–5597, 2018. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 4

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. 3, 4

[12] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *ICCV*, pages 2669–2678, 2017. 2, 5, 7

[13] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph LSTM. In *ECCV*, pages 125–143, 2016. 3

[14] Liang Lin, Xiaolong Wang, Wei Yang, and Jian-Huang Lai. Discriminatively trained and-or graph models for object shape detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):959–972, 2015. 3

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 3

[16] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. T-C3D: temporal convolutional 3d network for real-time action recognition. In *AAAI*, pages 7138–7145, 2018. 3

[17] Wei Liu, Yu-Gang Jiang, Jiebo Luo, and Shih-Fu Chang. Noise resistant graph ranking for improved web image search. In *CVPR*, pages 849–856, 2011. 3

[18] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016. 3

[19] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018. 3

[20] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. Multi-stream fusion model for social relation recognition from videos. In *MMM*, pages 355–368, 2018. 2, 5, 7

[21] You-Jin Park and Kun-Nyeong Chang. Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications*, 36(2):1932–1939, 2009. 1

[22] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for RGBD semantic segmentation. In *ICCV*, pages 5209–5218, 2017. 3

[23] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander N. Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *CVPR*, pages 3043–3053, 2016. 1, 2

[24] Vignesh Ramanathan, Bangpeng Yao, and Fei-Fei Li. Social role discovery in human events. In *CVPR*, pages 2475–2482, 2013. 1, 2

[25] Ashtosh Sapru and Hervé Bourlard. Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5):746–760, 2015. 2

[26] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *CVPR*, pages 435–444, 2017. 1, 2, 5, 7

[27] Gang Wang, Andrew C. Gallagher, Jiebo Luo, and David A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, pages 169–182, 2010. 1, 2

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 2, 3, 5, 6

[29] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 413–431, 2018. 3, 5

[30] Zhouxia Wang, Tianshui Chen, Jimmy S. J. Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *IJCAI*, pages 1021–1028, 2018. 1, 2, 3, 7

[31] Ting Yu, Ser-Nam Lim, Kedar A. Patwardhan, and Nils Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, pages 1462–1469, 2009. 1, 2

[32] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *ICCV*, pages 3631–3639, 2015. 1, 2, 5

[33] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 1