Understanding Social Relationship with Person-Pair Relations

#229

Abstract

Social relationships understanding is to infer the social relations among people in a given scene, which has attracted increasing attention in computer vision recently. Great progress has been made since the rise of deep learning. However, all of previous works infer the social relationship of each person pair in isolation, without taking into account the interaction of different pairs. It is natural to consider these interaction cues, *i.e.*, the mutual influence of multiple person pairs, in social relationship understanding. For instance, if two person pairs in an image are *Friends*, then the third pair is always *Friends* or at least other similar relations but not *No Relation*. Therefore, to capture these interaction cues, we propose a novel end-to-end trainable Person-Pair Relation Network (PPRN) using standard RNNs, a graph inference network that learns iteratively to improve its predictions via message passing among person pair nodes. Extensive experiments on PISC and PIPA-Relation show the superiority of our method over previous methods.

1 Introduction

Social relationships in either physical or virtual world form the basic of the social network in our daily life. Previous studies have shown that the implicit social relationships can be discovered from texts [Fairclough, 2003], images [Li *et al.*, 2017; Wang *et al.*, 2018; 2010; Zhang *et al.*, 2015b], and videos [Ding and Yilmaz, 2010; Ramanathan *et al.*, 2013; Vinciarelli *et al.*, 2009]. Images and videos have attracted increasing attention in computer vision community recently, and in this paper we mainly focus on still images.

The aim of *social relationships understanding* is to infer the social relations among people in a given scene such as a still image, which is a significantly important study in computer vision recently. For instance, nowadays with the increasing dependence of human on machines, understanding the social relationships enables the machines to blend in and make a better response in different situations. Besides, social relationships understanding is also helpful for avoiding the potential privacy risks via automatically parsing the information that may reveal social relations in many medias such as texts [Fairclough, 2003] and informing the users about this.

Many interesting studies has been made since the rise of deep learning. For example, [Sun *et al.*, 2017] used the information of head regions, body regions and human attributes to predict the person-pair's social relationships separately. [Li *et al.*, 2017] made use of the person pair and region proposals and allocated attention to each region to for various person pairs. [Wang *et al.*, 2018] built the Graph Reasoning Model (GRM) that incorporates common sense knowledge of the correlation between objects and a person pair.

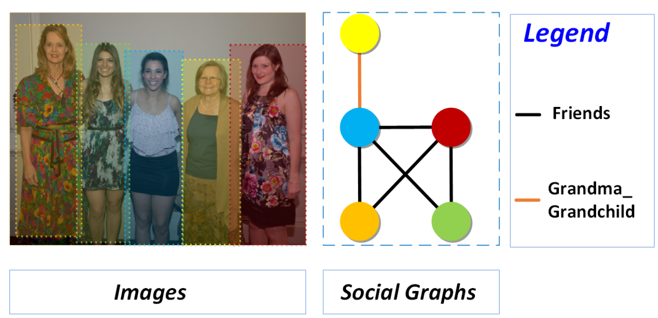

Figure 1: An example of an image and its social relation graph from PIPA-relation datasets. Each person under a mask corresponds to the node with the same color in the graph. An edge linked to two nodes denotes the social relation between them. There are many examples with stable social relationships in whole dataset that the same as this example, and the detail statistic is in Sec. 4.2.

However, all of previous works, a relationship between two people is predicted independently, and the interaction between those relationships, which can help the prediction, are ignored. There is an example from PIPA-relation [Sun *et al.*, 2017] shown in Figure 1, the social relationship of six person pairs are *Friends* except one is *Grandma_grandchild*, which is mostly a group of friends. Intuitively, if we want to infer the social relationships of a person pair and already know several others with same relationship such as *Friends*, then the person pair has a high probability of being *Friends*. Therefore, the cues of contextual social relationships play an important role in social relationships understanding. Furthermore, we made a detail analysis on PISC and PIPA-Relation datasets, and found that in many images, the relationships of person pairs in the same image tend to be similar, which inspires us to take into account the interaction among person pairs in this task.

It is still challenging to infer an unknown social relationship between two persons in an image. Based on the cues we mentioned above, a mechanism is needed to model the interaction between person pairs to predict known social relationships. Most of the previous works are incurring the region of the contextual object, and this contextual region is not always correctly detected by extra annotation. So, whether to introduce information about the objects in the scene has also become an important consideration for us. Specifically, the proposing method should avoiding introduce uncertain information, and the approach also model the interaction of various person pairs in an image well.

To address this problem, we propose a novel end-to-end trainable Person-Pair Relation Network (PPRN) which makes good use of the information of social relationships interaction in the same scene. PPRN consists of 3 parts: *feature extraction*, *message pooling*, and *message passing*. In the feature extraction module, we use a ResNet [He *et al.*, 2016] to extract features of each person and another ResNet to extract features of the person pairs in the image. In this module, the position of each person is also taken into account. In the message pooling module, we use an attention mechanism to make all the social relationships interact with each other well and the output of this module will be considered as the inputs of the GRU unit. In the message passing module, we use the RNNs contained GRU to make the social relationships message passing proceed iteratively. As the model proceeding, this message passing module and the message pooling module will have an interaction, and make the overall mechanism better. We use the hidden states of the last GRU as the social relationships detection results of the scene.

We evaluate our model in classic datasets: PIPA-Relation dataset and PISC dataset. The experiment results verify the superiority of our model over previous methods.[1] Specifically, our contributions are as follows:

- To our best knowledge, it is the first attempt to introduce the idea of multiple person pair's relationships in the same scene on the social relationship detection understanding;
- We design a novel interaction mechanism to model the interaction of social relationships in the same scene which gets the best result in this task;
- We analyze and verify that the role played by objects in the scene is not very big which is a very novel idea in the social relationship detection task.

2 Related Work

In this section we introduce the related work on social relationship understanding and message passing.

2.1 Social Relationship Understanding

The foundation of social network is the social relationships understanding, an important multidisciplinary problem that has attracted increasing attention in computer vision recently. A much number of studies that aim to infer social relationships from images [Li *et al.*, 2017; Wang *et al.*, 2018; 2010; Zhang *et al.*, 2015b] and videos [Ding and Yilmaz, 2010; Ramanathan *et al.*, 2013; Vinciarelli *et al.*, 2009] have been made since the rise of deep learning. For instance, motivated by psychological studies, [Zhang *et al.*, 2015b] and [Dibeklioglu *et al.*, 2013] exploit social relationships based on facial attributes such as expression and head pose, and affective behaviour analysis. Besides, [Li *et al.*, 2017] introduce the contextual regions that produced by Faster RCNN[Ren *et al.*, 2015] at first. [Li *et al.*, 2017] proposed a dual-glance model for social relationships, where the first glance makes a coarse relationship prediction for a given person pair and then the second one refines the prediction by using the regions around the pair; [Wang *et al.*, 2018] constructed a semantic-aware knowledge graph by detecting objects, and employed Gated Graph Neural Network (GGNN) [Li *et al.*, 2015] to integrate the graph into the Graph Reasoning Model (GRM), a graph reasoning network where a proper message propagation and graph attention mechanism are introduced to explore the interaction between person pair and the contextual objects.

Unlike the aforementioned works which mainly focus on facial attributes, neighboring regions or contextual objects, we studied the two classic datasets PISC [Li *et al.*, 2017] and PIPA-relation [Sun *et al.*, 2017] in Sec. 4.2 and found the social relationships in a still image are always stable. Based on this discovery, we designed a novel end-to-end trainable Person-Pair Relation Network (PPRN), a graph inference network to capture this semantic relevance cues via message passing among person pair nodes.

2.2 Message Passing

Graph inference is a kind of form of message passing and Conditional Random Fields (CRF) have been used extensively in this field. Johnson *et al.* used CRF to infer scene graph grounding distributions for image retrieval [Johnson *et al.*, 2015]. Yatskar *et al.* use a deep CRF model to propose situation-driven object and action prediction[Yatskar *et al.*, 2016] . Danfei Xu *et al.* use the GRU-RNNs to solve the scene graph generation problem iteratively[Xu *et al.*, 2017]. Our work is related to Graph-LSTM [Liang *et al.*, 2016] and the work of Danfei Xu *et al.*[Xu *et al.*, 2017] which formulate the message passing problem using RNN models. Danfei Xu *et al.*[Xu *et al.*, 2017] design primal graph and dual graph in their model while we just simplify the model and just use one graph to make social relationship messages to pool and achieve a better result. As what Danfei Xu *et al.*[Xu *et al.*, 2017] have done, our model iteratively refines the social relationship predictions through relationship message passing in the scene, whereas the Structural RNN model only makes one-time predictions along the temporal dimension, and thus cannot refine its past predictions[Xu *et al.*, 2017].

3 PPRN Model

In this section, we introduce the proposed PPRN for social relationships understanding. We formulate the relationships of an image as a *social graph*, each node denotes relationship of a person pair. Then, taking the same approach as

[1] Due to space limit, omitted data, code, and supporting materials are provided in the online appendix (https://tinyurl.com/IJCAI19-229).

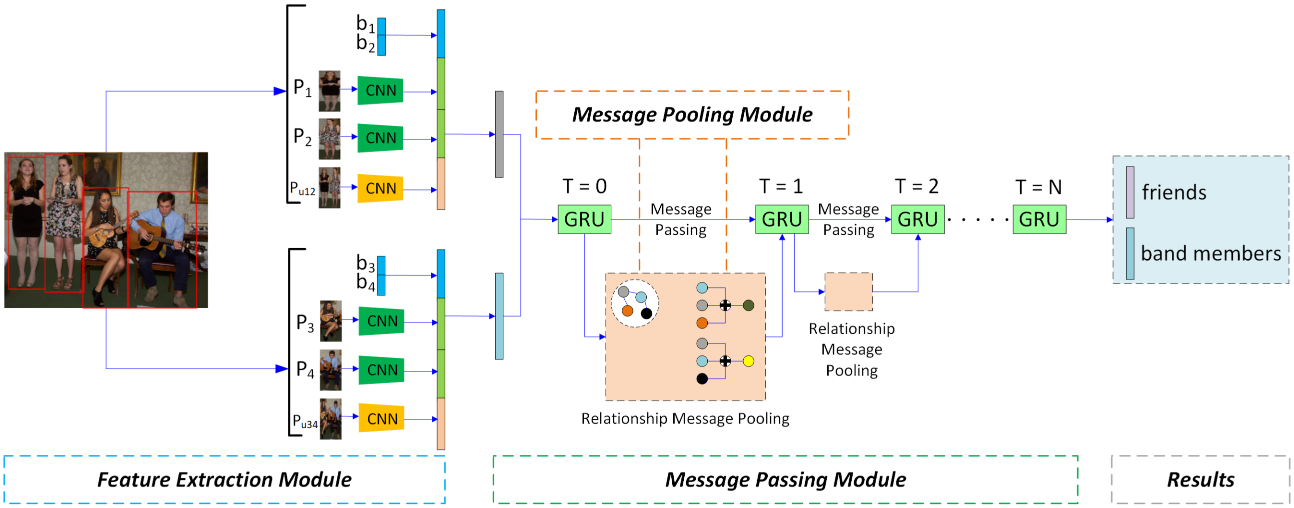

Figure 2: An illustration of our PPRN. The model first extracts the features of each person and person pairs in the Feature Extraction Module. P_i denotes the i-th person and P_{uij} denotes the union region of i-th person and j-th person. b_i denotes the position of the i-th person. In the Message Pooling Module, a message pooling function computes relationship messages that are passed to the node GRU in the next iteration from the hidden states. The $\oplus$ symbol denotes a learnt weighted sum. Then in the Message Passing Module, we iteratively updates the hidden states of the GRUs. We use the hidden states of the GRUs at the last iteration step, to predict the social relationships in the scene.

GRM[Wang *et al.*, 2018] to extract various features initialize these nodes. Person pair relation network employs Gated Recurrent Unit[Cho *et al.*, 2014] to explore the interaction of relationships with each other, and we employ the attention mechanism to adaptively exploit the most relevant nodes. It integrates multiple component to learn the representations of relationship between peoples in an image. The proposed framework is shown in Fig. 2.

3.1 Feature Extraction Module

Given an image $\mathbf{I}$ and the bounding box of peoples, we first crop three patches for each person pair, where the first two cover each person, $\mathbf{p_1}$ and $\mathbf{p_2}$, and one for union region, $\mathbf{p_u}$, that cover the both people and maintains the basic information for recognition. These patches are resized to 224×224 pixels and fed into three CNNs, These feature vectors from the last convolutional layer are flattened and concatenated. $\mathbf{p_1}$ and $\mathbf{p_2}$ share the same weights. In addition, geometry feature of bounding box is complementary to the visual appearance, as the relation *No Relation*, which is not easy to learn only from visual feature. We denote the geometry feature of bounding box i as $b_i^{pos} = \{x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, area_i\} \in \mathbf{R}^5$, where all the parameters are relative values. These features also concatenated with the CNN features for p_1, p_2 and p_u to form a single vector. Finally, both of them are fed into a fully connected layer to produce a 4096-dimension feature vector v_h.

3.2 Message Passing Module

In this subsection, we will talk about how the social relationships message will pass to each other. In this Message Passing Module, we use the Recurrent Neural Networks(RNNs) to finish the inference of the social relationship detection task. In contrast to Zheng *et al.* [Zheng *et al.*, 2015], our model use a generic RNN module to compute the hidden states. In addition, we select the best cell Gated Recurrent Units [Cho *et al.*, 2014] for RNNs. We use the hidden state of the t-th step to denote the relationship nodes of the current social graph, and the rest hidden state inheritance in the hidden state from the previous step. So the relationship node will be updated step by step as the GRU-based RNN going. Besides, the input of the GRU for each step comes from the output of the Message Pooling Module which has completed a social relationship interaction. In particular, the feature vector v_h which comes from Feature Extraction Module will be fed into the first GRU. On the whole, the Message Passing Module can be formulated as follows:

$$
\begin{aligned}
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]), \\
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t]), \\
\hat{\boldsymbol{h}_t} &= tanh(\boldsymbol{W}[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t]) \\
\boldsymbol{h}_t &= (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \hat{\boldsymbol{h}}_t
\end{aligned}
\tag{1}
$$

where σ and $tanh$ are the logistic sigmoid and hyperbolic tangent functions. In addition, $\odot$ symbol denotes the element-wise multiplication operation. $\boldsymbol{r}_t$ denotes the reset gate in t step and $\boldsymbol{z}_t$ denotes the update gate at t step. $\boldsymbol{W}_r$ and $\boldsymbol{W}_z$ denotes weights of reset gate and update gate. $\boldsymbol{W}_r$, $\boldsymbol{W}_z$ and $\boldsymbol{W}$ are trainable parameters. Specially, at t-th step, each GRU takes the previous hidden state $\boldsymbol{h}_{t-1}$ and the coming messages $\boldsymbol{x}_t$ as input, and produce a new hidden state. By the way, $\boldsymbol{x}_t$ computed by Message Passing Module. Each node in *social graph* holds the inner state in the corresponding GRU unit. At last, we use the last hidden states of the GRU unit to represent the social relationship of each node and output the results.

3.3 Message pooling Module

Sec 3.2 provides a way to solve reasoning problem using RNNs. As each GRU receives multiple incoming messages, we need an aggregation function that merge information from all messages into a meaningful representation. Intuitively, the methods of standard pooling can do this, such as average pooling and max pooling. But, it will be more effective to exploit various contextual cues and only preserve the appropriate parts when understanding social graph of an image. So, we utilized a message pooling function that computes the weights factors for each incoming message and aggregate the the messages using a weights sum.

Formally, given the current GRU hidden states of node $\boldsymbol{h_i}$, we takes the messages from other nodes as $\boldsymbol{m_{i,j\to i}}$, $j \to i$ represents the nodes j passing message to node i, and $\boldsymbol{h_{j\to i}}$ as the hidden states of other nodes in the social graph. $\boldsymbol{m_{i,j\to i}}$ is computed by function of its hidden state $\boldsymbol{h_i}$. To be more specific, $\boldsymbol{m_{i,j\to i}}$ are computed by the following message pooling functions:

$$\boldsymbol{m_{i,j\to i}} = \sum_j \sigma(\boldsymbol{w}^T[h_i, h_{j\to i}])h_{j\to i} \tag{2}$$

where [.] represents the operation in the concatenation of vectors. and $\boldsymbol{\sigma}$ denoted a sigmoid function. $\boldsymbol{w}$ is the parameter to be learn. Specifically, $m_{i,j\to i}$ will be the next GRU unit inputs that denotes as x_t in equation 1

3.4 Optimization

Given the predicted score $\mathbf{s}^{I,k} \in R^{|\mathcal{C}|}$ for the k-th person pair in image I, we use *softmax* function to get its corresponding probability $\mathbf{p}^{I,k} \in R^{|\mathcal{C}|}$

$$p_i^{I,k} = \frac{\exp s_i^{I,k}}{\sum_{j=1}^{|\mathcal{C}|} \exp s_j^{I,k}}, i = 1, 2, \ldots, |\mathcal{C}| \tag{3}$$

where $\mathcal{C}$ donates the classes set of social relationship and $|\mathcal{C}|$ is its size. The loss function is expressed as

$$\mathcal{L} = -\frac{1}{\sum_{I\in\mathcal{I}} \mathrm{N}(I)} \sum_{I\in\mathcal{I}} \sum_{k=1}^{\mathrm{N}(I)} \sum_{i=1}^{|\mathcal{C}|} \mathrm{L}(y_i^{I,k}, p_i^{I,k}) \tag{4}$$

where $\mathrm{N}(I)$ returns the number of person pair in image I, $\mathrm{L}(\cdot)$ is the cross entropy loss function, $\mathcal{I}$ is the image set.

4 Experiments

4.1 Experiment Setting

Datasets. In this work, two datasets are used to evaluate our proposed method. The first one is the large-scale People in Social Context (PISC) [Li *et al.*, 2017] with 22,670 images and contains two-level recognition tasks: **3 Coarse-level relationship**, namely *No Relation, Intimate Relation, None-Intimate Relation* and **6 Fine-level relationship**, i.e., *Friend, Family, Couple, Professional, Commercial, No Relation*. The second one is the People in Photo Album Relation (PIPA-Relation) [Sun *et al.*, 2017], an extension version of People in Photo Album (PIPA) [Zhang *et al.*, 2015a] with 37107 images. It also annotates 26,915 person pairs on two-level recognition tasks: **5 Social Domains** and **16 Social Relations** based on these domains. The train/val/test in PISC are 13,142/4,000/4,000 images with 14,536/25,636/15,497 person pairs on coarse level relationship, and 16,828/500/1,250 images with 55,400/1,505/3,691 person pairs on fine level relationship, respectively. In PIPA-relation, we follow [Wang *et al.*, 2018] and focus on recognizing its 16 relationships in the experiment. The train/val/test in it are 13,729/709/5,106 person pairs.

Implementation Details. During our work, We adopt the same strategy as previous works including [Li *et al.*, 2017] and [Wang *et al.*, 2018]. First, we fine-tune the ResNet-101 model [He *et al.*, 2016] , and we set a lower learning rate as 0.0001. After that, we first freeze the feature extraction module, , and for the message passing progation model, the dimension of hidden size is set as 512. The iteration time of RNNs T is set as 4 and learning rate as 0.0001. Similar to [Wang *et al.*, 2018], the fine-tuning model utilized SGD, and the all com message passing module is trained with ADAM.

4.2 Datasets Analysis

Table 1: Statistic on PISC and PIPA-relation. "Sui." denotes that the percentage of person pair in an image with multiple person pairs. while "Unsui." is its opposite. "Single Rel." represents proportion of the images that have only one category of social relationship, while "Multi Rel." is its opposite.

Dataset		Sui.	Unsui.	Single Rel.	Multi Rel.
PISC	coarse	87.1%	12.9%	79.9%	20.1%
	fine	83.9%	16.1%	86.4%	13.6%
PIPA-relation 16		71.9%	28.1%	94.9%	5.1%

In this subsection, we made an analysis on PISC and PIPA-relation. As shown in table 1, we first counted the relationship categories of each image. Specifically, on PISC coarse, 79.9% of images have only one coarse category, while 20.1% images have multiple coarse one. Second, 87.1% and 83.9% person pairs are the proportion that suitable for interaction in PISC coarse and PISC fine datasets.

4.3 Comparisons with State-of-the-Art Methods

We compare our proposed model with the existing state-of-the-art methods on both PISC and PIPA-Relation datasets. Formally, the compared methods are as followed:

Performance On the PISC dataset

UnionCNN Following [Lu *et al.*, 2016], it generates a single CNN model to predicate relations. In this task, we also feeds the union region of person pair to a single CNN for classification.

Pair CNN[Li *et al.*, 2017] consists of two equivalent CNNs with shared weights to extracted features for cropped image for two individuals.

Pair CNN + BBox + Union[Li *et al.*, 2017] incorporates spatial location information of two bounding box that based the previous Pair CNN and Union CNN.

Dual-glance[Li *et al.*, 2017] implements coarse and fine prediction which includes three and six relationships. Dual-glance employing Pair CNN + BBox + BBox + Union and utilized surrounding region proposal to refine the prediction.

Table 2: Recall-per-class and mean average precision (mAP) evaluating our PPRN model and previous methods on PISC (in %).

Methods	Coarse relationships				Fine relationships						
	Intimate	Non-Intimate	No Relation	mAP	Friends	Family	Couple	Professional	Commercial	No Relation	mAP
Union CNN [Lu *et al.*, 2016]	72.1	81.8	19.2	58.4	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair CNN [Li *et al.*, 2017]	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Pair CNN + BBox + Union [Li *et al.*, 2017]	71.1	81.2	57.9	72.2	32.5	62.1	73.9	61.4	46.0	52.1	56.9
Pair CNN + BBox + Global [Li *et al.*, 2017]	70.5	80.0	53.7	70.5	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-glance [Li *et al.*, 2017]	73.1	**84.2**	59.6	79.7	35.4	**68.1**	76.3	70.3	57.6	60.9	63.2
GRM [Wang *et al.*, 2018]	81.7	73.4	65.5	**82.8**	59.6	64.4	**58.6**	76.6	39.5	67.7	68.7
Ours	**81.9**	67.3	**74.7**	81.8	**61.0**	67.1	56.2	**76.9**	**46.0**	**68.1**	**69.7**

Table 3: Accuracy (in %) evaluating our PPRN model and previous methods on PIPA-relation.

Methods	accuracy
Two stream CNN [Zhang *et al.*, 2015a]	57.2
Dual-Glance [Li *et al.*, 2017]	59.6
GRM [Wang *et al.*, 2018]	62.3
Ours	**64.7**

GRM[Wang *et al.*, 2018] propose a graph reasoning model that unifies the frequency of co-concurrences of each relationship-object pair to facilitate social relationship Understanding.

Similar to the model of GRM, we also adopt the per-class recall and mean average precision(mAP) to evaluate our model. The experiments data are reported in Table 1. First, both Pair CNN + BBox + Global, Dual-glance and GRM are incur extra Faster-RCNN[Ren *et al.*, 2015] to extract the local contextual cues(object proposal). GRM utilized the object proposal to construct a semantic-aware knowledge graph for reason about the social relationship. It is notable that both of them incur extra detection annotations that contains noisies. Specifically, our model achieves an accuracy of 75.1% and mAP of 81.8% for the coarse-level recognition. the model also takes an accuracy of 65.6% and mAP of 69.7% for the fine-level dataset, our model beating previous best model in the fine-level recognition ,but slightly lower on coarse-level than the best model before.

Performance On the PIPA-Relation Dataset

On this dataset, we also compare our proposed model with the existing methods,i.e, Two stream CNN[Sun *et al.*, 2017],Dual-glance[Li *et al.*, 2017] and GRM[Wang *et al.*, 2018] that achieves the best performance before. Specifically, we directly reprint the experimental of several baselines from the literature. The result are presented in Table 2. Notably, our PPRN significantly outperforms previous methods. Still, our model outperforms all the baselines in PIPA-relation, and beating the best of them 2.4%.

Table 4: The mAP and accuracy result of RCNN, our model and our model with contextual region that implements in the same way as dual-glance (in %)

Methods	PISC coarse		PISC fine	
	accuracy	mAP	accuracy	mAP
RCNN	-	63.5	-	48.4
Ours(max)	74.3	80.8	64.1	68.3
Ours(average)	74.6	80.1	63.8	68.3
Ours(atten)	**75.1**	**81.8**	**65.7**	**69.7**
Ours(atten) + objects region	74.9	81.2	65.3	69.1

4.4 Analysis On Experimental Result

In this section, we first present the comparison result of message passing mechanism and analysis the reasons behind, and the result present in Table 3. Then, we conduct a conditional experiment to investigate the effectiveness of the factor of contextual object regions and the interaction between person-pair.

Significance Of Message Passing

In our framework, the core component is the introduction of message passing mechanism, and one of the key component is the message pooling functions that use the learnt weights sum to aggregate hidden states of other relationship nodes into message. To further investigate the improvement of our approach on recognizing social relationships, we evaluate variants of our model with standard pooling methods. The first is to use average pooling (avg. pool) instead of the learnt weighted sum to aggregate the hidden states. The second is similar to the first one, but uses max pooling (max pool).

Analysis Of Contextual Information

First, our model is to process the cue of contextual relationships without contextual object region that incurred in Dual-glance and GRM by extra detection annotations. From the object regions point of view, the effect of object regions are provide the information of the scene to constrain the result of

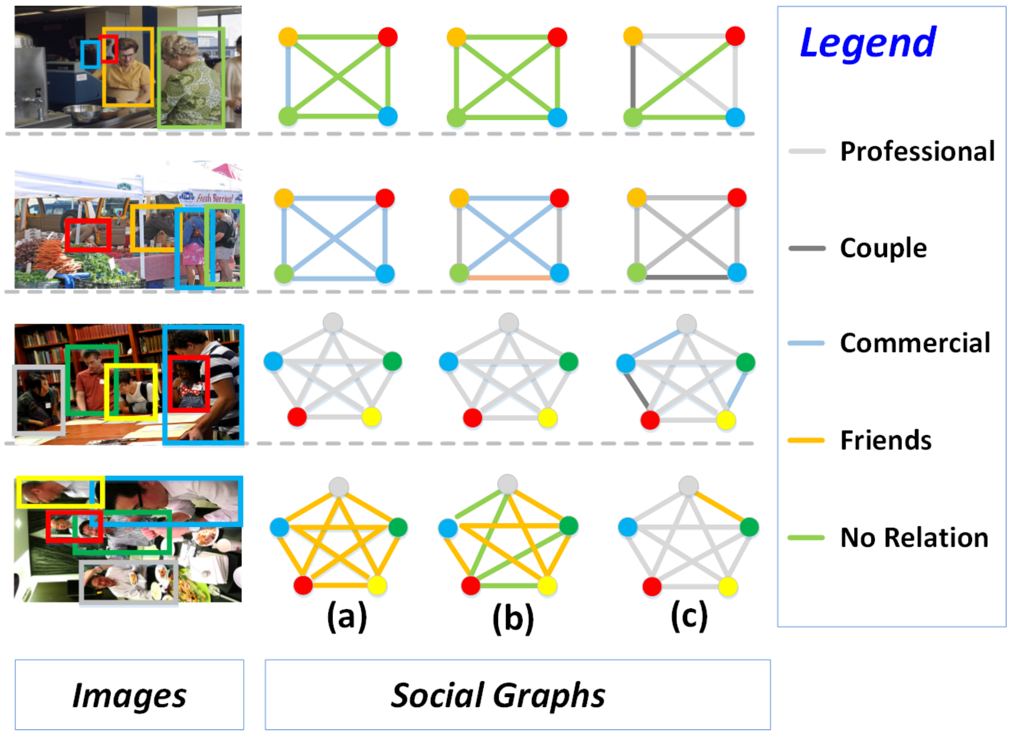

Figure 3: Comparison of social relation graphs from different sources: (a) PISC fine, (b) our PPRN and (c) GRM. People with boxes of various colors on the original image correspond to the same color node on right of the image. An edge linked to two nodes represents the social relationships between them and its meaning can be found in the right legend. From these example, all edges in (a) are almost identical, which suggests that the social relationships in a still image are always stable.

the relationship and the information of scene is single. Contextual object region provide information on the premise that they can be correctly identified, and the object regions are always wrong due to the defects of current object detection model and the complexity of image scene. what's more, if the object of laptop is useful to classify the relationship as professional while not contribute to the relationship of friend, and there exists a scene with the two relationships. But, the relationship of professional will helpful to exclude the relationships such as family or couple.eg . As table 3 reported, RCNN, which uses only object information, has the worst effect. [Li *et al.*, 2017] trained a model with attention mechanisms to exploit different object regions cues according different pairs of people. we also conduct a attention mechanisms as same as Dual-glance, but the performance do not improved as reported in Table 3. One possible reason is that the object regions cues is covered by contextual information of relationships cues, and the result of experiments proved the analysis before.

4.5 Case Study

Four examples in Figure 3 are shown to illustrate the ability of our PPRN to infer social relationships. We compared two predicted social relationships, one from our PPRN and the another from GRM [Wang *et al.*, 2018]. We found that compared to the social relation graphs in (c), the graphs in (b) are very similar to the graphs in (a), which means that our PPRN performs better than GRM. In addition, similar to (a), over half of the edges in (b) are identical, which strongly suggests that the contextual relationship cues are very significant for social relationship understanding and our PPRN can fully utilize it. Considering the first example, the true social relation between the person in an orange box and another in a green one is No Relation, which can be correctly predicted by our PPRN while being incorrectly predicted as Couple by GRM, and the accuracy of PPRN is 100% while 33.3% in GRM.

5 Conclusion

In this study, we propose a Person-Pair Relation Network(PPRN) that aims to solve social relationships recognition of an image. The proposed model incorporates the information of contextual relationships. The key challenge is to design a model for interaction between social relationships. PPRN consists of a reasoning module that propagates relationship message through the RNNs. In this way, it improves the quality of relationships prediction. Specifically, a attention also utilized to compute the weights factor for the connected nodes in a social graph, and these weights introduced to aggregates the messages. We also analysis the influence of contextual relationships and contextual object regions, and we found that the cue of contextual object region covered by contextual relationships. Extensive experiments on two large-scale benchmarks(PISC and PIPA-Relation) achieves better performance without incur extra detection annotations.

References

[Barr *et al.*, 2014] Jeremiah R. Barr, Leonardo A. Cament, Kevin W. Bowyer, and Patrick J. Flynn. Active clustering with ensembles for social structure extraction. In *IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, March 24-26, 2014*, pages 969–976, 2014.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111, 2014.

[Dibeklioglu *et al.*, 2013] Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *Proceedings of ICCV*, pages 1497–1504, 2013.

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

[Fairclough, 2003] Norman Fairclough. Analysing discourse : Textual analysis for social research / n. fairclough. 01 2003.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016.

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

[Li *et al.*, 2015] Li-Jia Li, David A. Shamma, Xiangnan Kong, Sina Jafarpour, Roelof van Zwol, and Xuanhui Wang. Celebritynet: A social network constructed from large-scale online celebrity images. *TOMCCAP*, 12(1):3:1–3:22, 2015.

[Li *et al.*, 2017] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of ICCV*, pages 2669–2678, 2017.

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

[Lu *et al.*, 2016] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *Proceedings of ECCV*, pages 852–869, 2016.

[Ramanathan *et al.*, 2013] Vignesh Ramanathan, Bangpeng Yao, and Fei-Fei Li. Social role discovery in human events. In *Proceedings of CVPR*, pages 2475–2482, 2013.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of NIPS*, pages 91–99, 2015.

[Sun *et al.*, 2017] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of CVPR*, pages 435–444, 2017.

[Vinciarelli *et al.*, 2009] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Comput.*, 27(12):1743–1759, 2009.

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

[Wang *et al.*, 2018] 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 *Proceedings of IJCAI*, pages 1021–1028, 2018.

[Xu *et al.*, 2017] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of CVPR*, pages 3097–3106, 2017.

[Yatskar *et al.*, 2016] Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of CVPR*, pages 5534–5542, 2016.

[Zhang *et al.*, 2015a] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir D. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of CVPR*, pages 4804–4813, 2015.

[Zhang *et al.*, 2015b] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of ICCV*, pages 3631–3639, 2015.

[Zheng *et al.*, 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of ICCV*, pages 1529–1537, 2015.