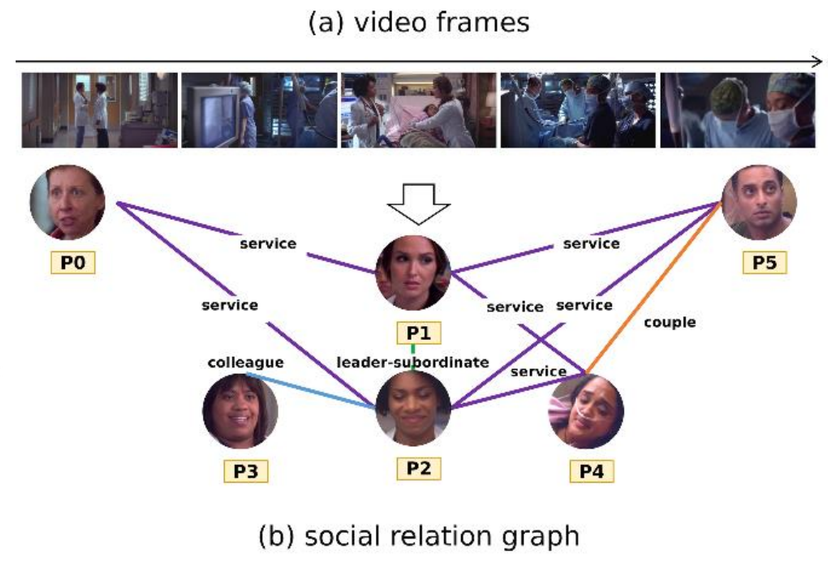
CAGNet: A Context-Aware Graph Neural Network for Social Relation Detection on Videos

Fan Yu, Yaqun Fang, Zhixiang Zhao, Jia Bei✉, Tongwei Ren, Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University

Introduction

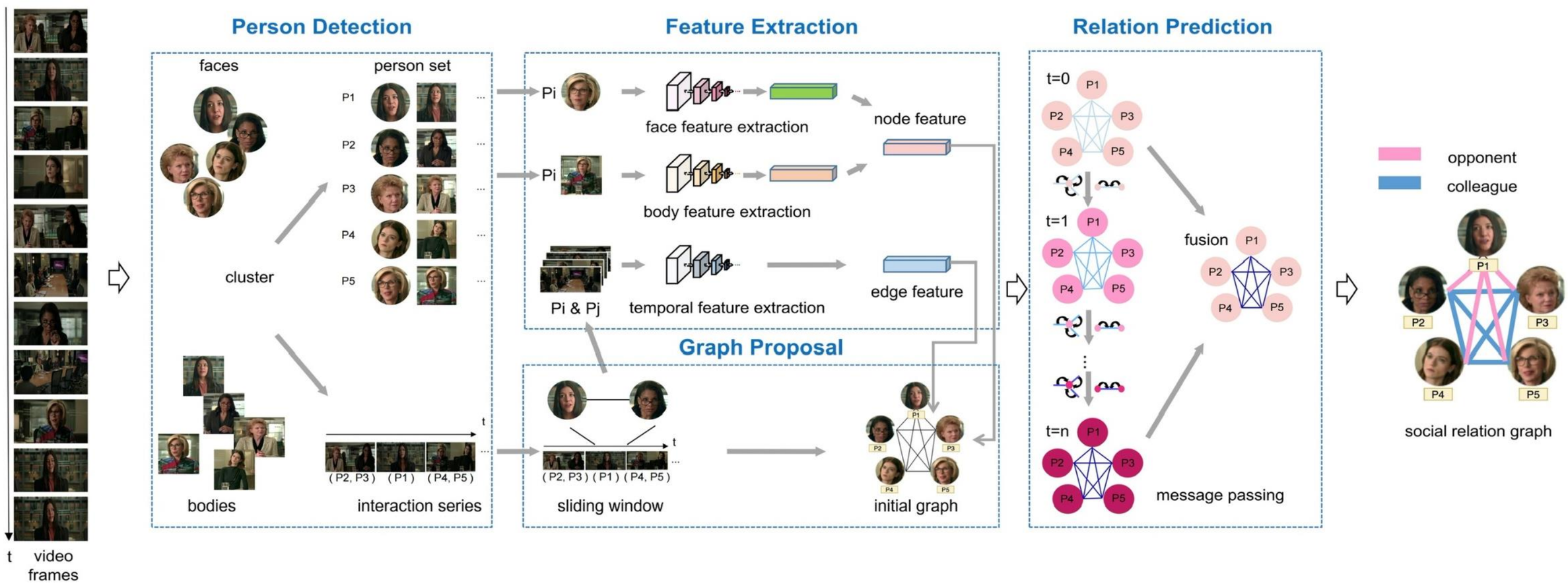
In this paper, we analogize social relation detection to scene graph generation and we call it **Video Social Relation Graph Generation (VSRGG)**, which involves generating a social relation graph for each video based on person-level relations. We propose a **Context-Aware Graph Neural Network (CAGNet)** to solve VSRGG, which effectively generates social relation graphs through message passing, capturing the context of the video. Additionally, we construct **a more challenging dataset, VidSoR**, to evaluate VSRGG, which contains 72 hours of videos with 6,276 person instances and 5,313 relation instances of eight relation types.



Method

We propose a novel social relation graph generation method, named **CAGNet**, which captures context information by a **graph message passing mechanism**.

- **Person Detection**: find main characters in the videos and detect both faces and bodies
- **Graph Proposal**: propose the potential edges to construct a sparse social relation graph
- **Feature Extraction**: face and body features are extracted and fused to generate node features and temporal features are extracted from key frames of person pair as edge features
- **Relation Prediction**: graph message passing mechanism is introduced, in which vertex feature and edge feature are iteratively updated with message from adjacent vertices and edges.



Experiments

Datasets: We construct **a more challenging dataset, VidSoR**, which contains 72 hours of videos from more than 300 different TV dramas.

Dataset	VidSoR	MovieGraphs
source	300+ TV dramas	51 movies
valid clips	1798	1551
avg. duration	2 min 24 s	44s
relation instances	5313	2329
avg. relation instances	2.95	1.50
relation types	8	106

Relation	Description
colleague	co-workers, classmates
couple	husband-wife, lovers
friend	friends
couple	husband-wife, lovers
leader-subordinate	boss-employee
opponent	enemy
parent-offspring	grandparent-grandchild, parent-child
service	waiter-customer
sibling	brothers, sisters

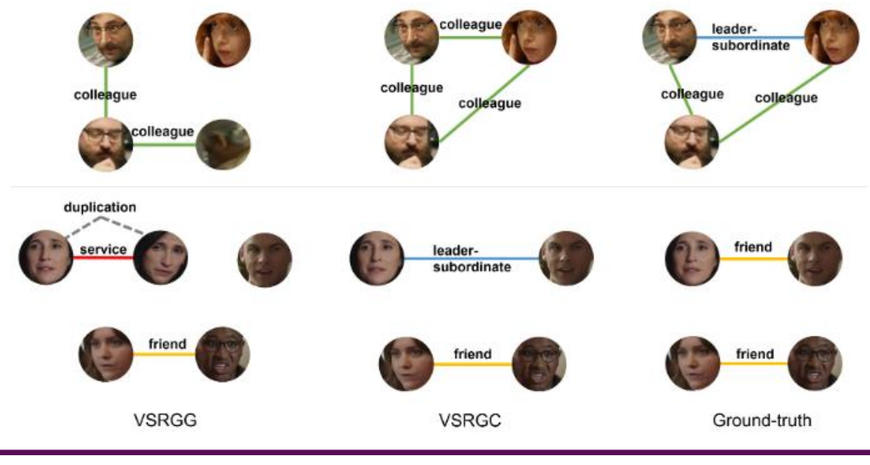
Evaluation setting

- **VSRGG**: only videos as input and output social relation graph among the main characters.
- **VSRGC**: take both videos and main characters with a collection of faces as input and generate social relation graph

Comparison with SOTA: We compare our CAGNet with different methods related to VSRGG, including methods for social relation recognition in images, social relation recognition in video-level, video social relation detection in person-level, along with methods for video scene graph generation. CAGNet can make **accurate predictions with a comparably high mRecall** in the case of only using visual features.

Ablation Study

- facial features + body features
- edge representations after message passing + vertex representations before message passing
- sliding window width 2



	VSRGG		VSRGC	
	mAP	mRecall	mAP	mRecall
UnionCNN [1]	7.09	1.10	17.01	11.54
PairCNN [25]	7.59	1.08	16.98	11.10
First-Glance [25]	9.54	1.56	16.93	11.45
Dual-Glance [25]	8.22	1.45	15.92	11.68
GRM [27]	5.88	2.23	16.54	10.99
SRGGN [29]	9.68	2.14	**17.49**	11.12
GSTEG [49]	8.4	2.22	9.56	4.15
STTran [50]	**10.76**	3.28	12.75	9.45
TRACE [51]	8.82	3.73	10.31	10.30
MSRT [4]	3.68	2.46	4.22	6.47
LIREC [5]	2.07	**7.98**	10.49	**13.61**
CAGNet(Ours)	**10.04**	**7.05**	**17.73**	**12.74**

	mAP	Recall
Face	14.05	11.58
Body	13.14	6.48
Body + Face	17.73	12.74

	mAP	Recall
ε_0	16.55	15.11
ε_n	15.13	14.67
$V_0 + \varepsilon_0$	16.13	15.51
$V_n + \varepsilon_n$	15.76	13.9
$V_n + \varepsilon_0$	13.93	15.27
$V_0 + \varepsilon_n$	17.73	12.74

南京大学 NANJING UNIVERSITY

yf@smail.nju.edu.cn

MAGUS MediA recoGnition and UnderStanding