ACMMULTIMEDIA 2019
21 - 25 October 2019, Nice, France

GC-29

MAGUS
MediA recoGnition and UnderStanding

Video Visual Relation Detection via Multi-modal Feature Fusion

Xu Sun, Tongwei Ren, Zi Yuan, Gangshan Wu

Introduction

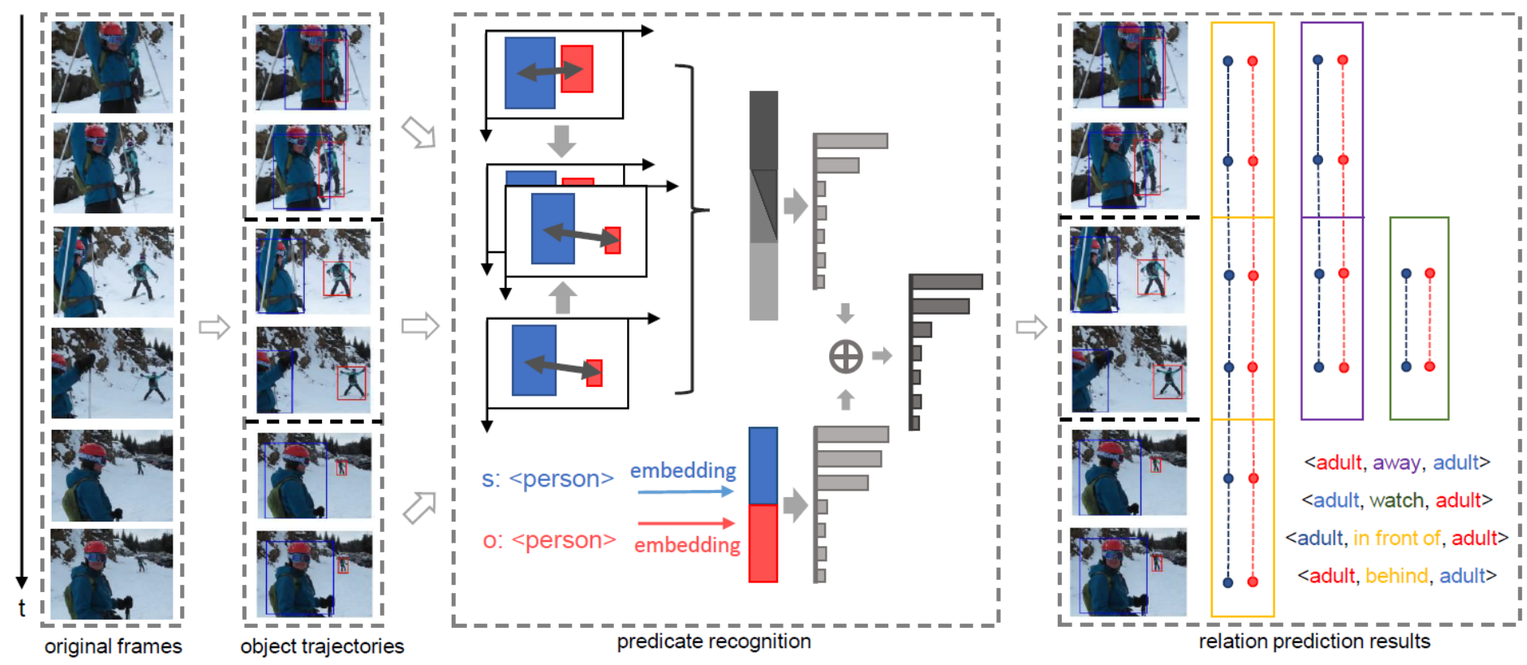
Video visual relation detection aims to capture dynamic interactions between co-occurrent objects in video with object trajectory pair and **⟨subject, predicate, object⟩** triplet.

Contribution: we explicitly combine **spatial-temporal feature** and **language context feature** with the assistance of **object trajectory detection**, and win **the first place** in the Visual Relation Detection task of Relation Understanding in Videos (VRU) Challenge.

Method

We propose a novel video visual relation detection method, which consists of two components:

- **object trajectory detection:** detect objects densely on individual frames with FGFA, generate short-term trajectories by associating the bounding boxes on individual frames with Seq-NMS, filter out the extremely short ones, and associate the short-term trajectories into complete ones with KCF tracker.
- **relation instance generation:** break the co-occurrent part into segments for each trajectory pair, predict predicates for the segments by fusing spatial-temporal feature and language context feature, and associate the segments with the same triplet predictions greedily.



Experiments

Dataset

The dataset consists of 10,000 videos on 80 object categories and 50 predicate categories.

The dataset is divided into three parts: 7,000 for training, 835 for validation, and 2,165 for final testing.

The average length of the videos in VidOR is 35.73 seconds.

Evaluation metrics

VRU official metric:
mAP, tagging precision@5

Additional metrics:
Recall@50, Recall@100, tagging precision@1

Comparison with the state-of-the-arts

Our method **is superior to** the method comes second and the baselines constructed with the state-of-the-art visual relation detection methods on image (OTD+CAI) and video (OTD+GSTEG) respectively.

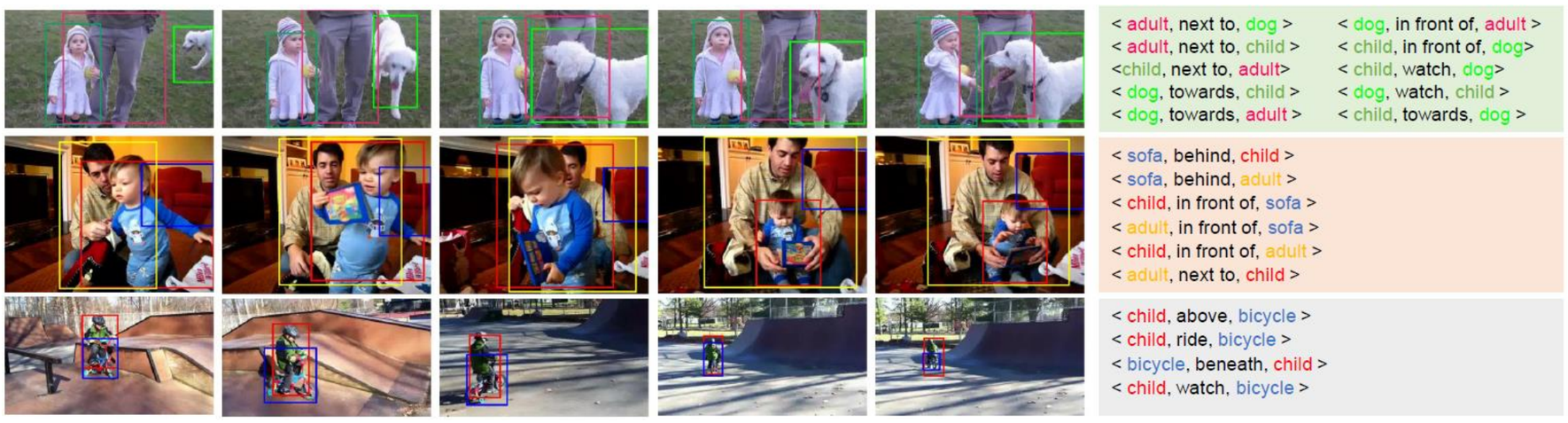
validation set

method	tagging precision@1	tagging precision@5	Recall@50	Recall@100	mAP
OTD+CAI	48.31	38.49	6.19	8.16	5.65
OTD+GSTEG	51.20	37.26	6.40	8.43	5.58
Ours	**51.20**	**40.73**	**6.89**	**8.83**	**6.56**

test set

method	tagging precision@5	mAP
RELAbuilder	23.60	0.546
Ours	**42.10**	**6.310**

Qualitative results



Xu Sun: sunx@smail.nju.edu.cn Tongwei Ren: rentw@nju.edu.cn