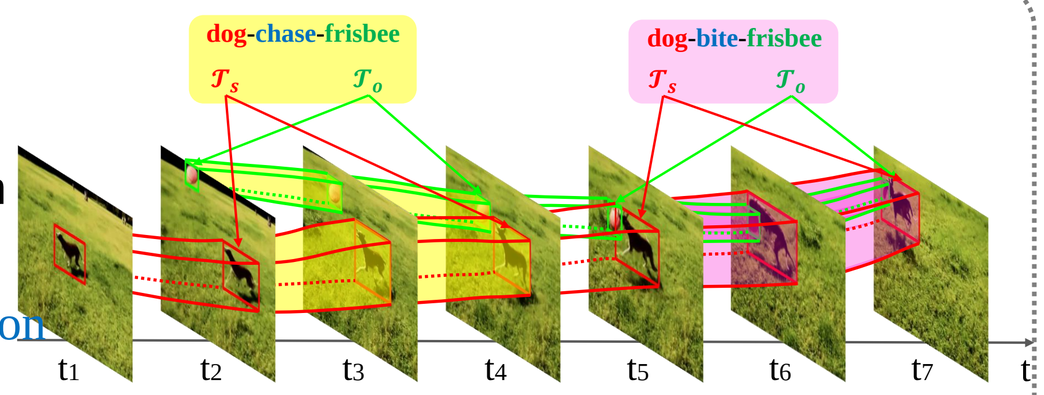
Video Visual Relation Detection

Xindi Shang[1], Tongwei Ren[2], Jingfan Guo[2], Hanwang Zhang[3], Tat-Seng Chua[1]

1. School of Computing, National University of Singapore, Singapore
2. State Key Lab for Novel Software Technology, Nanjing University, China
3. Department of Computer Science, Columbia University, USA

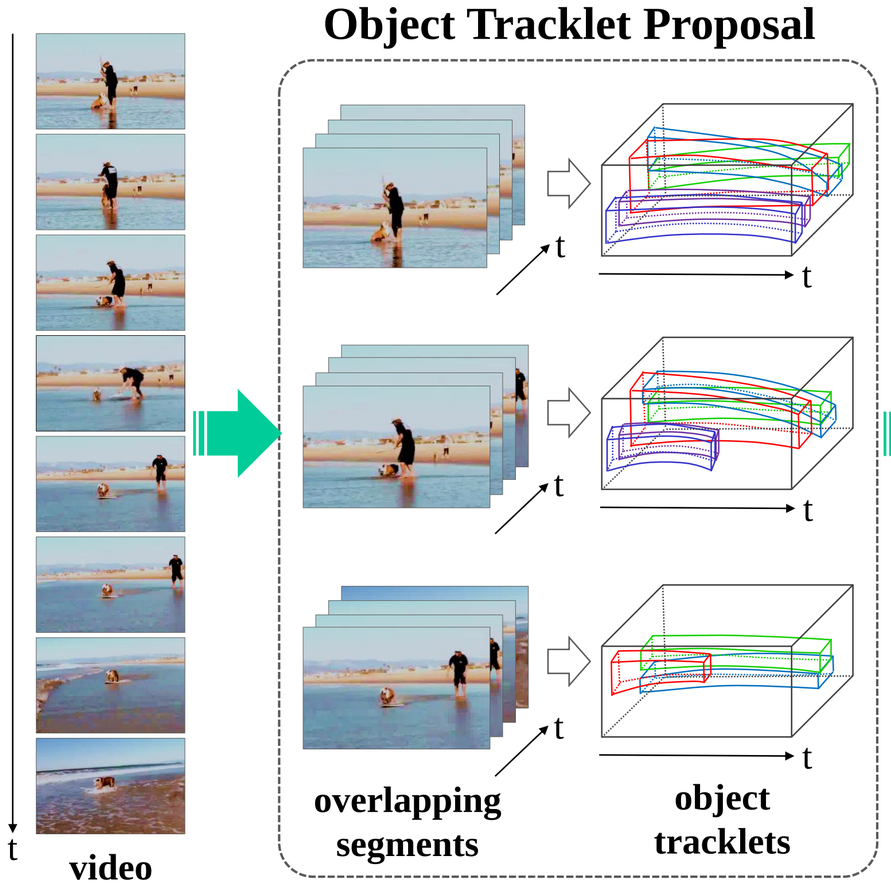
Introduction

Relation understanding is crucial for holistic understanding of video data, which can support complicated applications like automatic captioning and multimedia question answering. In this work, we aim to detect visual relations between objects, including the dynamic relations and temporally changing relations. Formally, a visual relation instance in a video is represented by a relation triplet <subject, predicate, object> with the trajectories of the subject and object (see figure).

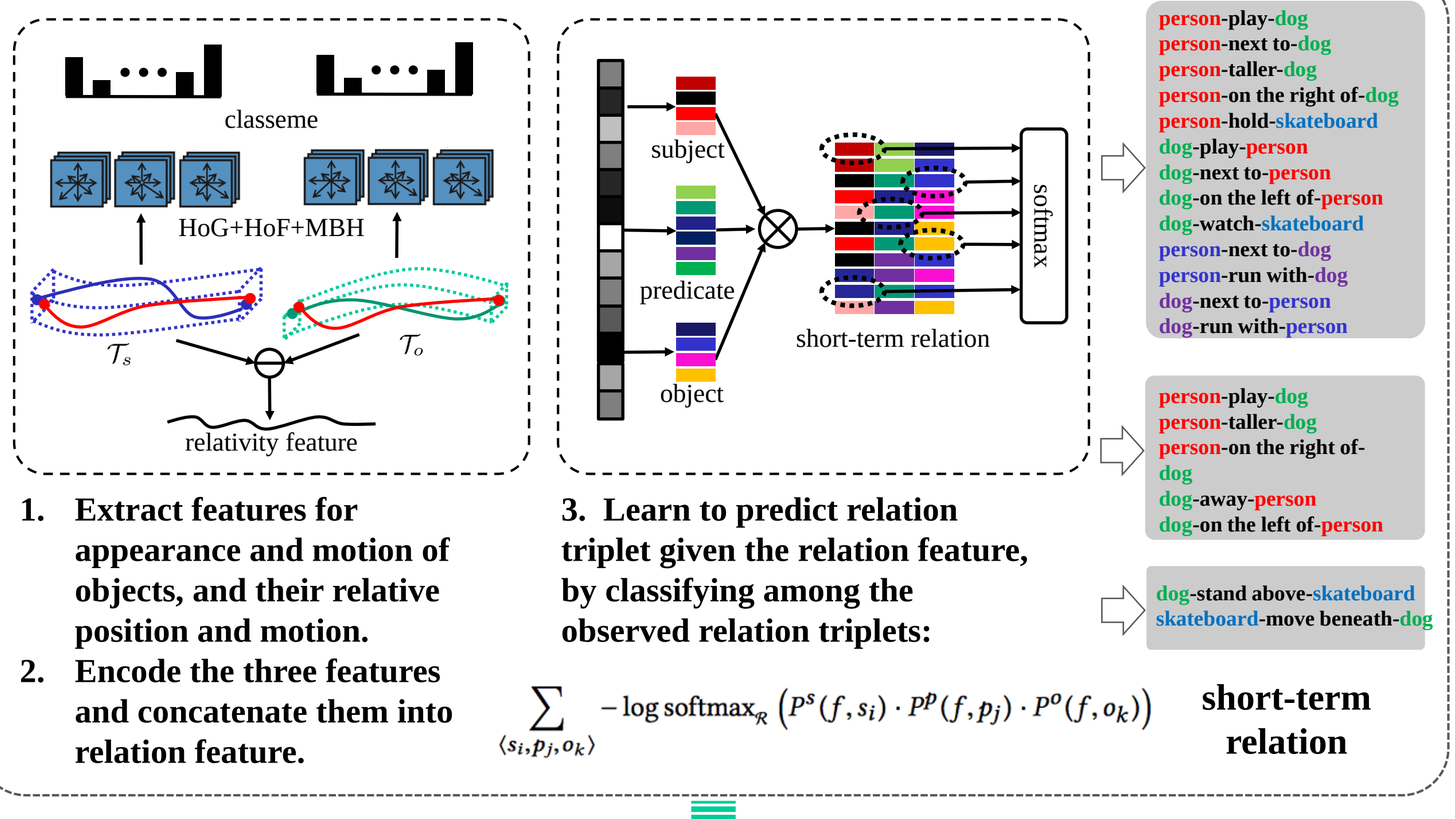


Definition of visual relation in videos

Object Tracklet Proposal



Relation Prediction



1. **Extract features for appearance and motion of objects, and their relative position and motion.**
2. **Encode the three features and concatenate them into relation feature.**
3. **Learn to predict relation triplet given the relation feature, by classifying among the observed relation triplets:**

$$\sum_{\langle s_i, p_j, o_k \rangle} -\log \text{softmax}_{\mathcal{R}} \left(P^s(f, s_i) \cdot P^p(f, p_j) \cdot P^o(f, o_k) \right)$$

short-term relation

Evaluation Results

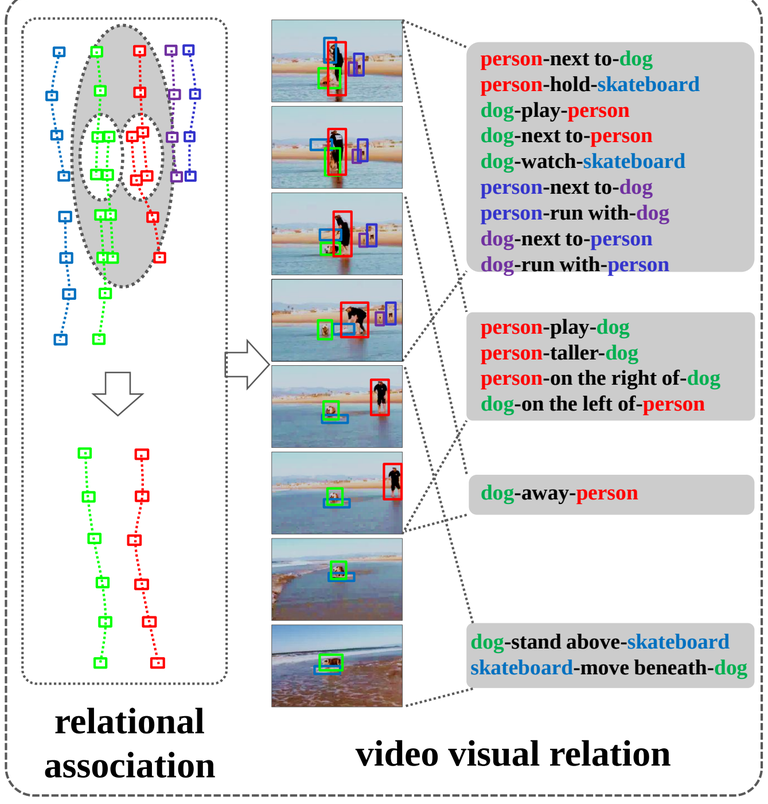
- Visual **relation tagging** evaluates the performance of relation prediction (no need to predict the object trajectories), which's useful for retrieval and visual QA.
- The **baselines** are adapted from image based methods by using our tracklet proposal, feature extraction and relational association methods.
- **Zero-shot** setting requires to predict unseen visual relations w.r.t. training

Method	relation detection			relation tagging		
	R@50	R@100	mAP	P@1	P@5	P@10
VP [32]	0.89	1.41	1.01	36.50	25.55	19.20
Lu's-V [22]	0.99	1.80	2.37	20.00	12.60	9.55
Lu's [22]	1.10	2.23	2.40	20.50	16.30	14.05
VTransE [42]	0.72	1.45	1.23	15.00	10.00	7.65
VidVRD	**5.54**	**6.37**	**8.58**	**43.00**	**28.90**	**20.80**

zero-shot setting

Method	relation detection		
	R@50	R@100	mAP
Lu's-V [22]	0.93	0.93	0.40
Lu's [22]	0.69	1.16	**0.47**
VTransE [42]	0.69	0.69	0.03
VidVRD	**1.62**	**2.08**	0.40

Greedy Relational Association



VidVRD Dataset

- The first video visual relation dataset, based on ILSVRC16-VID dataset.
- **Object trajectory annotation**: additional 5 object categories, *person, ball, sofa, skateboard and frisbee*.
- **Visual relation annotation**: 3,219 types of relation triplets in which 258 only appears in test set. The test set has 432 unseen instances.
- Download link: *https://lms.comp.nus.edu.sg/research/VidVRD.html*
- Loading and evaluation codes: *https://github.com/xdshang/VidVRD-helper*

	training set	test set
video	800	200
subject/object category	35	35
predicate category	132	132
relation triplet	2,961	1,011
visual relation instance (video-level)	-	4,835
segment	15,146	3,202
labeled segment	3,033	2,801
visual relation instance (segment-level)	25,917	29,714

A Joint Research Collaboration between NUS National University of Singapore, Tsinghua University, University of Southampton

Supported by National Research Foundation, Prime Minister's Office, Singapore

NExT++
NATIONAL UNIVERSITY OF SINGAPORE
13 Computing Drive , Singapore 117417