Joint learning of object and action detectors

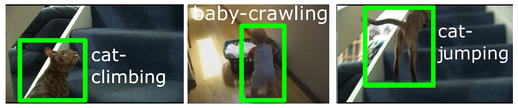
Vicky Kalogeiton[1,2], Philippe Weinzaepfel[3], Vittorio Ferrari[2] and Cordelia Schmid[1]

1 Inria
2 The University of Edinburgh
3 NAVER LABS EUROPE

ICCV17

Introduction

Joint object-action detection in videos



Related work

objects *or* actions

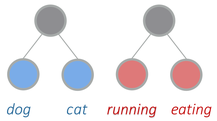
[Lu et al., ECCV16; Sadeghi et al., CVPR11; Xu et al., CVPR15,16]

[Gkioxari et al., CVPR15; Weinzaepfel et al., ICCV15; Peng and Schmid, ECCV16]

Contributions

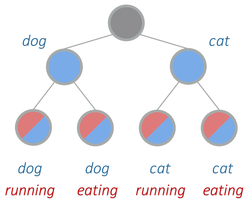
- End-to-end *multitask* architecture
- Enables *zero-shot learning, relationship detection* in images and object-action *semantic segmentation*
- *Fewer parameters* than alternatives:

Multitask



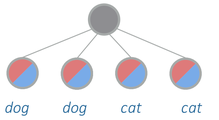
$\#params = (C_{in} + 1) \cdot (|\mathcal{O}| + |\mathcal{A}| + 2)$

Hierarchical



$\#params = (C_{in} + 1) \cdot (|\mathcal{O}| + |\mathcal{V}| + 1)$

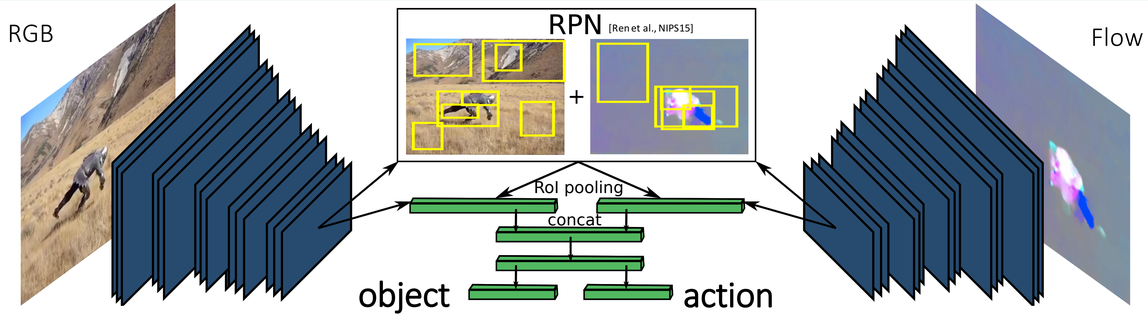
Cartesian



$\#params = (C_{in} + 1) \cdot (|\mathcal{V}| + 1)$

$\mathcal{O}$ object labels
$\mathcal{A}$ action labels
C_{in} input channels
$\mathcal{V} \subset \mathcal{O} \times \mathcal{A}$ valid object-action combinations

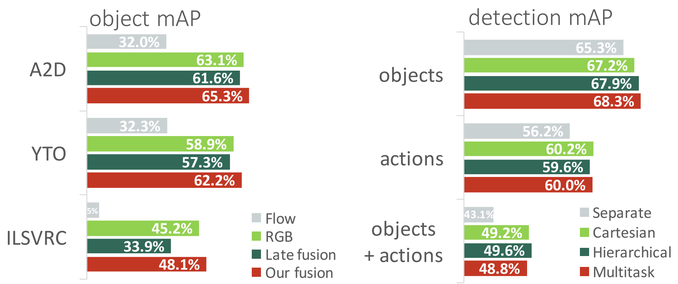
Joint object-action detection architecture



Joint object-action detection results

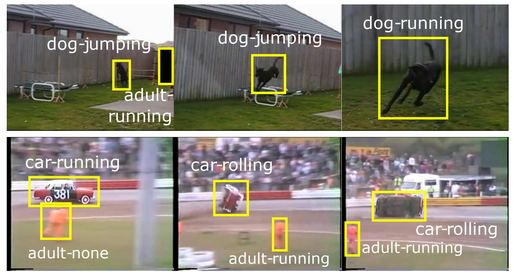
	A2D	YTO	ILSVRC VID
objects	✓	✓	✓
actions	✓	-	-
#annot	20k	6.5k	1.9M

End-to-end network



Joint learning

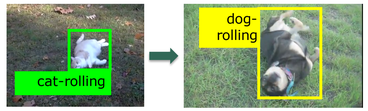
Qualitative results



✓ Better fusion ✓ Joint better than separate learning ✓ Less parameters

Zero-shot learning

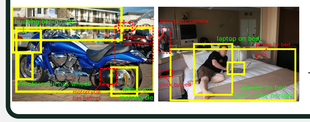
Remove action labels for one object category at training time but predict them at test time



✓ +5-15%mAP compared to chance level for object classes that share commonalities

Relationship detection in images

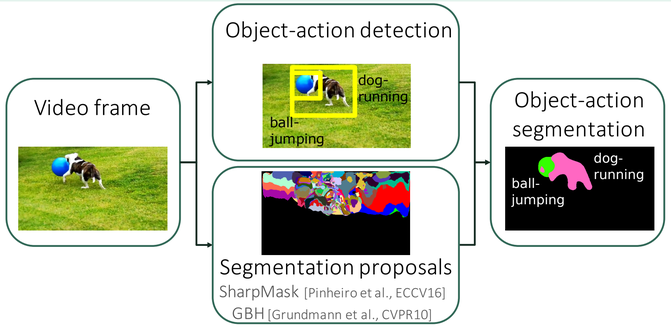
- VRD dataset: 38k object relationships in triplets: object1-interaction-object2
- Transform each triplet into two pairs: object and an interaction label



Phrase detection: zero-shot learning

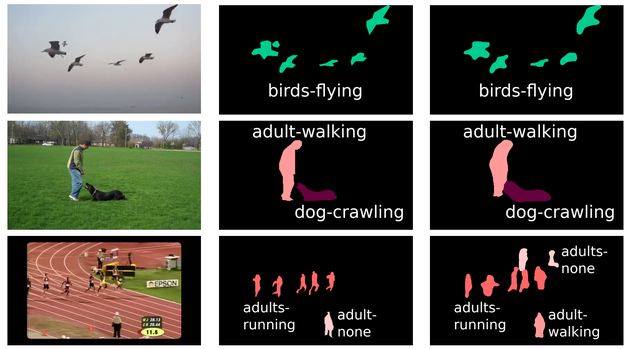
mod.	method	R@100	R@50
V	Lu et al., ECCV16	1.1	0.8
	Baseline	4.3	2.3
	Ours	5.5	3.4
V+L+F	Lu et al., ECCV16	3.8	3.4

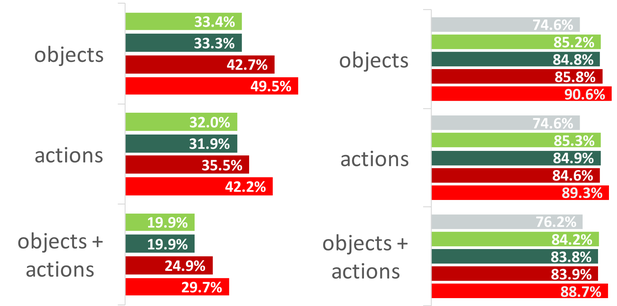
Object-action semantic segmentation



Qualitative results

Frame | Ground truth | Ours



Quantitative results

mean IoU segmentation accuracy | global pixel accuracy