

Self-Selective Context for Interaction Recognition

Mert Kilicay, Nourreddine Houssein, Efstratios Gavves, Arnold Smulders
QNVA Deep Vision Lab, University of Amsterdam
Netherlands
Email: kilicaym@gmail.com, {nhoussein, egavves, a.w.m.smulders}@uva.nl

Experiment 1

TABLE I: Self-Selection of The Single Context.

Context Feature	mAP(%)	Improvement $\Delta \uparrow$
Human-Object (HO)-only	62.50	-
HO + Body-part context	68.55	6.05
HO + Stuff context	68.20	5.70
HO + Surround context	68.44	5.94
HO + Deformation context	68.30	5.80

Problem: Human-object interaction recognition



Fig. 1: Human-object interactions come with many contexts that can help in recognition. In the example above, utilizing the body-part context (red box) and the stuff context (red box) can ease the recognition of *“ride, bicycle”*. However, background objects like the house can mislead the recognition. The proposed work can handle such cases by selecting the most useful context features and combining them as we did as a context selection scheme Self-Selective Context to rely only on the most discriminative contexts.

Experiment 2

TABLE II: Self-Selection of The Multiple Contexts.

Method	Dataset		
	HICO	V-COCO	CINT
HO-only	62.50	52.27	45.24
HO + Fusion	69.59	54.74	49.74
HO + SSC (Ours)	70.78	55.00	54.36

Method: Self-Selective Context Module

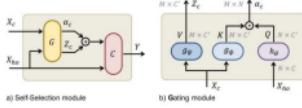


Fig. 2: Overview of our method. On the left, the Self-Selection module: It takes as an input the features X_{HO} , of N human-object pairs in a certain image, and M context features X_c , corresponding to the image. Then, it modulates the context features X_c using a selection module G . The output of G is then concatenated with the input X_{HO} and passed through a GLU (\odot) to produce the human-object interactions in the image. On the right, the Gating module G [1], inspired from the Self-attention [10]. The main purpose of $G(\cdot)$ is to embed the heterogeneous context features X_c into a compact representation Z_G . $G(\cdot)$ predicts the vectors a used for Self-Selection of the embedded context features Z_G .

Method: Proposed Context Features

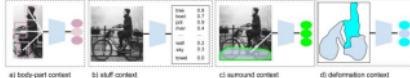
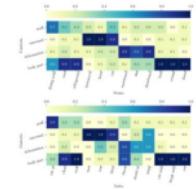
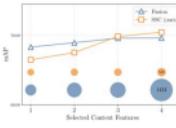


Fig. 3: The context features proposed in this paper. a.) Body-part context models the appearance of the human joints, b.) Stuff-context models the occurrence of stuff regions in the image, c.) Surround context models the appearance of local segments around humans, d.) Deformation context models the shape of the human-object posture.

Quantitative Analysis

TABLE IV: Contribution of Joint Conditioning.

Condition	Dataset		
	HICO	V-COCO	CINT
context-only	62.77	49.38	44.83
human-object & context	70.78	55.00	54.36



Experiment 3

TABLE V: Combining SSC with the State-of-the-art.

Method	Dataset		
	HICO	V-COCO	CINT
VGG-16 [16]	56.10	46.91	44.83
VGG-16 [16] + SSC	67.59	51.17	49.56
ConvFus [5] + SSC	63.47	51.36	46.72
ConvFus [5] + SSC	65.79	52.24	51.92
PairAtt [7] + SSC	65.10	53.62	48.99
PairAtt [7] + SSC	68.29	54.24	51.22
Human-Object	62.50	51.60	47.85
Human-Object + SSC	70.78	55.00	54.36

Experiments

Datasets: HICO, V-COCO, CINT (ours)

Evaluation: mAP

Qualitative Analysis



Fig. 6: Qualitative examples from CINT dataset for PairAtt [7]. Top row shows images with 100% detection results. Bottom row shows images where the context is unexpected (top), however may decrease the result if the context is not visible or too noisy (bottom).

Conclusion

This paper tackled human-object interaction recognition from a single image

We treated the task as a context selection task via Self-Selective Context

Experiments on three benchmarks demonstrate the effectiveness of the technique