Weakly Supervised Affordance Detection

Johann Sawatzky*[1] Abhilash Srikantha*[2] Juergen Gall[1]

[1]University of Bonn [2]Carl Zeiss AG

UNIVERSITÄT BONN

Summary

- Affordances
  - Affordances are attributes of object parts which indicate functionality
  - Detecting affordances can be viewed as a multichannel image segmentation problem
- Problems
  - Affordance segmentation is more difficult than object segmentation due to object parts associated with multiple affordances and the higher level of abstraction: A system should detect affordances of previously unseen classes
  - No datasets of object affordances in context available
  - Pixel wise annotations are expensive, a weakly supervised training method is desirable
- Contributions
  - CAD120 affordance dataset introduced: Pixelwise annotated object affordances in human context
  - Weakly and strongly supervised affordance segmentation methods proposed, both generalize affordances to object classes absent in train set
  - Weakly supervised method is the state of the art in affordance segmentation, outperforming existing image segmentation methods
- Code and new data set publicly available

CAD120 affordance dataset

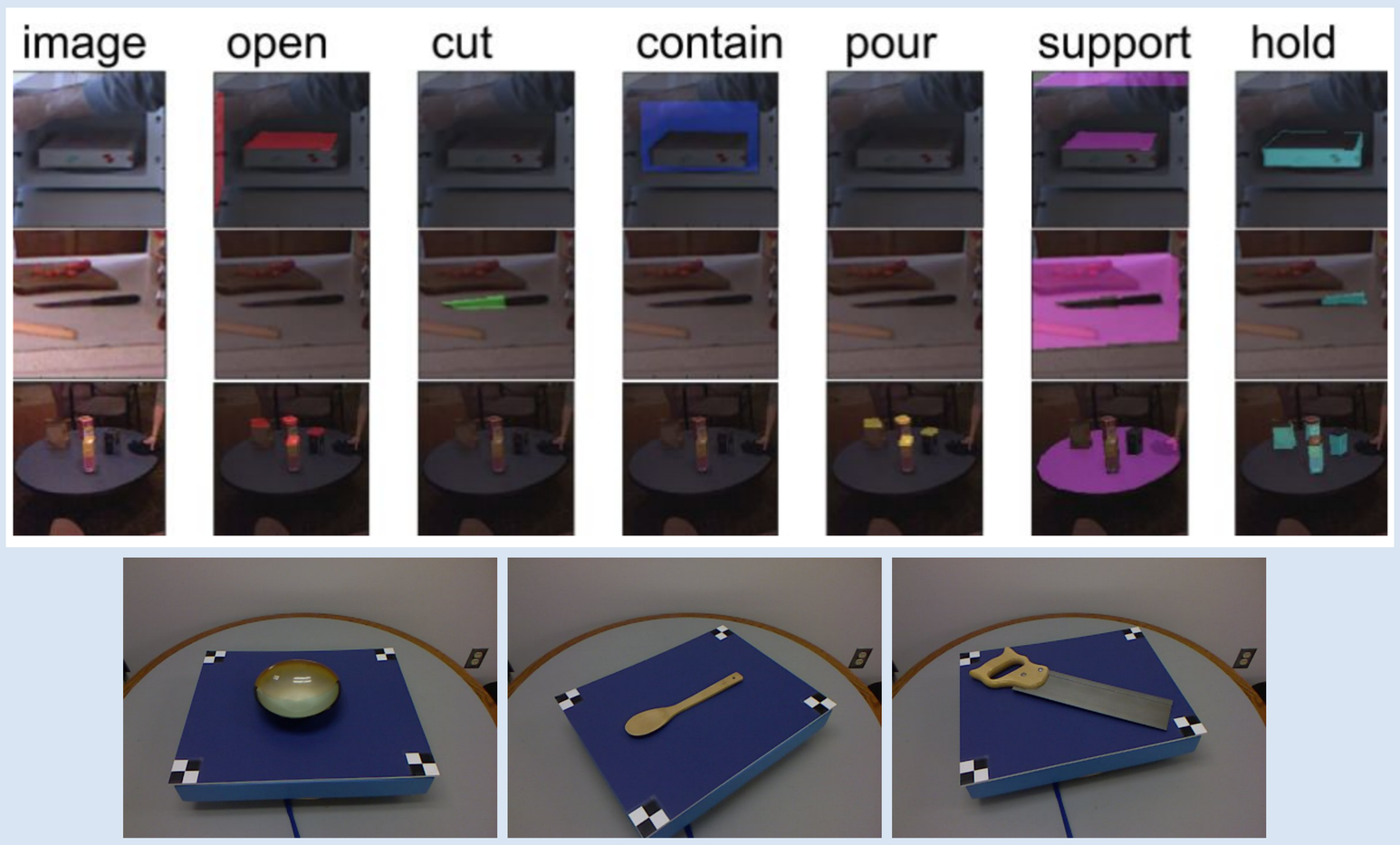


Figure 1: Examples of pixel wise annotations from our CAD120 affordances dataset (left) and example images of the UMD part affordance dataset [4] (right)

We propose a new dataset showing objects in context with pixel wise affordance annotations

- 3090 video frames from the CAD 120 dataset [5] depicting 9916 object instances in context of human interaction
- 6 affordances: open, cut, pour, contain, support, hold
- 12 object classes: table, palte, thermal cup, medicine box, microwave, bowl, coffee pot, bottle, knife, can, paper box, mug
- Human poses available for all frames
- Available at https://zenodo.org/record/495570

For evaluation we also use the UMD part affordance dataset [4].

Strongly Supervised Method

We adapt the VGG and ResNet architectures from [6] by exchanging the final softmax layer for a sigmoid layer.
The prediction is given by

$$P(y_{i,l}|I;\theta) = \frac{1}{1+\exp\left(-f_{i,l}\left(y_{i,l}|I;\theta\right)\right)}, \tag{1}$$

where $f_{i,l}\left(y_{i,l}|I;\theta\right)$ is the output of the CNN at pixel x_i and affordance l without the softmax layer.
Training means minimizing

$$J(\theta) = \log P(Y|I;\theta) = \sum_{i=1}^{n}\sum_{l\in\mathcal{L}} \log P(y_{i,l}|I;\theta), \tag{2}$$

Where Y is the ground truth labeling of the pixels, $\mathcal{L}$ is the set of affordances, θ are the CNN parameters and I is the image.
For inference we threshold at 0.5:

$$\hat{y}_{i,l} = \begin{cases} 1 & \text{if } P(y_{i,l}|I,Z_x) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

*contributed equally sawatzky@iai.uni-bonn.de, abhilash.srikantha@zeiss.com, gall@iai.uni-bonn.de

Weakly Supervised Affordance Detection

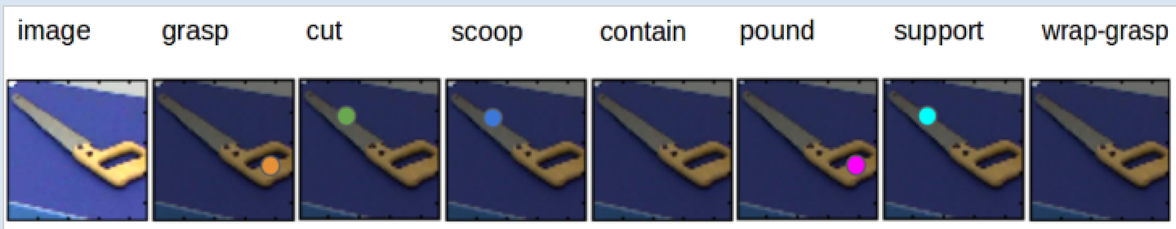
Johann Sawatzky*[1] Abhilash Srikantha*[2] Juergen Gall[1]

[1]University of Bonn [2]Carl Zeiss AG

UNIVERSITÄT BONN

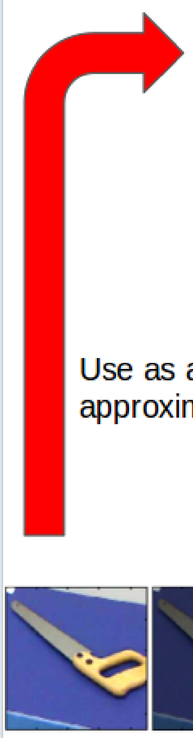
Weakly Supervised Method

Our EM-algorithm, in practice we use two M-steps and 1 E-step.
Where l: affordance class, i: pixel, k: key point, x_i: spatial pixel coordinate, x_k: spatial coordinate of affordance key point, Z_x: set of affordance key points.



Initialization of pixelwise labels:

$$\hat{y}_{i,l} = \begin{cases} 1 & \text{if } |\{(z_k, x_k) \in Z_x : z_k = l \wedge |x_k - x_i| \leq \sigma\}| > 0 \\ 0 & \text{otherwise} \end{cases}$$



M-step

E-step

$$P(Y|I, Z_x) = \sum_{l \in \mathcal{L}} P(Y, l|I, Z_x) = \sum_{l \in \mathcal{L}} P(Y|l, I, Z_x) P(l|I, Z_x) \approx \sum_{l \in \mathcal{L}} P(Y_l|I; \theta) P(l|Z_x).$$

$$P(l|Z_x) = \begin{cases} 1 & \text{if } |\{(z_k, x_k) \in Z_x : z_k = l\}| > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Estimated pixelwise labels for training images

Results intra object class affordance generalization

In this setup, all object classes occur in the train as well as in the test set.

UMD dataset (category split)	Grasp	Cut	Scoop	Contain	Pound	Support	Wgrasp	mean
Fully Supervised IoU category split								
HMP + SVM [4]	0.57	0.37	0.70	0.77	0.41	0.49	0.79	0.59
DEP + SRF [4]	0.35	0.15	0.38	0.65	0.18	0.26	0.80	0.40
Proposed (VGG)	0.66	0.77	0.85	0.84	0.64	**0.73**	0.82	**0.76**
Proposed (ResNet)	**0.71**	**0.79**	**0.86**	**0.86**	**0.72**	0.55	**0.84**	**0.76**
Weakly Supervised IoU category split								
Proposed (VGG)	0.46	0.48	**0.72**	**0.78**	0.44	**0.53**	0.65	0.58
Proposed [7] (VGG)	0.55	0.48	**0.72**	0.76	**0.49**	0.48	0.67	0.59
Proposed (ResNet)	0.42	0.35	0.67	0.70	0.44	0.44	**0.77**	0.54
Proposed [7] (ResNet)	**0.57**	**0.54**	0.71	0.70	0.43	0.54	0.69	**0.60**
Image label [3]	0.06	0.19	0.04	0.22	0.12	0.02	0.08	0.10
Area constraints [3]	0.06	0.04	0.10	0.14	0.22	0.04	0.37	0.14
SEC [1]	0.39	0.16	0.27	0.13	0.35	0.19	0.07	0.22
WTP [2]	0.16	0.14	0.20	0.20	0.01	0.07	0.13	0.13

CAD120 affordance dataset (actor split)	Bck	Open	Cut	Contain	Pour	Support	Hold	Mean
Fully Supervised IoU category split								
Proposed (VGG)	0.81	0.67	0.00	0.54	0.42	0.70	0.64	0.54
Proposed (ResNet)	**0.86**	**0.71**	0.00	**0.61**	**0.45**	**0.79**	**0.70**	**0.59**
Weakly Supervised IoU category split								
Proposed (VGG)	0.61	0.33	0.00	0.35	0.30	0.22	0.43	0.32
Proposed [7] (VGG)	0.71	0.47	0.0	0.36	0.37	0.56	0.49	0.42
Proposed (ResNet)	0.60	0.25	0.00	0.35	0.30	0.17	0.42	0.30
Proposed [7] (ResNet)	**0.77**	**0.50**	0.00	**0.43**	**0.39**	**0.64**	**0.56**	**0.47**
SEC [1]	0.53	0.43	0.00	0.25	0.09	0.02	0.20	0.22
WTP [2]	0.53	0.13	0.00	0.10	0.08	0.11	0.22	0.17
Image label [3]	0.55	0.05	0.01	0.09	0.10	0.02	0.21	0.15
Area constraints [3]	0.53	0.11	**0.02**	0.09	0.09	0.07	0.15	0.15

Table 1: Evaluation of fully and weakly supervised approaches for affordance detection on the UMD part affordance dataset (category split) and the CAD 120 affordance dataset (actor split). Evaluation metric is IoU. We could improve our method in [7].

Results with cross object class generalization

None of the object classes from the test set is present in the train set.

UMD dataset (novel split)	Grasp	Cut	Scoop	Contain	Pound	Support	Wgrasp	mean
Fully Supervised IoU novel split								
HMP + SVM [4]	0.29	0.10	0.61	**0.74**	0.03	0.24	0.63	0.38
DEP + SRF [4]	0.32	0.04	0.23	0.42	**0.16**	0.22	0.81	0.31
Proposed (VGG)	**0.37**	0.35	0.65	0.62	0.10	**0.52**	**0.85**	**0.50**
Proposed (ResNet)	0.33	**0.51**	**0.69**	0.52	0.09	0.51	**0.85**	**0.50**
Weakly Supervised IoU novel split								
Proposed (VGG)	0.27	0.14	0.55	**0.58**	0.02	0.37	0.67	0.37
Proposed [7] (VGG)	0.31	0.18	0.56	0.49	**0.08**	0.41	0.66	0.38
Proposed (ResNet)	0.25	0.21	**0.62**	0.50	**0.08**	**0.43**	0.67	0.40
Proposed [7] (ResNet)	**0.34**	**0.34**	0.58	0.40	0.07	0.42	**0.77**	**0.42**
Image label [3]	0.04	0.00	0.09	0.16	0.01	0.02	0.32	0.09
Area constraints [3]	0.05	0.00	0.04	0.16	0.00	0.01	0.32	0.09
SEC [1]	0.12	0.03	0.06	0.23	0.07	0.12	0.25	0.13
WTP [2]	0.11	0.03	0.18	0.11	0.00	0.02	0.23	0.10

CAD120 affordance dataset (object split)	Bck	Open	Cut	Contain	Pour	Support	Hold	Mean
Fully Supervised IoU object split								
Proposed (VGG)	0.76	0.10	0.27	0.60	0.45	0.66	**0.60**	0.49
Proposed (ResNet)	**0.80**	**0.22**	**0.50**	**0.62**	**0.48**	**0.75**	**0.60**	**0.57**
Weakly Supervised IoU object split								
Proposed (VGG)	0.62	0.08	0.08	0.24	0.22	0.20	0.46	0.27
Proposed [7] (VGG)	0.68	0.10	**0.23**	0.44	0.36	0.50	0.47	0.40
Proposed (ResNet)	0.69	0.11	0.09	0.28	0.21	0.36	**0.56**	0.33
Proposed [7] (ResNet)	**0.74**	**0.15**	0.21	**0.45**	**0.37**	**0.61**	0.54	**0.44**
SEC [1]	0.54	0.04	0.09	0.13	0.09	0.08	0.13	0.16
WTP [2]	0.57	0.01	0.00	0.02	0.09	0.03	0.19	0.13
Image label [3]	0.58	0.00	0.00	0.00	0.00	0.00	0.23	0.12
Area constraints [3]	0.59	0.03	0.03	0.01	0.02	0.02	0.28	0.14

Table 2: Evaluation of fully and weakly supervised approaches for affordance detection on the UMD part affordance dataset (novel split) and the CAD 120 affordance dataset (object split). Evaluation metric is IoU. We could improve our method in [7].

References

1 A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In ECCV, 2016.
2 A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. Whats the Point: Semantic Segmentation with Point Supervision. ECCV, 2016.
3 G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In ICCV, 2015.
4 A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In ICRA, 2015.
5 H. S. Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. In ECCV, 2014.
6 L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and L. A. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv:1506.02106v5, 2016
7 J.Sawatzky and J.Gall. Adaptive Binarization for Weakly Supervised Affordance Segmentation. arXiv:1707.02850v1, 2017

*contributed equally sawatzky@iai.uni-bonn.de, abhilash.srikantha@zeiss.com, gall@iai.uni-bonn.de