EyeGuide - From Gaze Data to Instance Segmentation

Jacqueline Kockwelp[1,2,3], Jörg Gromoll[2], Joachim Wistuba[2], Benjamin Risse[1,3]

[1]*Institute for Geoinformatics (ifgi)*
University of Münster
Münster, Germany

[2]*Centre of Reproductive Medicine and Andrology (CeRA)*
University Hospital Münster
Münster, Germany

[3]*Faculty of Mathematics and Computer Science*
University of Münster
Münster, Germany

Introduction

Precise segmentation, e.g. the identification and labeling of complex and non-convex components in images, is a challenging task in machine learning. To facilitate the annotation process, we propose the novel guidance strategy "EyeGuide", which uses a remote eye tracking system passively recording gaze information from a human annotator during inspection. The information gained is used as additional input for neural network training for automatic prediction of segmentation masks. We found gaze data acquisition to be faster and more convenient with fewer annotations being necessary to generate the proper segmentation, and overall better generalization compared to state-of-the-art techniques was achieved.

Method

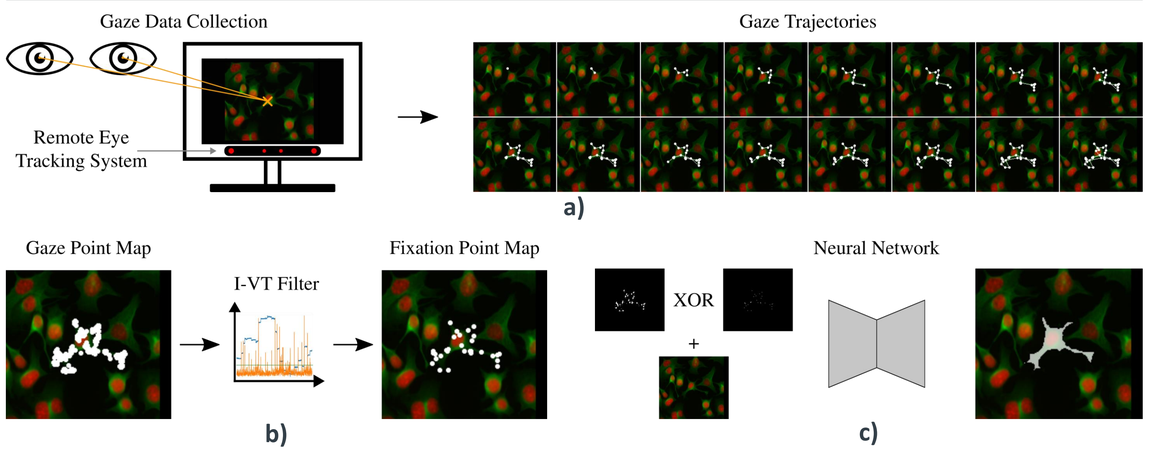

Figure 1: EyeGuide overview. a) Gaze information of a human annotator is recorded for individual object instances. b) Raw gaze information is filtered using an I-VT filter to generate fixation points. c) The raw gaze or fixation point map is concatenated with the RGB image and serves as a 4-channel input for a neural network to predict instance segmentations.

Experimental Design:

- Tobii Pro Fusion screen-based remote eye tracker at 120 Hz
- To inform the user about the object of interest the bounding box and the ground truth polygon mask were presented for 0.5 seconds
- For very small objects initial zoom was applied
- No time limit was given for inspection
- User signals the object inspection start and end by pressing a button
- Possibility to repeat the observation and to move freely in the presented image (dragging, scrolling and zooming)

Figure 2: Comparison between task-based (contour tracing) and task-free image inspection. a) Ground truth mask. b) Gaze point map with explicit task. c) Gaze point map with task-free inspection. d) Training and test loss for task-based vs. task-free inspection.

Datasets

- PascalVOC2012 train [1]: 1,645 images with 3,507 objects divided into 20 classes
- Cellpose [2]: 500 cell instances of the same type

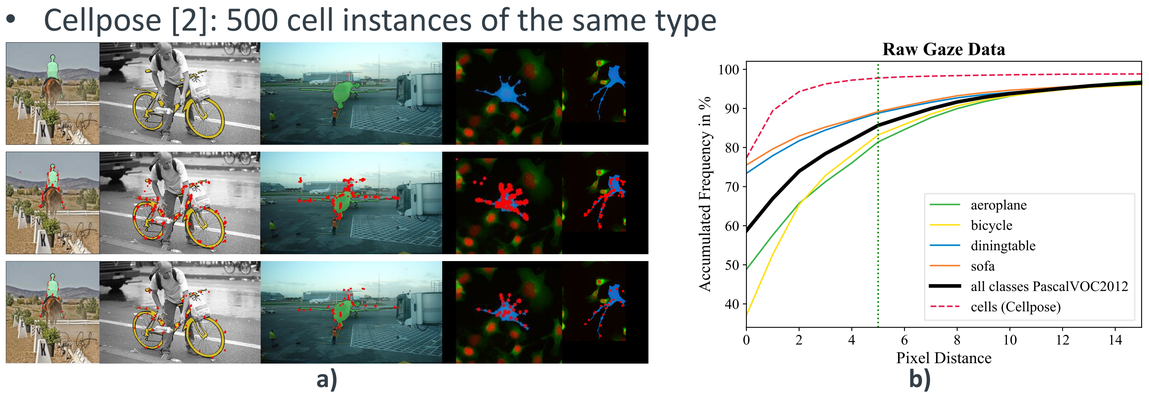

Figure 3: a) Dataset examples for PascalVOC2012 and Cellpose. First row: Images with ground truth mask. Second row: Images with gaze point map visualization. Third row: Images with fixation point map visualization. b) Raw Gaze Data grouped by their distance to the ground truth mask of PascalVOC2012 (mean of all classes and four selected classes) and Cellpose dataset.

Results

Architecture	fixation filter	blurred	gaze data jitter	gaze data dropout	mIoU PascalVOC2012	mIoU Cellpose
FCN+ResNet-50 (without cropping)					53.2	-
FCN+ResNet-50					75.4	82.9
FCN+ResNet-50	x				74.6	82.7
FCN+ResNet-50		x			75.8	83.3
FCN+ResNet-50			x		75.9	81.8
FCN+ResNet-50			x	x (30%)	75.7	-
FCN+ResNet-50				x (30%)	76.3	83.4
FCN+ResNet-101					76.2	-
DeepLabv3+ResNet-50					75.9	-
FCN+ResNet-101				x (30%)	**76.4**	-
FCN+ResNet-50		x		x (30%)	76.2	**83.6**
DEXTR [3]					70.3	78.9
DEXTR (pretrained)					75.9	82.7

Table 1: Overview experiments. Evaluation of different architectures and model configurations for PascalVOC2012 and Cellpose.

- **Annotation Efficiency:** average annotation time was 6.19 seconds for PascalVOC2012 and 5.13 seconds for Cellpose (for comparison: polygon mask ~79 seconds, bounding box ~35 seconds, extreme points ~7.5 seconds [4])
- Segmentation performance evaluation depending on the convexity c (see Figure 4), defined as:

$$c = \frac{area\ of\ gt\ mask}{convex\ hull\ area\ of\ gt\ mask}, c\ \in [0,1]$$

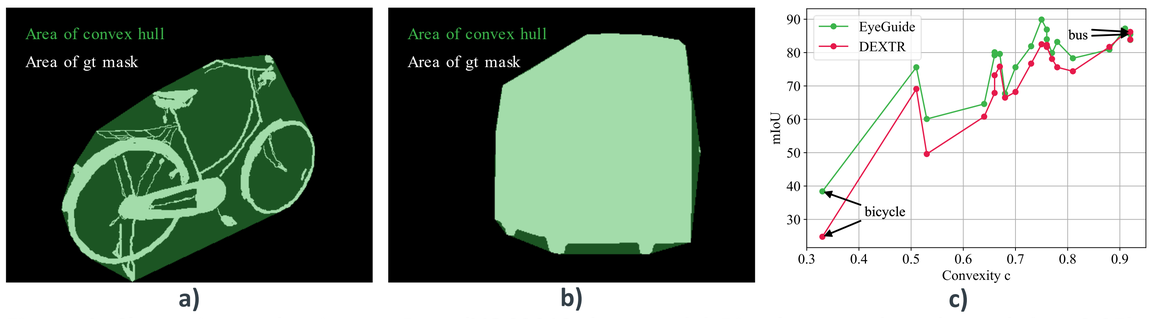

Figure 4: Class-wise evaluation on PascalVOC2012 dataset. (a) Bicycle ground truth mask (c = 0.25). (b) Bus ground truth mask (c = 0.95). (c) Convexity mIoU plot for every class.

	seen classes						unseen classes											
	train			test			train			test			train			test		
	# img	mIoU Ours	mIoU DEXTR	# img	mIoU Ours	mIoU DEXTR	# img	mIoU Ours	mIoU DEXTR	# img	mIoU Ours	mIoU DEXTR	# img	mIoU Ours	mIoU DEXTR	# img	mIoU Ours	mIoU DEXTR
person_net	696	77.8	68.3	174	**75.1**	63.2	-	-	-	852	**65.6**	60.7	847	**78.2**	64.5	938	**65.2**	59.2
vehicle_net	685	76.5	70.4	167	**71.3**	64.8	870	**67.8**	56.8	-	-	-	847	**73.8**	64.1	938	**63.2**	58.4
animal_net	680	85.2	78.9	167	**81.1**	74.3	870	**71.3**	61.2	852	**64.1**	62.0	-	-	-	938	**63.7**	59.2
indoor_net	752	72.1	66.1	186	**66.6**	61.5	870	**69.0**	58.6	852	**63.2**	60.1	847	**73.9**	62.5	-	-	-
all_net	2,813	79.6	74.6	694	**76.4**	70.3	-	-	-	-	-	-	-	-	-	-	-	-

Table 2: Evaluation of the generalisation capabilities of EyeGuide vs. DEXTR [3] to unseen classes. mIoU results for training on a subset of classes and testing on the subsets that were left out.

Outlook

With EyeGuide we demonstrate that the use of eye tracking offers great potential for the annotation of new datasets. In the future, we want to investigate this in more detail, including the integration of eye tracking as a prompt for existing foundation models and exploiting eye tracking specific properties such as temporal information in the form of more specialized deep learning architectures.

References

[1] M. Everingham et al. "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results."

[2] C. Stringer et al. "Cellpose: a generalist algorithm for cellular segmentation". Nature Methods, 18(1):100–106, 2021

[3] K.-K. Maninis et al. "Deep extreme cut: From extreme points to object segmentation". In Computer Vision and Pattern Recognition (CVPR), 2018

[4] D. P. Papadopoulos et al. "Extreme clicking for efficient object annotation". In IEEE International Conference on Computer Vision (ICCV), pages 4940–4949, 2017

j.kockwelp@uni-muenster.de
b.risse@uni-muenster.de