Supplemental Material

Instance Segmentation of Event Camera Streams in Outdoor Monitoring Scenarios

Tobias Bolten[1][a], Regina Pohle-Fröhlich[1][b] and Klaus D. Tönnies[2]
[1]*Institute for Pattern Recognition, Hochschule Niederrhein, Krefeld, Germany*
[2]*Department of Simulation and Graphics, University of Magdeburg, Germany*
{*tobias.bolten, regina.pohle*}*@hs-niederrhein.de, klaus@isg.cs.uni-magdeburg.de*

1 DATASETS

1.1 Event Count Statistics

Histograms of the event counts contained in the extracted adaptive region of interest are shown in Figure 1. Additional statistical event count parameters are as follows:

DVS-iOUTLAB dataset aRoIs:

Mean:	1830.6182
Std Dev:	1663.9163
Minimum:	126.0000
Maximum:	24700.0000
Quartile:	664.0000
Median:	1265.0000
Quartile:	2452.0000

N-MuPeTS dataset aRoIs:

Mean:	2341.1705
Std Dev:	2291.8931
Minimum:	1.0000
Maximum:	28219.0000
Quartile:	864.0000
Median:	1700.0000
Quartile:	2967.0000

1.2 Splitting Details

In selecting a 60/20/20 % split of the built time blocks of the N-MuPeTS dataset, care was also taken to ensure that the selected activity annotations of the dataset were approximately equally represented in the created splits. The resulting distribution of annotations per set is shown in detail in Table 1.

The intentionally challenging subset of test data from N-MuPeTS is composed of time windows in which at least one actor is labeled with one of the following dataset annotations:

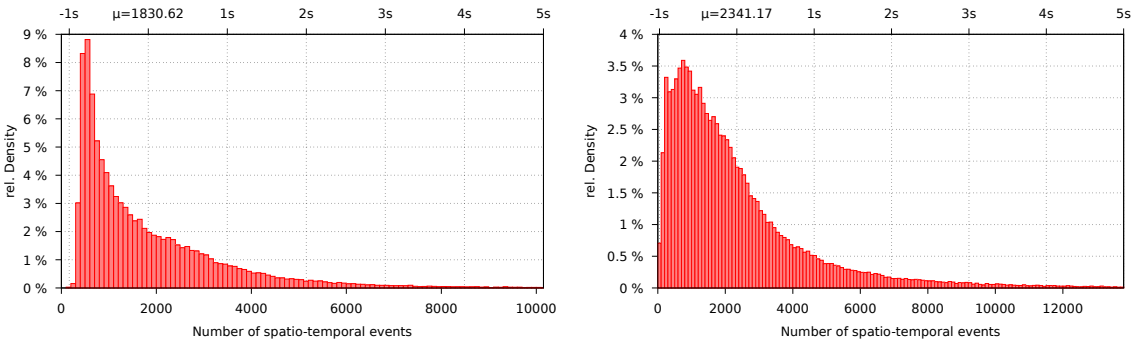
- OCCLUSION
- CROSSING
- MEET
- SIDEBYSIDE
- HELIX
- FAR

2 EXTENDED RESULTS

For the multi-class, multi-instance scenario of the DVS-iOUTLAB dataset, we also report the metrics per class. Table 2 shows the semantic quality, while Table 3 shows the instance segmentation metrics.

Regarding the experiments performed on the N-MuPeTS dataset, Table 4 shows the results on the *complete* test set.

[a] https://orcid.org/0000-0001-5504-8472
[b] https://orcid.org/0000-0002-4655-6851



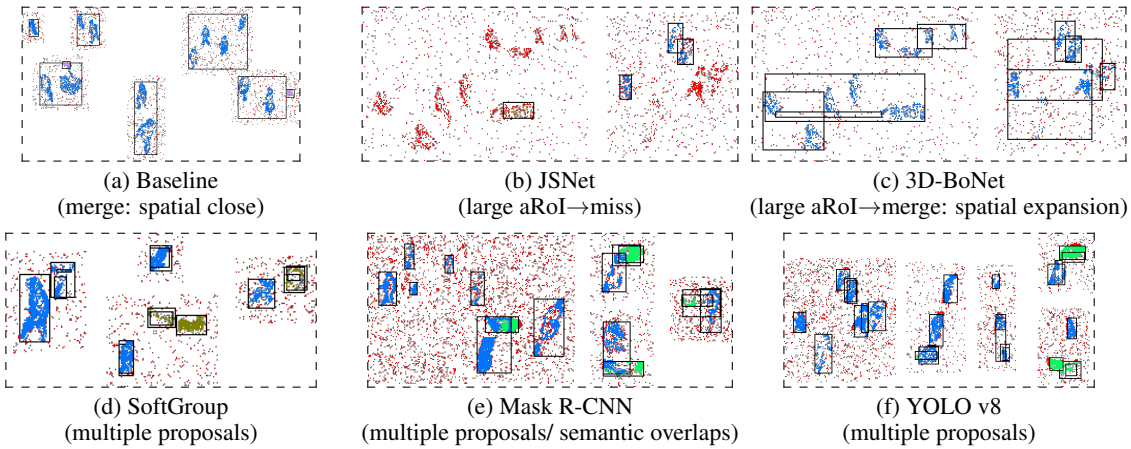
(a) DVS-iOUTLAB dataset

(b) N-MuPeTS dataset

Figure 1: Event count histograms for time Δ=10 ms filtered aRoIs (60 ms event frame windows).

Table 1: Percentage of dataset annotation labels in N-MuPeTS split (based on labels per actor given in 25 ms dataset time windows).

annotation	train	test	validation
(a) Object occlusion			
OCCLUSION	57.40	31.54	11.06
(b) Pose & Movement			
EXERCISING	61.96	30.05	7.99
WAVING	33.97	22.54	43.49
WALKING	67.57	18.87	13.55
RUNNING	69.41	17.14	13.45
RANDOM	85.13	9.62	5.25
(c) Object interaction			
CROSSING	64.37	22.48	13.15
MEET	66.40	23.26	10.34
SIDEBYSIDE	64.66	18.83	16.50
HELIX	62.15	10.80	27.05
(d) Object Size			
FAR	60.87	20.72	18.41
(e) Over-All			
mean	63.08	20.53	16.39



(a) Baseline (merge: spatial close)
(b) JSNet (large aRoI→miss)
(c) 3D-BoNet (large aRoI→merge: spatial expansion)
(d) SoftGroup (multiple proposals)
(e) Mask R-CNN (multiple proposals/ semantic overlaps)
(f) YOLO v8 (multiple proposals)

Figure 2: Typical prediction *error cases* on DVS-iOUTLAB displayed as false-color aRoI-montage images (best viewed in color and digital zoomed).

Table 2: Semantic segmentation results per class on DVS-iOUTLAB dataset (60 ms event time window).

Network	Configuration	Semantic Quality; weighted F1-score				
		PERSON	DOG	BICYCLE	SPORTSBALL	NOISE
(a) Baseline method: PointNet++ with spatial clustering						
PointNet++	in 2048 events	0.95	0.87	0.86	0.94	0.93
Clustering	in 1024 events	0.95	0.85	0.85	0.94	0.93
(b) Space-Time Event Cloud-based methods						
JSNet	4 layers in 2048 events	0.96	0.90	0.88	0.97	**0.95**
	4 layers in 1024 events	0.94	0.80	0.80	0.96	0.93
3D-BoNet	4 layers in 2048 events	0.96	0.86	0.90	0.93	0.93
	4 layers in 1024 events	0.95	0.85	0.89	0.93	0.92
(c) Voxel-based method						
SoftGroup	voxel grid $(768 \times 640 \times 60)$	**0.98**	**0.98**	**0.97**	**0.98**	0.94
(d) Frame-based methods						
Mask R-CNN	polarity in (768×640) px	0.94	0.94	0.94	0.94	0.87
	MTC in (768×640) px	0.94	0.94	0.93	0.95	0.87
YOLO v8	polarity in (768×640) px	0.94	0.90	0.93	0.86	0.86
	MTC in (768×640) px	0.94	0.86	0.93	0.85	0.86

Table 3: Instance segmentation results per class on DVS-iOUTLAB dataset (60 ms event time window)

Network	Configuration	Instance Quality: mIoU PERSON	mIoU DOG	mIoU BICYCLE	mIoU SPORTSBALL	$AP^{0.95}_{0.5}$ PERSON	$AP^{0.95}_{0.5}$ DOG	$AP^{0.95}_{0.5}$ BICYCLE	$AP^{0.95}_{0.5}$ SPORTSBALL
(a) Baseline method: PointNet++ with spatial clustering									
PointNet++ Clustering	in 2048 events	0.82	0.70	0.73	0.85	0.47	0.52	0.53	0.77
	in 1024 events	0.83	0.74	0.74	0.88	0.47	0.52	0.53	0.78
(b) Space-Time Event Cloud-based methods									
JSNet	4 layers in 2048 events	**0.91**	0.91	0.72	0.93	**0.84**	0.71	0.79	0.91
	4 layers in 1024 events	0.89	**0.92**	0.57	**0.94**	0.77	0.55	0.69	0.81
3D-BoNet	4 layers in 2048 events	0.85	0.77	0.78	0.87	0.75	0.63	0.67	0.78
	4 layers in 1024 events	0.84	0.80	0.76	0.88	0.72	0.63	0.66	0.77
(c) Voxel-based method									
SoftGroup	voxel grid $(768 \times 640 \times 60)$	0.86	0.88	**0.90**	0.93	**0.84**	**0.87**	**0.87**	**0.94**
(d) Frame-based methods									
Mask R-CNN	polarity in (768×640) px	0.78	0.79	0.79	0.84	0.51	0.60	0.63	0.74
	MTC in (768×640) px	0.77	0.75	0.78	0.83	0.50	0.61	0.59	0.73
YOLO v8	polarity in (768×640) px	0.80	0.73	0.82	0.66	0.64	0.65	0.66	0.45
	MTC in (768×640) px	0.81	0.83	0.80	0.67	0.65	0.57	0.69	0.42

Table 4: Segmentation results on *complete* N-MuPeTS test set (60 ms event time window).

Network	Configuration	Semantic quality weighted F1-score NOISE	Semantic quality weighted F1-score PERSON	mIoU	PERSON Instance Quality $AP^{0.95}_{0.5}$	PERSON Instance Quality $AP^{0.5}$	PERSON Instance Quality $AP^{0.75}$
(a) Baseline method: PointNet++ with spatial clustering							
PointNet++	in 2048 events	**0.93**	**0.96**	0.84	0.52	0.66	0.55
Clustering	in 1024 events	**0.93**	**0.96**	0.85	0.52	0.65	0.54
(b) Space-Time Event Cloud-based methods							
JSNet	4 layers in 2048 events	**0.93**	**0.96**	0.86	0.70	0.87	0.75
	4 layers in 1024 events	0.92	0.95	0.84	0.64	0.82	0.70
3D-BoNet	4 layers in 2048 events	**0.93**	**0.96**	0.86	**0.72**	0.86	**0.76**
	4 layers in 1024 events	0.91	0.95	0.83	0.62	0.78	0.66
(c) Voxel-based method							
SoftGroup	voxel grid ($768 \times 640 \times 60$)	0.88	0.95	**0.88**	0.71	0.83	0.75
(d) Frame-based methods							
Mask R-CNN	polarity in (768×640) px	0.83	0.91	0.76	0.50	0.87	0.56
	MTC in (768×640) px	0.83	0.91	0.76	0.52	0.88	0.58
YOLO v8	polarity in (768×640) px	0.84	0.93	0.75	0.64	**0.91**	0.75
	MTC in (768×640) px	0.84	0.93	0.76	0.64	**0.91**	0.64