EV-Mask-RCNN: Instance Segmentation in Event-based Videos

RQ: Can we train deep networks to do instance segmentation on event-based cameras?

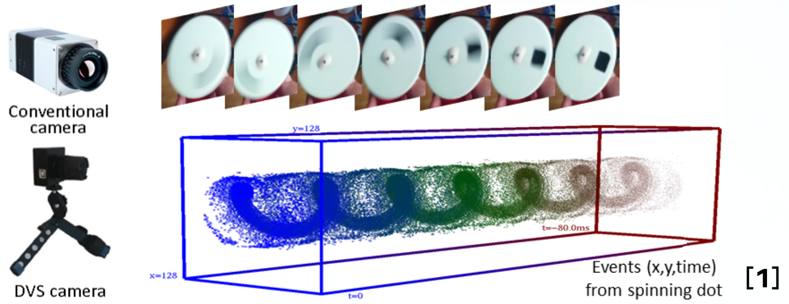
Ana Băltărețu

Nergis Tömen
Ombretta Strafforello
Xin Liu

TU Delft

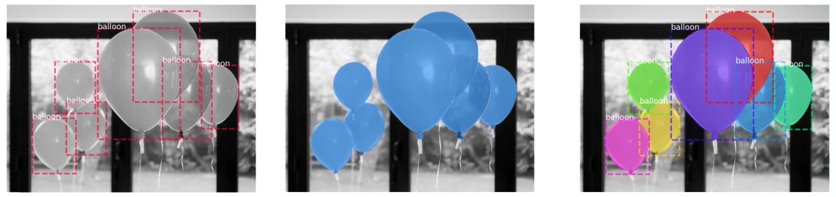
1 Introduction

Event-based cameras are asynchronous sensors that detect changes in light intensity at every pixel. An event with coordinates (x, y), polarity (p) and timestamp(t) is denoted as e(x, y, p, t).



[1]

Instance segmentation combines bounding box detection and semantic segmentation producing a mask for each individual instance of a class.



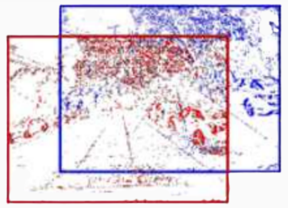
Bounding Box Detection Semnatic Segmentation Instance Segmentation [2]

2 Concept

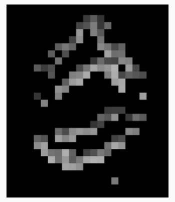
The **goal** was to figure out how to transform event-based data such that a **deep network** could be trained to detect **masks** for each instance of an object.

1) Model choice: **Mask R-CNN**
2) Data representation: **Two channel images** and **depth frames**
3) Performance evaluation: **Acc, mIoU, mAP**

Mask R-CNN [3]



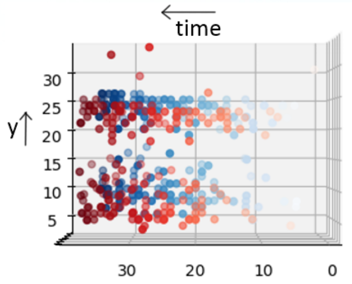
Two-channel image [4]



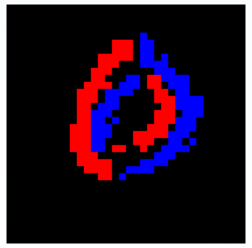
Depth frame

Contact
a.baltaretu@student.tudelft.nl

3 Methodology

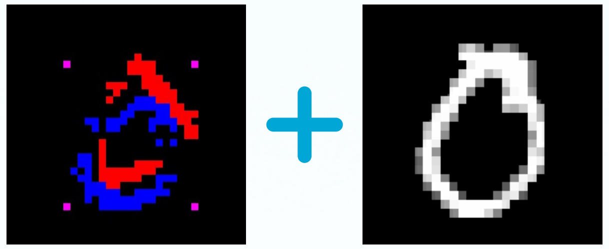


Events time stream



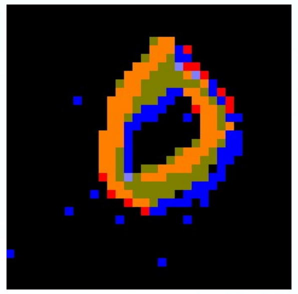
Denoised frame

1) Split the events time stream into frames of fixed time windows.



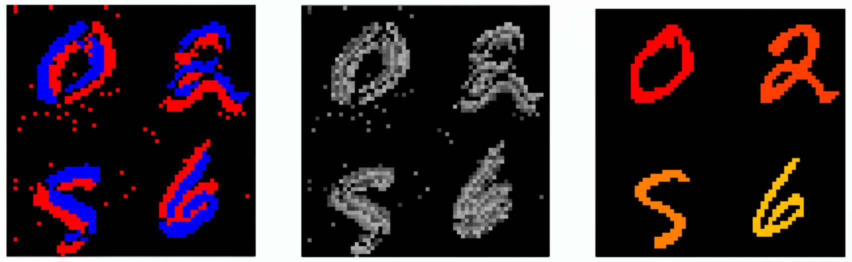
Digit extremes **MNIST mask**

2) Find extremities of digit and its corresponding MNIST mask.



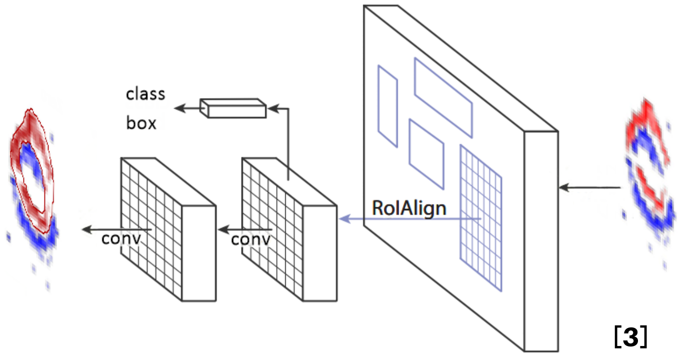
Aligned mask

3) Calculate overlap score and align the mask on top of the negative events.



Frame **Depth** **Mask**

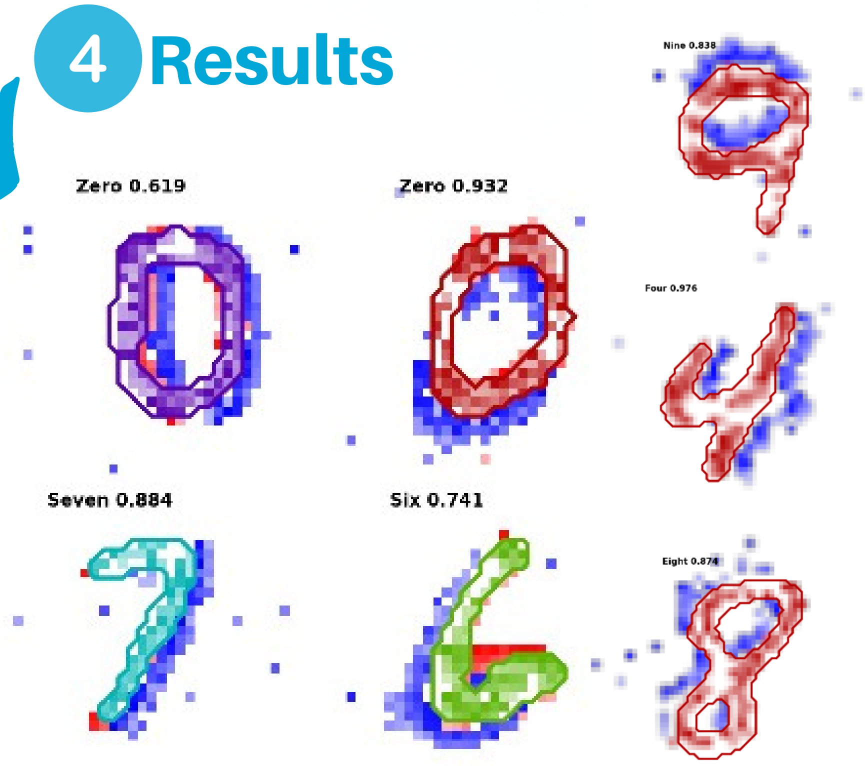
4) Save noised frames, depth frames and masks.



[3]

5) Train Mask R-CNN using the generated RGB-D images and masks.

4 Results



5 Conclusion

The qualitative and quantitative results are **promising** and comparable to others from literature [5], [6].

Future work:
1) Generate a dataset with thicker objects, similar to N-MNIST.
2) Label DDD17 dataset [7] and see how this model compares to results from other papers.
3) Compare event-based models directly to frame-based models.

References

[1] F. Barranco, C. L. Teo, C. Fermuller, and Y. Aloimonos, "Contour detection and characterization for asynchronous event sensors," in Proceedings of the IEEE International Conference on Computer Vision, pp. 486–494, 2015.

[2] W. Abdulla, "Splash of color: Instance segmentation with mask r-cnn and tensorflow," Mar 2018.

[3] K. He, G. Gkioxari, P. Doll ́ar, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international conference on computer vision, pp. 2961–2969, 2017.

[4] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza, "End-to-end learning of representations for asynchronous event-based data," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5633–5643, 2019.

[5] I. Alonso and A. C. Murillo, "Ev-segnet: Semantic segmentation for event-based cameras," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0, 2019.

[6] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. Yoon, "Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation," 2021.

[7] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd17: End-to-end davis driving dataset," arXiv preprint arXiv:1711.01458, 2017.