Visual Tracking under Fast Motions with an Event Camera

Irene Pérez Salesa, Rodrigo Aldana López, Carlos Sagüés Blázquiz

Robótica, Percepción y Tiempo Real (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: i.perez@unizar.es

Abstract

Visual object tracking under fast motions can be hindered by conventional cameras, which are prone to producing blurry images. Event cameras are better suited for this task, but applying them to object tracking is non-trivial. We propose a framework to take advantage of these sensors by using deep learning techniques.

Introduction

Object tracking is a widely studied problem in the field of robotics. The use of neural networks for tracking-by-detection provides new application opportunities. However, robotic problems featuring highly dynamic environments need vision sensors able to produce clear images in these conditions. Conventional cameras tend to produce frames with motion blur, on which the detection task might fail. Event cameras show low latency, high dynamic range and robustness to motion blur. However, they asynchronously capture light intensity changes in each independent pixel, rather than a full image of absolute intensity. For this reason, traditional computer vision algorithms cannot be directly applied to their output. Some works have tried to combine information from conventional frames and events simultaneously [1] or to adapt neural networks to detect objects in the event domain [2]. The scarcity of available labelled event data is a problem for these approaches. In addition, the event-domain representation suffers from great variability. Recently, neural networks to reconstruct intensity frames from event information have been proposed [3]. We take advantage of them to create an object-tracking framework that is robust to motion blur.

Framework

Our proposal uses an image reconstruction neural network fed by events to create absolute intensity frames. In a later step, an object detection network is used to locate the object within the reconstructed frame. Such detection is used as a measurement in a multi-rate Kalman filter in order to obtain a full estimate of the state of the target.

For the image reconstruction task, we use the E2VID network, which takes event data as an input and produces grayscale intensity frames [3]. The events are grouped into tensors of a fixed number of elements. This means that we create a new reconstruction only when significant change has happened in the scene, given that events represent changes in light intensity. Therefore, no redundant images are produced and the frequency of the reconstruction adapts to the dynamics of the scene. For object detection, we use the YOLOv5 neural network [4] to locate objects within the reconstructed frames. The measurement of the target's position is considered to be the center of the detected bounding box. Finally, we use a Kalman filter to estimate the position and velocity of the target. Since a new measurement is obtained only when a certain amount of events is reached, the tracker updates its estimate asynchronously. We consider one target, but our approach can be extended to multiple object tracking in the same way as traditional frame-based trackers-by-detection.

Experiments

The experiments have been performed on sequences from the VisEvent and Event Camera datasets [1, 5], which provide paired conventional frames and event data captured with a DAVIS camera. The VisEvent dataset also includes groundtruth annotations of the object's location on the conventional frames. The sequences include challenging conditions such as fast motions, low light and high dynamic range, as well as background events due to camera motion.

First, we compare the quality of detection on conventional frames and reconstructed ones from the event data. Figure 1 shows a qualitative comparison of the detection with both options. Figure 2 shows the precision and recall metrics, computed on sequences with motion blur. Even though reconstructed frames show a slightly lower precision, they allow to detect the object in a greater amount of frames, providing better recall values. This means that, for the same temporal sequence, a significant increase in the

number of detections is achieved using reconstructed frames. The impact of using different amounts of events in each tensor to create a reconstruction is also captured (the *number of events per pixel* is used to make the values agnostic to sensor dimensions).

Once the detection task has been tested, we compare our tracking framework to a baseline tracker that uses conventional frames. For a sequence with strong motion blur, our tracker is able to produce a better estimate, given the decrease in blurriness that the reconstructed frames provide. Figure 3 shows the estimation results. Since very few detections are achieved on the conventional frames, the error and uncertainty of the estimation increase greatly. Note that the estimate is plotted from the first detection: initially, there is a time lapse for which no detections are found on the conventional frames, due to motion blur. By using reconstructed frames from event data, we manage to mitigate these issues.

Conclusions

We have presented a framework to breach the gap between events and deep learning for object detection and tracking applications. A detection-based object tracker that relies solely on event information has been implemented, by reconstructing images from events via an E2VID network and then performing detection with a YOLO neural network. It has been shown that this framework maintains the advantages of event cameras: the event tracker shows a superior performance to conventional frames in scenes where fast motions are present, while conventional cameras are not able to produce clear images for detection.

REFERENCES

[1]. WANG, X., LI J., ZHU L., ZHANG, Z., CHEN, Z., LI, Y., WANG, Y.T. and WU, F. VisEevent: Reliable object tracking via collaboration of frame and event flows. 2021. Available from: doi:arxiv-2108.05015.

[2]. IACONO, M., WEBER, S., GLOVER, A. and BARTOLOZZI, C. Towards event driven object detection with off-the-shelf deep learning. *IEEE International Conference on Intelligent Robots and Systems.* 2018, pp. 6277–6283.

[3]. REBECQ, H., RANFTL, V., KOLTUN, V. and SCARAMUZZA, D. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021, 43(6), pp. 1964–1980.

[4]. JOCHER, G. et al., YOLOv5n 'Nano' models, Roboflow integration, Tensor-Flow export, OpenCV DNN support. 2021. Available from: doi:10.5281/zenodo.5563715.

[5]. MUEGGLER, E., REBECQ, H., GALLEGO, G., DELBRUCK, T. and SCARAMUZZA, D. The Event-Camera Dataset and Simulator: Event-based data for pose estimation, visual odometry, and SLAM. *The International Journal of Robotics Research.* 2016, 36(2), pp. 142-149.

Figure 1. Comparison of detection on conventional frames (top) and reconstructed frames from event data (bottom).

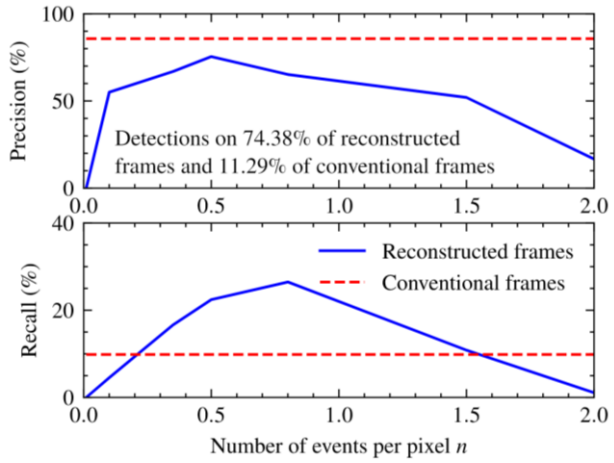

Figure 2. Metrics for detection quality on conventional and reconstructed frames.

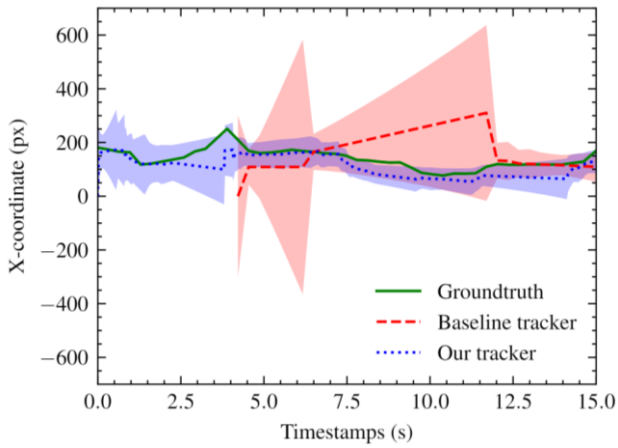

Figure 3. Tracking results for our framework and a conventional frame-based tracker.