Tracking-Learning-Detection

Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas

Abstract—This paper investigates long-term tracking of unknown objects in a video stream. The object is defined by its location and extent in a single frame. In every frame that follows, the task is to determine the object's location and extent or indicate that the object is not present. We propose a novel tracking framework (TLD) that explicitly decomposes the long-term tracking task into tracking, learning, and detection. The tracker follows the object from frame to frame. The detector localizes all appearances that have been observed so far and corrects the tracker if necessary. The learning estimates the detector's errors and updates it to avoid these errors in the future. We study how to identify the detector's errors and learn from them. We develop a novel learning method (P-N learning) which estimates the errors by a pair of "experts": 1) P-expert estimates missed detections, and 2) N-expert estimates false alarms. The learning process is modeled as a discrete dynamical system and the conditions under which the learning guarantees improvement are found. We describe our real-time implementation of the TLD framework and the P-N learning. We carry out an extensive quantitative evaluation which shows a significant improvement over state-of-the-art approaches.

Index Terms—Long-term tracking, learning from video, bootstrapping, real time, semi-supervised learning.

1 Introduction

Consider a video stream taken by a hand-held camera depicting various objects moving in and out of the camera's field of view. Given a bounding box defining the object of interest in a single frame, our goal is to automatically determine the object's bounding box or indicate that the object is not visible in every frame that follows. The video stream is to be processed at frame rate and the process should run indefinitely long. We refer to this task as *long-term tracking* and illustrate it in Fig. 1.

To enable long-term tracking, there are a number of problems which need to be addressed. The key problem is the detection of the object when it reappears in the camera's field of view. This problem is aggravated by the fact that the object may change its appearance, thus making the appearance from the initial frame irrelevant. Next, a successful long-term tracker should handle scale and illumination changes, background clutter, partial occlusions, and operate in real time.

The long-term tracking can be approached either from tracking or from detection perspectives. Tracking algorithms estimate the object motion. Trackers only require initialization, are fast, and produce smooth trajectories. On the other hand, they accumulate error during runtime (drift) and typically fail if the object disappears from the camera view. Research in tracking aims at developing increasingly robust trackers that track "longer." The postfailure behavior is not directly addressed. Detection-based algorithms estimate the object location in every frame independently. Detectors do not drift and do not fail if the object disappears from the camera view. However, they require an offline training stage and therefore cannot be applied to unknown objects.

The starting point of our research is the acceptance of the fact that neither tracking nor detection can solve the long-term tracking task independently. However, if they operate simultaneously, there is potential to benefit one from another. A tracker can provide weakly labeled training data for a detector and thus improve it during runtime. A detector can reinitialize a tracker and thus minimize the tracking failures.

The first contribution of this paper is the design of a novel framework (*TLD*) that decomposes the long-term tracking task into three subtasks: tracking, learning, and detection. Each subtask is addressed by a single component and the components operate simultaneously. The tracker follows the object from frame to frame. The detector localizes all appearances that have been observed so far and corrects the tracker if necessary. The learning estimates detector's errors and updates it to avoid these errors in the future.

While a wide range of trackers and detectors exist, we are not aware of any learning method that would be suitable for the TLD framework. Such a learning method should: 1) deal with arbitrarily complex video streams where the tracking failures are frequent, 2) never degrade the detector if the video does not contain relevant information, and 3) operate in real time.

To tackle all these challenges, we rely on the various information sources contained in the video. Consider, for instance, a single-patch denoting the object location in a single frame. This patch defines not only the appearance of the object, but also determines the surrounding patches, which define the appearance of the background. When tracking the patch, one can discover different appearances of the same object as well as more appearances of the background. This is in contrast to standard machine

- *Z. Kalal and K. Mikolajczyk are with the Centre for Vision, Speech, and Signal Processing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom. E-mail: zdenek.kalal@gmail.com, K.Mikolajczyk@surrey.ac.uk.*
- *J. Matas is with the Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Karlovo namesti 13, 121 35 Praha 2, Czech Republic. E-mail: matas@cmp.felk.cvut.cz.*

Manuscript received 10 Mar. 2011; revised 24 Aug. 2011; accepted 18 Oct. 2011; published online 7 Dec. 2011.
Recommended for acceptance by V. Pavlovic.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference. IEEECS Log Number TPAMI-2011-03-0153.
Digital Object Identifier no. 10.1109/TPAMI.2011.239.