Tracking by Identification Using Computer Vision and Radio: Supplementary Material

Rok Mandeljc, Stanislav Kovačič, Matej Kristan and Janez Perš

October 22, 2012

Abstract

Supplementary material for this paper consists of five video files, which, together with the commentary in this document, aim to illustrate the results from Section 6.3 in the paper. Video files are encoded using H.264 video codec and Matroska (MKV) container. To view them, please make sure that the video codec is installed on your system or use a video player with out-of-the-box support for it, e.g. VLC[1].

Introduction

Each video shows visualization of tracking results for a system, obtained on frames 3721-7761 of the dataset that we use in our evaluation[2] (see the paper for details). In all videos, top view of the room is shown. *Colored diamonds* denote manually-annotated *ground-truth points*, while *colored circles* denote identified detection hypotheses made by the system. For better impression of the movement, 25-frame trails are drawn behind detections and ground-truth points, denoted by *colored crosses* and *x-signs*. For additional video-specific annotations, see below.

Video #1: radio-based localization and tracking

In this video, results for the Ubisense *radio-based localization system* are visualized. As can be seen, the detections have strong (reliable) identitiy information, but their localization is rather poor. However, it is worth noting that most of the time, the spatial configuration of detections corresponds to that of ground-truth points. The trails belonging to detections (colored x-marks) are relatively sparse compared to trails of ground-truth points (colored crosses), which is due to fact that radio tags have update frequency of around 5 *Hz*, whereas ground-truth data has update frequency of 20 *Hz*.

Video #2: camera-based identification by tracking (25 cm)

In this video, results for state-of-the-art *computer-vision-based identification by tracking* are visualized. *Gray crosses* denote anonymous detections, which are produced

[1] `http://www.videolan.org/vlc/index.html`
[2] `http://vision.fe.uni-lj.si/research/mvl_lab5/`

by the detection algorithm and serve as the input to the tracker. The resolution of the underlying occupancy map is 25 *cm*. Each trajectory is assigned an identity of the person on whom it was initialized; after that, identities are propagated along the trajectories. This leads to propagation of identity switches, with first one occuring as early as 0:05, when identities of persons #4 and #5 are switched. After 0:34, all identities are switched and largely remain so for the rest of the sequence; any correction is purely by chance. At 1:18, there is a very notable example of a tracker drifting around (in this case tracker with identity #1, the green circle). This is caused by prolonged period of a missing detection (note the absence of gray cross near the ground-truth point) and an occurrence of a phantom detection during it. It can be seen that during normal operation, the computer-vision-based system offers much better localization than radio-based one, but has issues with propagated identity switches. Note that the update frequency of detections is 20 *Hz*, same as that of the ground-truth data. However, detections' trails are still a bit sparser than those of ground-truth points, due to limited resolution of the underlying occupancy map (25 *cm*).

Video #3: proposed tracking by identification (25 cm)

In this video, results for the *proposed tracking by identification* are visualized. *Gray crosses with numbers* denote identified detections produced by the second stage of our fusion, on top of which separate instances of tracker are run. The resolution of the underlying occupancy map is 25 *cm*. As same tracker is used as for pure computer-vision-based system, there is still drifting when input detections are missing. When individuals come close together, their identities might end up being switched (e.g. individuals #1 and #2 at 0:54). However, when they disperse again, their identities are correctly re-established. The proposed system does not propagate the identity switches.

Video #4: camera-based identification by tracking (10 cm)

This video shows the visualization of results for the state-of-the-art *computer-vision-based identification by tracking*, this time with underlying occupancy map resolution of 10 *cm*. As a result of higher grid resolution, trails of detections are less sparse. Additionally, trajectories are less jittery, and somewhat less prone to identity switches, although they still occur. The first one occurs at 0:32 (individuals #1 and #2), although it is corrected (by chance) at 1:01. Additional propagations of identity switches occur at 1:20, and at 2:42 and 3:15; at the end of the sequence, four identities are switched.

Video #5: proposed tracking by identification (10 cm)

This video shows the visualization of results for the *proposed tracking by identification*, this time with underlying occupancy map resolution if 10 *cm*. Denser grid results in better localization and less jittery trajectories, and relying on identified detections successfully prevents propagation of identity switches.