Workshop on "RGB-D: Advanced Reasoning with Depth Cameras"
In conjunction with Robotics: Science and Systems (RSS'11)
Los Angeles, USA; 2011

Learning to Detect and Track People in RGBD Data

Matthias Luber Luciano Spinello Kai O. Arras

Extended Abstract

Introduction

People detection and tracking is an important and fundamental component for many robots, interactive systems and intelligent vehicles. Previous works have used cameras and 2D and 3D range finders for this task. In this paper, we present a 3D people detection and tracking approach using RGB-D data. Given the richness of the data, we learn target appearance models for the purpose of improved detection and data association. This is a new aspect for range-based target tracking which usually deals with objects of identical appearance. To this end, we take an on-line boosting approach similar to [2] and learn a strong target classifier based on three types of RGB-D features. This results in an on-line people detector which is combined with a novel people detector based on a generic learned model, both integrated into a multi-hypothesis tracking system. The new generic detector, called Combo-HOD, fuses the image and depth information in the data. It is composed of the HOG detector (Histogram of Oriented Gradients) [1] applied on the RGB image and a newly introduced HOD approach (Histogram of Oriented Depths), inspired from the former, applied on the depth image. HOD locally encodes the direction of depth changes and relies on an depth-informed scale-space search that leads to a 3-fold acceleration of the detection process. Combo-HOD is general in that it neither relies on background learning nor on a ground plane assumption. For the evaluation we collect RGB-D data in a populated indoor environment with a setup of three Microsoft Kinect sensors with a joint field of view. The experiments demonstrate reliable 3D detection and tracking of people in RGB-D data up to $8\,m$ from the sensor and further show how the on-line detector improves the overall tracking performance.

This paper advances the state-of-the-art in the following aspects. First, we address the novel problem of detecting people in RGB-D data in distances far beyond the recommended sensor operating range (called *adequate play space*, see Fig. 2), second, we perform tracking of people in 3D data with a multi-hypothesis tracker (MHT), and third, we propose an online-learning method of target appearances along with its integration into the MHT.

A. Detection of People in 3D Range Data

The HOG detector belongs to the most successful methods for finding people in images. Opposed to visual HOG, HOD descriptors are computed in the *depth image* and encode directions of depth changes based on a physically grounded sensor model that accounts for the hyperbolic depth resolution loss in the Kinect sensory data. HOD further relies on an depth-informed scale-space search using integral images over scales that leads to a dramatic reduction of the search at improved detection rates. Finally, we propose Combo-HOD that fuses the RGB- and D-sensor modalities via a weighted mean of the likelihoods obtained by a sigmoid fitted to the SVM outputs. See Fig. 1, top, for example detections.

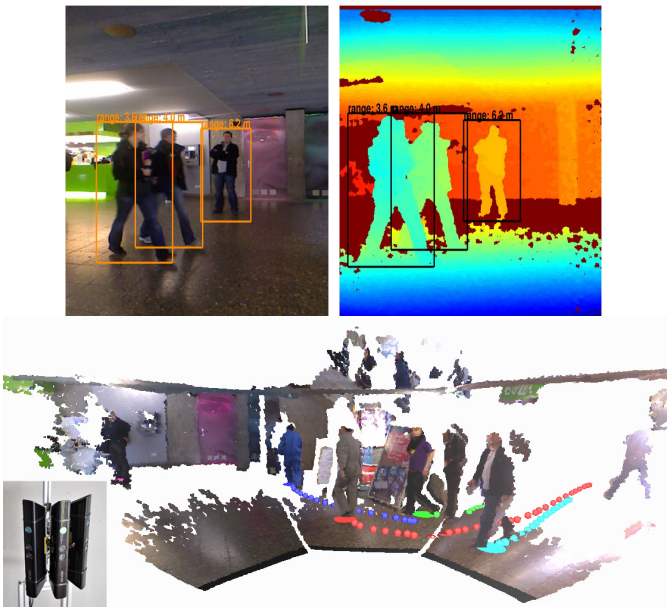

Fig. 1. **Top:** People detected in RGB-D data: dense depth data and color image data are simultaneously used for detecting a group of people. Note that the method neither relies on background learning nor on a ground plane assumption. **Bottom:** People tracking in RGB-D data by using a triple Kinect setup. The colored circles and dots in the point cloud show the positions and trajectories of five tracked persons.

B. Tracking with Online-Boosted Target Models

For various tasks such as motion planning among people or human activity recognition, the robot requires more than single-frame detections but motion trajectories of the surrounding people. To this end, we employ the MHT framework to produce full 3D estimates from the Combo-HOD detections. The richness of the RGB-D data enables us to learn target-specific appearance models that we will use to improve both detection and data association.

We take an on-line boosting approach for this purpose based on three types of RGB-D features namely Haar-like features and Lab color features in the RGB image and

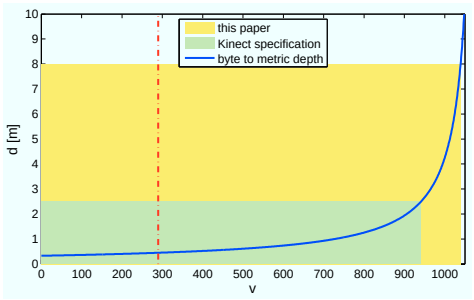

Fig. 2. Kinect depth characteristics. The blue line is the function that relates the byte values of the range image to metric depth, the red line is the sensor's minimal measurable depth. The dark green area indicates the *adequate play space* recommended in the Kinect User Manual, the yellow area is the range considered in this paper for detecting and tracking people. Notice that we address the problem of people detection at nearly 4× the suggested working range, where depth resolution is becoming increasingly coarser.

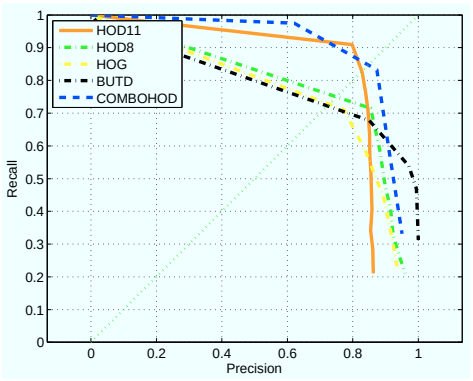

Fig. 3. Precision recall curves for depth-based, image-based, and combined RGB-D based detection methods. The most performant detector is Combo-HOD that combines both sensory cues. HOD is evaluated at different depth discretizations, 8bits and 11bits. HOD11 is the best depth-based detection method. Visual-based HOG detector underperforms due to unfavorable light conditions. BUTD underperforms due to the hyperbolic depth resolution loss of kinect data.

Haar-like features in the depth image, all at randomized positions and scales in the bounding box of each target.

We propose a novel strategy to integrate an online-learned detector with a generic (a priori) detector and a tracking system. The integration includes:

i) Unlike [2] where the new region is bootstrapped from the previous detection, we use the bounding box position of the generic detector to recenter the on-line detector. This avoids a key problem of on-line adaptation namely drifting of the model to background or clutter.

ii) The on-line classifier adds an appearance likelihood for data association that expresses how much the observed target's appearance matches the learned model. We thus have a joint likelihood that accounts for both motion state and appearance.

iii) In each cycle, the MHT assigns measurements to tracks and interprets measurements as new tracks or false alarms and tracks as occluded or deleted. This information is used to either initialize a new strong classifier in case of a new track, to update and recenter the on-line classifier in case of a match, stop on-line adaptation in case of an occlusion, and fill gaps of missing detections from the generic detector.

I. Experiments

We collected and annotated a large-scale indoor data set with unscripted behavior of people. The data set has been taken in the lobby of a large university canteen at lunch time using a sensory setup with three Kinect sensors as shown in Fig. 1. A total of 1648 instances of people in 1088 frames and 31 tracks have been labeled. The data set will be made publicly available.

For the evaluation of the generic people detector, we compared Combo-HOD with four alternative approaches: visual HOG on the RGB data, two versions of HOD (HOD8 using 8 bits to encode depth, and HOD11 using the full Kinect resolution of 11 bits), and a recently proposed people detector for 3D range data (BUTD) [3] that has been proven very accurate for sparse point clouds. Combo-HOD achieves an equal error rate (EER) of 85% and outperforms all other approaches with EERs of 73% for visual HOG, 75% for HOD8, 83% for HOD11, and 72% for BUTD. For the evaluation of the computational speedup over a non-informed (pyramidal) scale space search, we reached a 3-fold acceleration thanks to a significant reduction in tested scales and positions. The BUTD approach expectedly did not perform well due to the resolution loss of the sensory data at large distances that heavily compromises the geometric details of the point cloud. At close range, with good data quality, BUTD performs as well as HOD11, the best range-only method.

To assess the impact of the on-line boosting onto the tracking performance we run the tracker with adaptation and with Combo-HOD only. The on-line detector and its integration strategy leads to an improvement in the MOTA performance metric (from 58% to 63%) and a reduction of the numbers of track identifier switches (from 42 to 38). These improvements are mainly due to a reduction of the number of misdetections made possible by the gap-filling capacity of the integrated on-line detector and the data association guided by the joint likelihood.

Future work will focus on the collection and annotation of more RGB-D data sets and an extension of the on-line detector towards joint learning of target models over all tracks to even better distinguish them from each other.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.

[2] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR'06*.

[3] L. Spinello, M. Luber, and K. O. Arras. Tracking people in 3D using a bottom-up top-down people detector. In *ICRA'11*.