Close-Range Human Detection for Head-Mounted Cameras

Dennis Mitzel
mitzel@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group
RWTH Aachen University
Aachen Germany

Motivation. Robust multi-person detection and tracking is an important prerequisite for many applications. Examples include the use of mobile service robots in busy urban settings or mobile AR applications such as Google's project Glass. In this paper, we address the problem of stereo based person detection from the perspective of a moving human observer wearing a head-mounted stereo camera system. From this viewpoint many pedestrians in a crowded scenario are only partially visible due to occlusions at the image boundaries. In such situations, standard full-body object detectors such as [3] are not well-suited, since they cannot deal with the large degree of occlusion. On the other hand, we can take advantage of the elevated viewpoint of a head-mounted camera, which typically leaves the head-shoulder region of close-by pedestrians well visible.

Taking inspiration from a recently proposed human upper body detector for Kinect RGB-Depth data by [2], this paper proposes an improved stereo depth-template based approach which can quickly and reliably detect close-by pedestrians. Similar to [1] we generate regions of interest (ROIs) based on the stereo data in order to reduce the search space of the detector. Our approach learns a continuous normalized depth template from annotations of human upper bodies and slides this template over the extracted depth ROIs at several scales in order to compute a normalized distance score. The output of this process are distance matrices whose entries represent the distance between the template and the overlayed segment of the ROI for each scale. After non-minimum suppression (NMS) in the distance matrices we obtain several detections (from different scales) for a person that are pruned to a set of final detections by a second, template-based NMS stage. We systematically evaluate this approach and characterize the effects of its parameters. In addition, we show how it can be integrated into a mobile multi-person tracking framework.

Approach. In Fig. 1 we illustrate a compact overview of our proposed detection and tracking framework. For each new frame, given the stereo pair and the corresponding depth map, we project the 3D points onto a (automatically estimated) ground plane and extract the ROIs using connected components on the ground projection image. For each extracted 3D ROI, we generate the corresponding ROI in the image plane by back-projecting the ROIs from the ground plane to the image. The 2D ROIs are passed to the detector, which slides the learned upper body template over the ROIs and computes the distance matrix by taking the Euclidean distance between the template and the overlayed, normalized depth image segment. Using a minimum filter on the distance matrix, we obtain possible bounding box hypotheses for the upper bodies. These hypotheses are further pruned to a final detection set by using a template based intersection-over-union (IoU) NMS stage, where the detection with the lowest distance is chosen first and all other detections within a certain overlap area are removed iteratively.

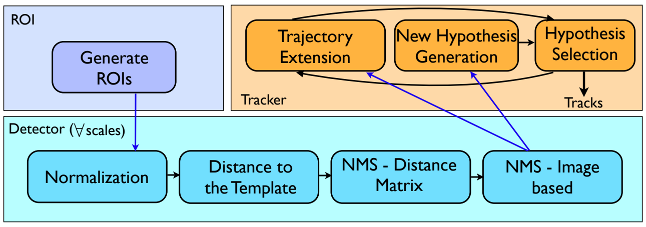

Figure 1: Overview of the different modules of the proposed approach.

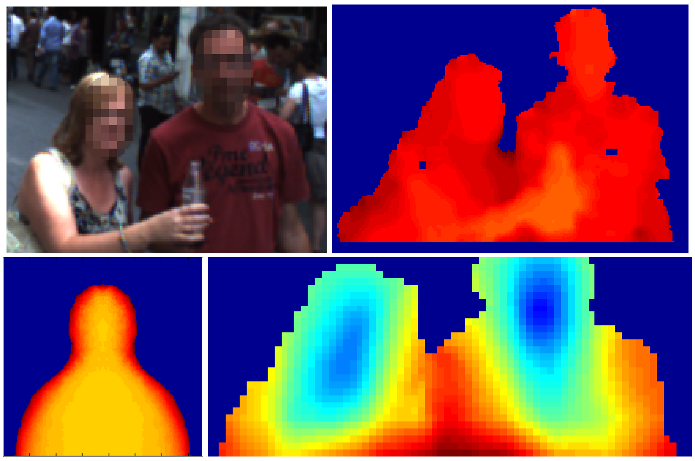

Figure 2: (upper left) Original ROI from the left image. (upper right) Input depth from the ROI. (bottom left) Depth template learned based on 600 upper body annotations. (bottom right) The resulting distance matrix for the initial scale.

Depth Template Detector. The pipeline for the detector consists of the following steps. First, for each ROI in the image plane, we discard the pixels which are not in the depth range of the ROIs in 3D by setting them to zero, as illustrated in Fig. 2. Then starting from an initial template size that is one third of the ROI height, we slide the template over the ROI. At each position, the segment of the ROI that is overlayed with the template is normalized with its median depth and then the distance between the template and the segment is computed. As a final result we obtain a distance matrix that contains for each position of the template the corresponding distance to the segment in the depth image, see Fig. 2 (bottom right). For Multi-Scale Handling we need to rescale the template and slide it over ROI again, because the initial scale estimation based on the height of the 2D ROI might not be representative for all pedestrians in the group. The multi-scale approach introduces several additional detections on a person for a number of neighboring scales, as the scale stride is usually small. To reduce this set to only one representative detection for each pedestrian, we perform an image based NMS.

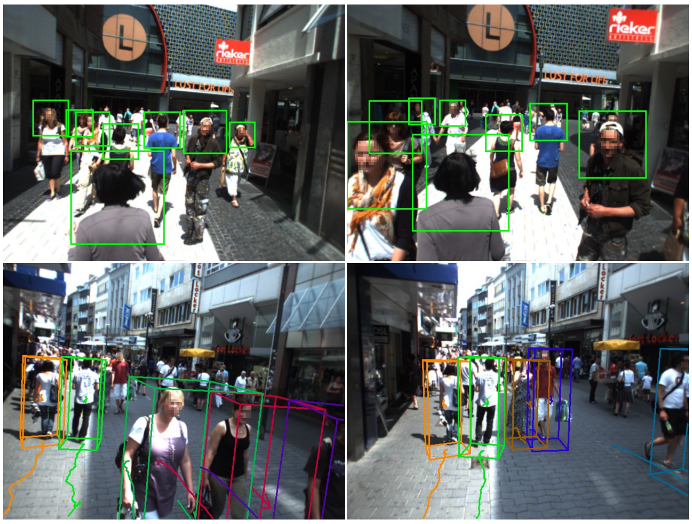

Figure 3: Experimental detection and tracking results

[1] M. Bansal, S. H. Jung, B. Matei, J. Eledath, and H. S. Sawhney. A real-time pedestrian detection system based on structure and appearance classification. In *ICRA*, 2010.

[2] W. Choi, C. Pantofaru, and S. Savarese. Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion. In *CORP*, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.