Probability Occupancy Maps for Occluded Depth Images

Timur Bagautdinov[1], Francois Fleuret[2], Pascal Fua[1].
[1]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. [2]IDIAP Research Institute, Switzerland.

The advent of the original Kinect camera [4] and its sucessors has sparked a tremendous regain of interest for RGB-D sensors, which were formerly perceived as being either cumbersone or expensive. They have been used with great success for motion capture [7, 8] and are becoming increasingly popular for people detection in robotics applications [5, 6]. However, the former requires the algorithms to be trained on very large training databases, which may not always be easy to create, to achieve the desired level of performance while the latter usually do not take into account possible occlusions between people.

In this paper, we propose an approach that relies on a generative model to evaluate the probability of target objects being present in the scene while explicitly accounting for occlusions, which prevents such failures. It is inspired by an earlier approach to estimating these probabilities from background subtraction results from multiple cameras with overlapping fields of view [3]. In the paper, we use a single depth map and approximate probabilities of occupancy at separate locations by choosing these probabilities so that the lower bound on the model likelihood is maximized (by using variational methods). In contrast to many other approaches, ours does statistical reasoning jointly, that is, knowledge about one piece of image evidence helps us to reason about the rest. This allows in particular to properly infer the presence of a severely occluded target from the presence of a small fragment.

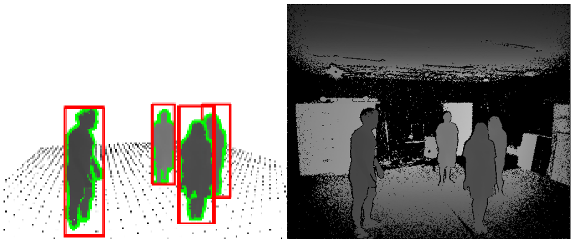

(a)
(b)

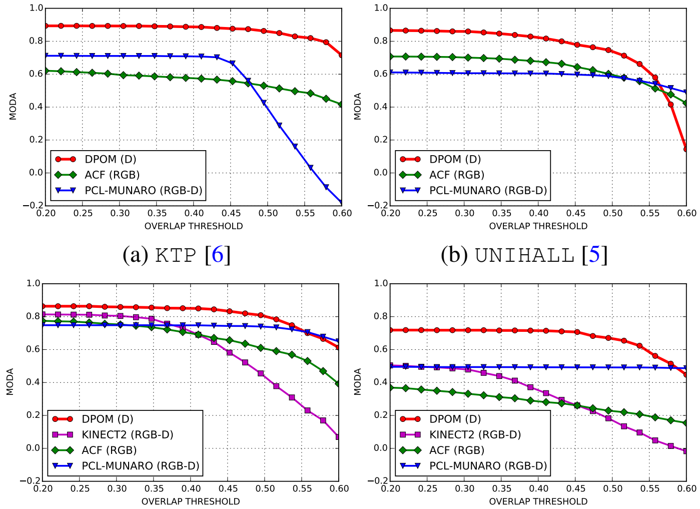
Figure 1: A generative model for depth maps. (a) Objects can be thought of as boxes, and images of objects are outlines within their rectangular projections. (b) Background is modeled explicitly, with a depth distribution for each pixel.

In our model, we consider a finite number of locations on the ground. An object of interest, e.g. a pedestrian, located at one of these, is represented by a flat free-shape inside a rectangular bounding box (see Fig. 1). In practice, with each location $k \in \mathcal{K}$ we therefore associate two random variables. The first is a Boolean X_k that denotes the presence or absence of the object at a location k. The second, a Boolean mask M_k, represents the 2D contour of that object, i.e. object segmentation. We then model the measured depths Z_i at each pixel $i \in \mathcal{L}$ in the image as conditionally independent given these variables, and distributed around the mean depth of the *closest object*, or according to the *background distribution* (bg) if no object is present:

$$l^* = \operatorname*{arg\,min}_{l:\{X_l=1,i\in M_l,l\in\mathcal{K}\}\cup\{\mathrm{bg}\}} \langle \theta_{l,i}(z|z\neq z^\infty)\rangle, \qquad (1)$$

$$Z_i \sim \theta_{l^*,i}(z). \qquad (2)$$

where $\theta_{k,i}(z)$ are distributions over depth for a specific location and pixel, $\langle\cdot\rangle$ denotes expectation and z^∞ encodes missing data. We use variational inference to get estimates of the posterior over object locations X_k. For M_k, a point estimate is obtained by a simplistic segmentation procedure. In the paper, we provide a complete formal description of our model, as well as details on inference and implementation.



(a) KTP [6]
(b) UNIHALL [5]
(c) EPFL-LAB
(d) EPFL-CORRIDOR

Figure 2: MODA scores on various datasets.

We evaluate our method on two available datasets [5, 6], as well as our own challenging sequences. In Fig. 2, we demonstrate performance comparison of our algorithm (DPOM) with multiple baselines (RGB detector, ACF [2], RGB-D detector, PCL-MUNARO [6], latest Kinect, KINECT2 [4]) in terms of MODA-scores [1]. In the paper, we demonstate that our method is able to outperform all of the baselines, even though it does not require neither large amounts of training data nor RGB signal.

The conclusion is that, with a use of an explicit generative model, and joint reasoning about the evidence, our method demonstrates state-of-the-art performance for pedestrian detection in depth maps, while being flexible enough to be used with a different object type, which we demonstrate on quadrocopters.

Acknowledgment

This work was supported in part by a Swiss National Science Foundation grant.

[1] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: the Clear Mot Metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008.

[2] P. Dollár, R. Appel, and W. Kienzle. Crosstalk Cascades for Frame-Rate Pedestrian Detection. 2012.

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-Camera People Tracking with a Probabilistic Occupancy Map. 30(2):267–282, February 2008.

[4] Kinect. Kinect for Windows SDK 2.0, 2014. http://www.microsoft.com/en-us/kinectforwindows/.

[5] M. Luber, L. Spinello, and K.O. Arras. People Tracking in Rgb-D Data with On-Line Boosted Target Models. pages 3844–3849, 2011.

[6] M. Munaro and E. Menegatti. Fast RGB-D People Tracking for Service Robots. *Autonomous Robots*, 37(3):227–242, 2014.

[7] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake. Real-Time Human Pose Recognition in Parts from a Single Depth Image. 2011.

[8] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM*, 56(1):116–124, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.