# Tracking Revisited using RGBD Camera: Unified Benchmark and Baselines

Shuran Song Jianxiong Xiao
Princeton University

## Abstract

*Despite significant progress, tracking is still considered to be a very challenging task. Recently, the increasing popularity of depth sensors has made it possible to obtain reliable depth easily. This may be a game changer for tracking, since depth can be used to prevent model drift and handle occlusion. We also observe that current tracking algorithms are mostly evaluated on a very small number of videos collected and annotated by different groups. The lack of a reasonable size and consistently constructed benchmark has prevented a persuasive comparison among different algorithms. In this paper, we construct a unified benchmark dataset of 100 RGBD videos with high diversity, propose different kinds of RGBD tracking algorithms using 2D or 3D model, and present a quantitative comparison of various algorithms with RGB or RGBD input. We aim to lay the foundation for further research in both RGB and RGBD tracking, and our benchmark is available at http://tracking.cs.princeton.edu.*

**Motivation and background** Visual object tracking is an important but challenging task. For example, in the standard tracking-by-detection pipeline, a slight offset in one frame may be reinforced after an online learning step, resulting in the so-called model drift problem. Besides, occlusion of target objects occurs quite often in real world scenarios. To address these issues, over the last decade, object tracking algorithms have evolved significantly in both their sophistication and quality of results. However, all these approaches are evaluated on a very small number of videos collected and annotated by different groups over the years (*e.g.* 8 videos used for evaluating). There is no consistent evaluation metric, especially when occlusion happens. Furthermore, all ground truth annotation is publicly available, which makes it even worse in terms of parameter overfitting. Many practitioners in the field find that it is hard to generalize some of these approaches to other videos because of parameter sensitivity. The lack of a reasonable size and consistently constructed benchmark for tracking has been preventing persuasive comparisons.

Figure 1. Examples from our Princeton Tracking Benchmark.

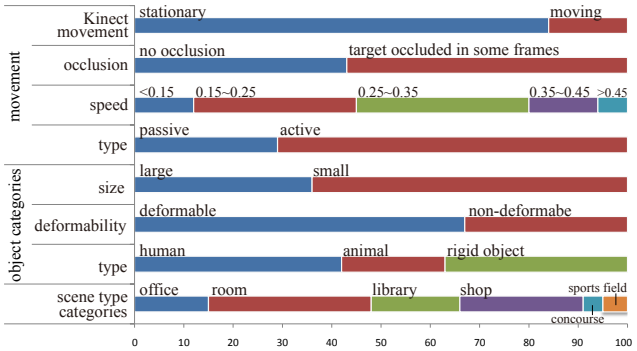
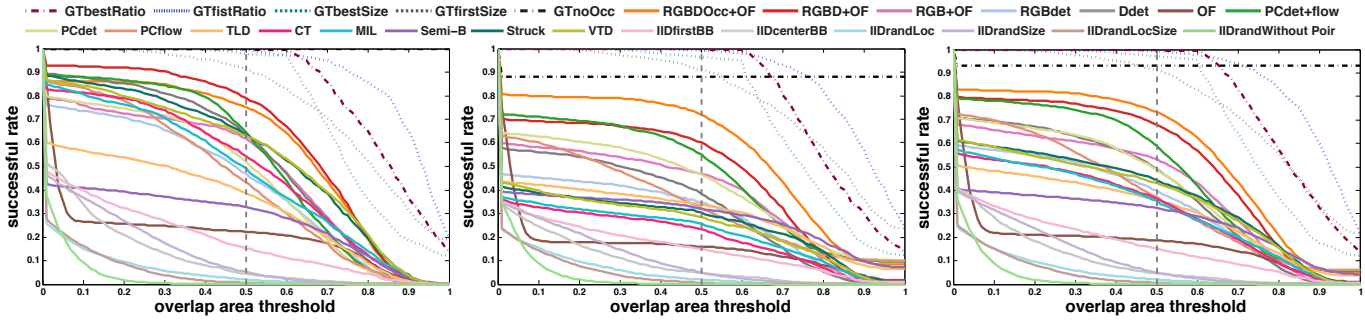



Figure 3. Statistics of our RGBD tracking benchmark dataset.

Meanwhile, great popularity of affordable depth sensors, such as Microsoft Kinect, make depth acquisition very easy. Reliable depth maps can provide valuable additional information to significantly improve tracking results with robust occlusion and model drift handling. How much does depth information help in tracking? Will the availability of depth significantly change the design of the standard tracking pipeline? What is a reasonable baseline algorithm for tracking with RGBD data? And how do the state-of-the-art RGB tracking algorithms perform compared with these new RGBD algorithms? This paper seeks to answer these




(a) Test cases without occlusion (b) Test cases with occlusion (c) All test cases

Figure 2. Average success rate vs. threshold of overlap ratio ($r_t$) evaluated on different categories of test cases.

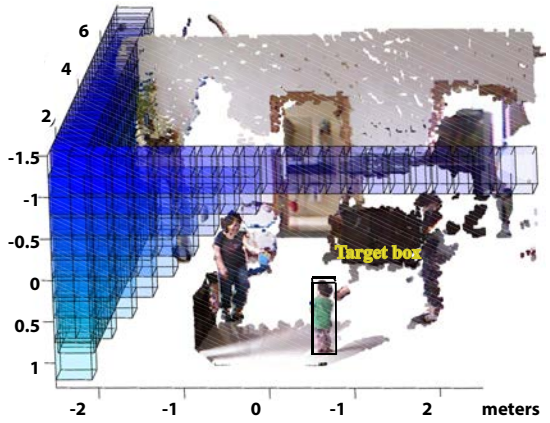



Figure 4. 3D point cloud with 3D sliding window detection.

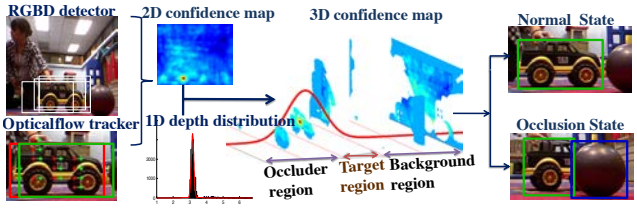



Figure 5. RGBD tracking algorithm based on 2D image patch, and occlusion handling algorithm

questions by conducting a quantitative benchmark evaluation, and proposing various simple but powerful baseline algorithms.

**Unified tracking benchmark** To establish a unified benchmark, we construct a RGBD dataset of 100 videos captured using a standard Microsoft Kinect 1.0 and manually annotate the ground truth of all frame. Our dataset includes deformable objects, various occlusion conditions, moving camera, and different scenes. Figure 1 summarize the statistics of our dataset. In our on-line evaluation system, 5 videos out of 100 are used for parameter tuning, and the remaining 95 are used for evaluation.

**Baseline algorithms** To build a set of diverse baseline algorithms, we design several tracking algorithms incorporating depth information to reduce model drift, including traditional 2D image patch based tracking with additional depth features, 3D point cloud based tracking algorithm and other baseline lagorithms use optical flow or ICP. We also design a simple occlusion handling algorithm based on the depth map. Overview of the 2D and 3D based algorithms is shown in Figure 5 and 4 respectively. To understand the impact of model assumptions, we use the ground truth to design several performance upper bounds under different model assumptions, such as fixed box size, aspect ratio, or target being always visible . To understand the impact of dataset bias to the evaluation, we also obtained performance lower bounds from several trivial image-independent algorithms as preformence lower bounds.

**Result and conculsion** Figure 2 measures the success rate $R$ while varying the threshold $r_t$. Compared with other state-of-the-art RGB trackers: TLD[4], CT[6], MIL[1], semi-B[2], Struck[3] and VTD[5] the results demonstrate that by incorporating depth data, trackers can achieve better performance and handle occlusion much more reliably. We hope that our unified benchmark provides new insights to the field, by making experimental evaluation more standardized and easily accessible.

## References

[1] B. Babenko, M.-H. Yang, and S. Belongie. Visual Tracking with Online Multiple Instance Learning. In *CVPR*, 2009. 2

[2] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *ECCV*, 2008. 2

[3] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011. 2

[4] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *PAMI*, 2012. 2

[5] J. Kwon and K. M. Lee. Visual tracking decomposition. In *CVPR*, 2010. 2

[6] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *ECCV*, 2012. 2