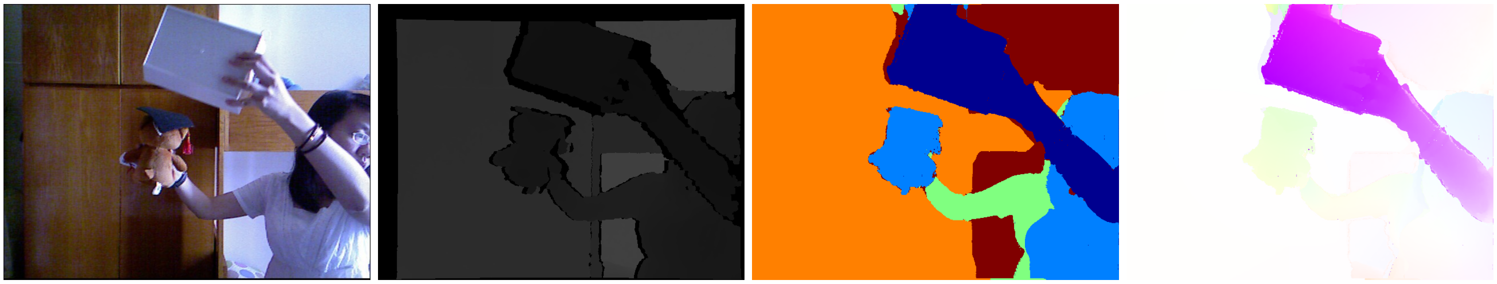
# Layered RGBD Scene Flow Estimation

Deqing Sun[1], Erik B. Sudderth[2], Hanspeter Pfister[1]
[1]School of Engineering and Applied Sciences (SEAS), Harvard University. [2]Department of Computer Science, Brown University.

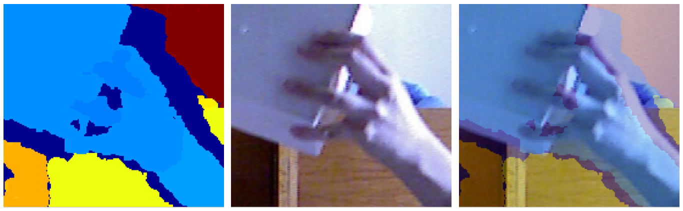


(a) First RGB image (b) First depth (c) Segmentation (d) 2D motion

Figure 1: Our layered approach can well handle multiple independently moving objects and reliably estimate their motion. The depth map provides the depth ordering of layers and solves a computational bottleneck for layered models. (Please see the main paper for our detected occlusions and the estimated motion by a comparison method).

**Introduction.** We address the problem of segmenting a scene into moving layers and estimating the 3D motion of each layer from an RGBD image sequence. Although the depth information simplifies some common challenges in vision, its noisy nature brings new challenges for RGBD scene flow estimation. In particular, the depth boundaries are not well-aligned with RGB image edges, as shown in Figure 2. As a result we cannot accurately localize the 2D motion boundaries using depth. Furthermore, depth is missing around occlusion and disocclusion regions. Occlusions are a major source of errors to both the 2D optical flow formulation and its 3D extension to the RGBD data.

Layered models are a promising approach to model occlusions in RGB image sequences [7]. Recent RGB layered methods [5, 6] have obtained promising results in the presence of occlusions, but are computationally expensive. One bottleneck is to infer the depth ordering of layers, which often requires searching over a combinatorial space ($K!$ possibilities for $K$ layers). We propose using the depth information from RGBD data to solve the depth ordering problem for layered models. Our work is based on a surprisingly simple observation that *depth determines depth ordering*, as shown in Figure 1. The depth further allows us to estimate the 3D rotation and translation for each independently moving layer. The resultant 3D rigid motion serves as a stronger prior to constrain the per-layer motion than pairwise Markov random field (MRF) models.



(a) Color-coded depth (b) Image (c) Overlaid

Figure 2: Crop of region around the left hand in Figure 1. The depth boundaries are not well-aligned with object boundaries in the 2D image plane. However, the noisy depth is sufficient to decide the depth ordering of the foreground and the background.

**Model.** Given a sequence of images and depth $\{\mathbf{I}_t, \mathbf{z}_t\}, 1 \le t \le T$, we want to segment the scene into moving layers and estimate the motion for each layer. We assume that the scene consists of $K$ independently moving layers, ordered in depth [3, 4]. To model the layer segmentation, we use $K-1$ continuous support functions for the first $K-1$ layers. The support function for the $k$th layer at frame $t$, $\mathbf{g}_{tk}$, encodes how likely every pixel is visible in that layer at frame $t$. The $K$th layer is background and has no support function. The motion for each layer includes both the 2D motion in the image plane $\{\mathbf{u}_t, \mathbf{v}_t\}$ and the change in depth $\mathbf{w}_t$.

Probabilistically we want to solve for the most likely layer support functions and motion fields given the image evidences. We can decompose the posterior distribution using Bayesian rules as

$$\begin{aligned} p(\mathbf{u}_t, \mathbf{v}_t, \mathbf{w}_t, \mathbf{g}_t | \mathbf{I}_t, \mathbf{I}_{t+1}, \mathbf{z}_t, \mathbf{z}_{t+1}, \mathbf{g}_{t+1}) \propto & \quad (1) \\ p(\mathbf{I}_{t+1} | \mathbf{I}_t, \mathbf{u}_t, \mathbf{v}_t, \mathbf{g}_t) p(\mathbf{z}_{t+1} | \mathbf{z}_t, \mathbf{u}_t, \mathbf{v}_t, \mathbf{w}_t, \mathbf{g}_t) & \\ p(\mathbf{u}_t, \mathbf{v}_t, \mathbf{w}_t | \mathbf{g}_t, \mathbf{I}_t) p(\mathbf{g}_{t+1} | \mathbf{g}_t, \mathbf{u}_t, \mathbf{v}_t) p(\mathbf{g}_t | \mathbf{I}_t), & \end{aligned}$$

where the first/second term describes how the next image/depth depends on the current image/depth, the motion, and the layer support. The third term uses semi-parametric models to capture both the global behavior and local deviations of the scene flow, the fourth term describes how the layer support non-rigidly evolves over time, and the last term uses conditional random field to describe how the layer support depends on the RGB images.

**Result.** We evaluate our layered method using several datasets. On the widely used Middlebury benchmark, the average error for our approach is less than half of the state-of-the-art SRSF method [2]. On a variety of real sequences, our method obtains visually comparable or better results, particularly for scenes with several independently moving objects and large occlusions. To further innovations, we make our MATLAB code publicly available at the first author's webpage [1].

Table 1: Average root mean squared (RMS) error and average angular error (AAE) results on the Middlebury 2002 dataset.

| | Teddy | | Cones | |
|---|---|---|---|---|
| | RMS | AAE | RMS | AAE |
| Our method | **0.09** | **0.17** | **0.12** | **0.13** |
| SRSF [2] | 0.49 | 0.46 | 0.45 | 0.37 |

[1] http://people.seas.harvard.edu/~dqsun/.

[2] J. Quiroga, Thomas Brox, F. Devernay, and J. Crowley. Dense semi-rigid scene flow estimation from rgbd images. In *European Conference on Computer Vision*, 2014.

[3] E. Sudderth and M. Jordan. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Neural Information Processing Systems*, pages 1585–1592, 2009.

[4] D. Sun, E. B. Sudderth, and M. J. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *Neural Information Processing Systems*, pages 2226–2234, 2010.

[5] D. Sun, E. B. Sudderth, and M. J. Black. Layered segmentation and optical flow estimation over time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1768–1775, 2012.

[6] D. Sun, J. Wulff, E. B. Sudderth, H. Pfister, and M. J. Black. A fully-connected layered model of foreground and background flow. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2451–2458, 2013.

[7] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE TIP*, 3(5): 625–638, September 1994.

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