Depth Extraction from Videos Using Geometric Context and Occlusion Boundaries

S. Hussain Raza[1]
hussain.raza@gatech.edu
Omar Javed[2]
omar.javed@sri.com
Aveek Das[2]
aveek.das@sri.com
Harpreet Sawhney[2]
harpreet.sawhney@sri.com
Hui Cheng[2]
hui.cheng@sri.com
Irfan Essa[3]
irfan@cc.gatech.edu

[1] School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA

[2] SRI International
Princeton, New Jersey, USA

[3] School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA

http://www.cc.gatech.edu/cpl/projects/videodepth

Abstract

We present an algorithm to estimate depth in dynamic video scenes. We propose to learn and infer depth in videos from appearance, motion, occlusion boundaries, and geometric context of the scene. Using our method, depth can be estimated from unconstrained videos with no requirement of camera pose estimation, and with significant background/foreground motions. We start by decomposing a video into spatio-temporal regions. For each spatio-temporal region, we learn the relationship of depth to visual appearance, motion, and geometric classes. Then we infer the depth information of new scenes using piecewise planar parametrization estimated within a Markov random field (MRF) framework by combining appearance to depth learned mappings and occlusion boundary guided smoothness constraints. Subsequently, we perform temporal smoothing to obtain temporally consistent depth maps. We present a thorough evaluation of our algorithm on our new dataset and the publicly available Make3d static image dataset.

1 Introduction and Approach

Methods exploiting visual and contextual cues for depth can be used to provide an additional source of depth information to the structure from motion or multi-view stereo based depth estimation systems. In this paper, we focus on texture features, geometric context, motion boundary based monocular cues along with co-planarity, connectivity and spatio-temporal consistency constraints to predict depth in videos. We assume that a scene can be decomposed into planes, each with its own planar parameters. We over-segment a video into spatio-temporal regions and compute depth cues from each region along with scene structure from geometric contexts. These depth cues are used to train and predict depth from features. However, such appearance to depth mappings are typically noisy and ambiguous. We incorporate the independent features to depth mapping of each spatio-temporal region within in a MRF framework that encodes constraints from scene layout properties of co-planarity, connectivity and occlusions. To model the connectivity and co-planarity in a scene, we explicitly learn occlusion boundaries in videos. To further remove the inconsistencies from temporal depth prediction, we apply a sliding window to smooth the depth prediction. Our approach doesn't require camera translation or large rigid scene for depth estimation. Moreover, it provides a source of depth information that is largely complementary to triangulation based depth estimation methods [5]. **The primary contributions of our method to extract depth from videos are:**

- Adoption of a learning and inference approach that explicitly models appearance to geometry mappings and piecewise scene smoothness;
- Learning and estimating occlusion boundaries in videos and utilizing these to constrain smoothness across the scene;
- There is no requirement of a translating camera or a wide-baseline for depth estimation;
- An algorithm for video depth estimation that is complementary to traditional structure from motion approaches, and that can incorporate these approaches to compute depth estimates for natural scenes;

2 Experiments and Results

We perform extensive experiments on video depth data to evaluate our algorithm. We perform 5-fold cross-validation over 36 videos ($\sim$ 6400 frames). We compute average log-error $|\log d - \log \hat{d}|$ and average relative error $|\frac{d-\hat{d}}{d}|$ to report the accuracy of our method. We achieve an accuracy of 0.153 log-error and 0.44 on relative error (Table 1). Figure 1 shows some example scenes from our dataset with ground truth and predicted depth. Our approach for depth estimation can also be applied to images. We applied our algorithm over a publicly available Make3d depth image dataset [6]. Table 2 gives the comparison of the single image variant of our approach with the state of the art and we achieve competitive results. It should be noted that our algorithm depends on occlusion boundary detection and geometric context (for which motion based features are important and is not optimized to extract depth from single images.

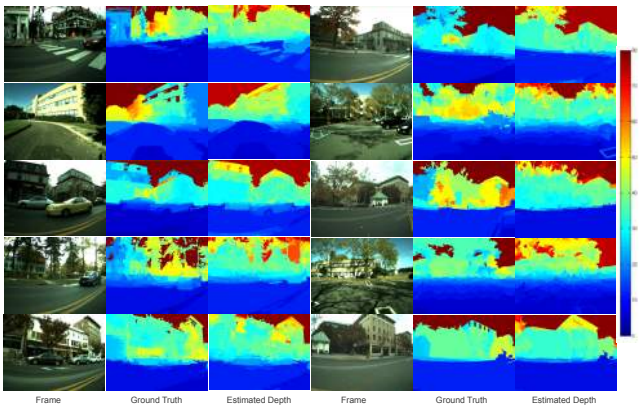

Figure 1: Examples of videos scenes, ground truth, and predicted depth by our method. Legend shows depth range from 0m (blue) to 80m (red).

Features	*log*10	rel-depth
ALL	0.153	0.44
App.+Flow	0.176	0.533
Appearance	0.175	0.512

Table 1: Performance of our algorithm on video dataset, combining appearance, flow, and surface layout features give best accuracy.

Algorithm	*log*10	rel-*log*
SCN [4]	0.198	0.530
HEH [1]	0.320	1.423
Baseline [6]	0.334	0.516
PP-MRF [6]	0.187	0.370
Depth Transfer [2]	0.148	0.362
Sematic Labels [3]	0.148	0.379
***Geom. Context Occl. Bound.**	**0.159**	**0.386**

Table 2: Our approach can also be applied to images. We apply it to Make3d depth image dataset [6].

[1] D. Hoiem, A.A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.

[2] Kevin Karsch, Ce Liu, and Sing Kang. Depth extraction from video using non-parametric sampling. In *ECCV 2012*.

[3] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[4] Ashutosh Saxena, Sung H Chung, and Andrew Ng. Learning depth from single monocular images. In *NIPS*, 2005.

[5] Ashutosh Saxena, Jamie Schulte, and Andrew Y Ng. Depth estimation using monocular and stereo cues. In *IJCAI*, 2007.

[6] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009.