Eurographics Symposium on Geometry Processing (2017)
J. A. Bærentzen and K. Hildebrandt (Editors)

DEPTHCUT: Improved Depth Edge Estimation Using Multiple Unreliable Channels

Paul Guerrero[1] Holger Winnemöller[2] Wilmot Li[2] Niloy J. Mitra[1]

[1]University College London [2]Adobe Research

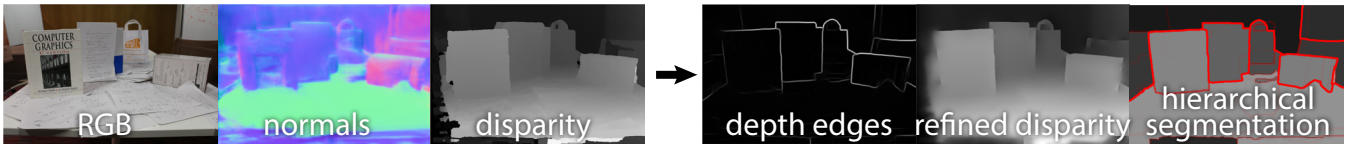

Figure 1: *We present* DEPTHCUT, *a method to estimate* depth edges *with improved accuracy from unreliable input channels, namely: RGB images, normal estimates, and disparity estimates. Starting from a single image or pair of images, our method produces depth edges consisting of depth contours and creases, that separate regions of smoothly varying depth. Complementary information from the unreliable input channels are fused using a neural network trained on a dataset with known depth. The resulting depth edges can be used to refine a disparity estimate or to infer a hierarchical image segmentation.*

Abstract

In the context of scene understanding, a variety of methods exists to estimate different information channels from mono or stereo images, including disparity, depth, and normals. Although several advances have been reported in the recent years for these tasks, the estimated information is often imprecise particularly near depth contours or creases. Studies have however shown that precisely such depth edges *carry critical cues for the perception of shape, and play important roles in tasks like depth-based segmentation or foreground selection. Unfortunately, the currently extracted channels often carry conflicting signals, making it difficult for subsequent applications to effectively use them. In this paper, we focus on the problem of obtaining high-precision depth edges by jointly analyzing such unreliable information channels. We propose* DEPTHCUT, *a data-driven fusion of the channels using a convolutional neural network trained on a large dataset with known depth. The resulting depth edges can be used for segmentation, decomposing a scene into segments with relatively smooth depth, or improving the accuracy of the depth estimate near depth edges by constraining its gradients to agree with these edges. Quantitative experiments show that our depth edges result in an improved segmentation performance compared to a more naive channel fusion. Qualitatively, we demonstrate that the depth edges can be used for superior segmentation and an improved depth estimate near depth edges.*

1. Introduction

A central task in scene understanding from a single image (mono) or pairs of images (stereo) is to extract information about scene geometry. Current methods to compute depth or normals from the input image(s) still suffer from imperfections (see columns 1-5 of Figure 2). In this work, instead of aiming for precise depth estimates, we focus on identifying depth contours and creases, which we refer to as *depth edges*. Studies [Gib86, BKP*13] have shown that precisely such depth edges carry critical cues for the perception of shapes, and play important roles in tasks like depth-based segmentation or foreground selection. Due to the aforementioned imperfections, current methods mostly produce poor depth edges. In contrast to absolute depth, depth edges often correlate with edges in other channels. Our main insight is that we can obtain better depth edges by fusing together multiple channels, each of which may, in isolation, be unreliable due to misaligned features, errors, and noise (see columns 5 and 6 of Figure 2). We propose a data-driven fusion of the channels using DEPTHCUT, a convolutional neural network (CNN) trained on a large dataset with known depth. We use the extracted depth edges for segmentation, decomposing a scene into segments with relatively smooth depth, or improving the accuracy of the depth estimate by constraining its gradients to agree with the estimated depth edges.

2. Method Overview

Starting from either mono or stereo images, we investigate fusing three different channels: color, estimated disparity, and estimated normals (see Figure 2), all of which are estimated from the input images using existing methods. These channels carry complimentary cues. For example, color channel carry good edge information, but fail to differentiate between depth and texture edges, while disparity channels allow us to make this distinction, but are unreliable near depth edges. Combining these channels is challenging, since different locations on the image plane require different combinations, depending on their context. Additionally, it is hard to formulate explicit rules how to combine channels.

We designed DEPTHCUT to combine these unreliable channels to obtain robust depth edges. The network fuses multiple depth cues in a context-sensitive manner by learning what channels to rely on in different parts of the scene. It is implemented as an autoencoder with skip connections that maps multiple unreliable input channels to a more reliable depth edge estimate. The network is trained on two

DOI: 10.2312/sgp.20171202

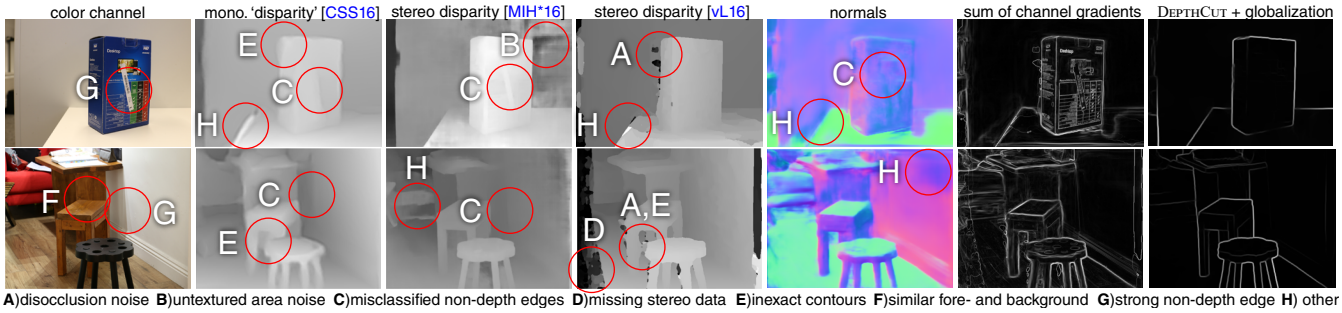

Figure 2: *Unreliable input channels. The channels we use as input for depth edge estimation contain various sources of noise and errors. We train* DEPTHCUT *to combine these channels for a cleaner set of depth edges, shown in the last column with an optional globalization that only retains the most salient edges.*

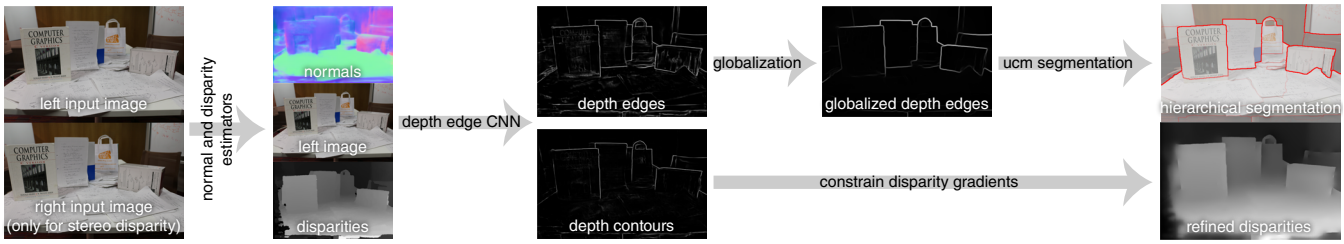

Figure 3: *Overview of our method and two applications. Starting from a stereo or mono image(s), we estimate our three input channels using existing methods and combine them in a data-driven fusion using our CNN to get a set of depth edges. These are used in two applications, segmentation and refinement of the estimated disparity. For segmentation, we also perform a globalization step that keeps only the most consistent contours.*

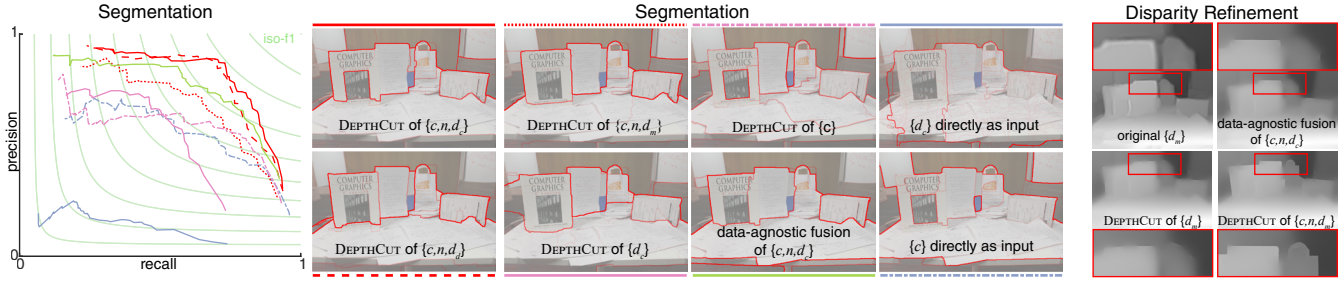

Figure 4: *Qualitative and quantitative comparisons to several baselines. Precision and recall is given for the baselines shown in the center using subsets of channels: color (c), normal (n), and disparity from several methods (monocular [CSS16] d_m, Dispnet stereo [MIH*16] d_d and MC-CNN stereo [vL16] d_c). Refinement of monocular disparity compared to three baselines is shown on the right. A data-driven fusion of multiple channels benefits both applications.*

datasets with known high-quality depth, the Middlebury 2014 Stereo dataset [SHK*14] and a custom synthetic indoor scenes dataset, with ground truth depth edges computed directly from the known depth. We use a mean-square loss, and add weighting to remove a bias in the training towards false positives caused by strong texture or shadow edges.

The estimated depth edges may be noisy and are not closed. For segmentation, we adapt the non-semantic segmentation method proposed by Arbeláez et al. [AMFM11] to get a set of closed contours. To refine a depth estimate with our edges, we constrain the depth gradient magnitude and direction using the estimated depth edge magnitude and direction. Figure 3 summarizes all steps.

3. Results

We compare to several baselines in Figure 4. The baselines use subsets of the channels as inputs and either perform no fusion, a data-agnostic fusion with simple manually crafted rules or our DEPTHCUT fusion. We conclude from these experiments that CNNs can be used for a data-driven fusion of noisy channels that generally performs better than single channels or a data-agnostic fusion.

References

[AMFM11] ARBELÁEZ P., MAIRE M., FOWLKES C., MALIK J.: Contour detection and hierarchical image segmentation. *IEEE PAMI 33*, 5 (2011), 898–916. 2

[BKP*13] BANSAL A., KOWDLE A., PARIKH D., GALLAGHER A., ZITNICK L.: Which edges matter? In *IEEE ICCV* (2013), pp. 578–585. 1

[CSS16] CHAKRABARTI A., SHAO J., SHAKHNAROVICH G.: Depth from a single image by harmonizing overcomplete local network predictions. In *NIPS* (2016), pp. 2658–2666. 2

[Gib86] GIBSON J. J.: *The Ecological Approach to Visual Perception*. Routledge, 1986. 1

[MIH*16] MAYER N., ILG E., HAUSSER P., FISCHER P., CREMERS D., DOSOVITSKIY A., BROX T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE CVPR* (2016), pp. 4040–4048. 2

[SHK*14] SCHARSTEIN D., HIRSCHMÃIJLLER H., KITAJIMA Y., KRATHWOHL G., NESIC N., WANG X., WESTLING P.: High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR* (2014), vol. 8753 of *Lecture Notes in Computer Science*, Springer, pp. 31–42. 2

[vL16] ŽBONTAR J., LECUN Y.: Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res. 17*, 1 (2016), 2287–2318. 2