IMAGE SEGMENTATION USING CONTOUR, SURFACE, AND DEPTH CUES

Xiang Fu[⋆] *Chen Chen*[⋆] *Jian Li*[⋆] *Changhu Wang*[†] *C.-C Jay Kuo*[⋆]

[⋆] University of Southern California
3740 McClintock Ave., Los Angeles, California, United States, CA 90089
[†] Toutiao AI Lab
No.43, North 3rd Ring West Road, Haidian District, Beijing, China, 100098

ABSTRACT

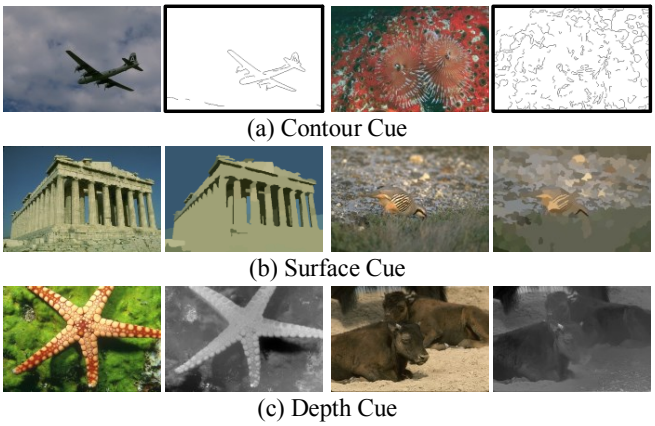
We target at solving the problem of automatic image segmentation. Although 1D contour and 2D surface cues have been widely utilized in existing work, 3D depth information of an image, a necessary cue according to human visual perception, is however overlooked in automatic image segmentation. In this paper, we study how to fully utilize 1D contour, 2D surface, and 3D depth cues for image segmentation. First, three elementary segmentation modules are developed for these cues respectively. The proposed 3D depth cue is able to segment different textured regions even with similar color, and also merge similar textured areas, which cannot be achieved using state-of-the-art approaches. Then, a content-dependent spectral (CDS) graph is proposed for layered affinity models to produce the final segmentation. CDS is designed to build a more reliable relationship between neighboring surface nodes based on the three elementary cues in the spectral graph. Extensive experiments not only show the superior performance of the proposed algorithm over state-of-the-art approaches, but also verify the necessities of these three cues in image segmentation.

Index Terms— Image Segmentation, Spectral Graph, Depth Estimation

1. INTRODUCTION

Automatic image segmentation is a fundamental problem in computer vision, and plays a significant role in diverse applications, such as object detection, and image retrieval. It automatically partitions an image into several (usually over two) disjointed coherent groups, which is different from figure-ground segmentation. The target is to achieve maximal inter-variance and minimal intra-variance of the clusters with a small number of segments, so that the segmentation result is as close to the understanding of humans as possible.

Most of existing algorithms on automatic image segmentation can be roughly classified into two categories, i.e., region-based and contour-based methods, depending on whether 2D surface cue or 1D contour cue plays a key role. Region-based methods find the similarity among spatially connected pixels and group them together using the surface properties like luma and chroma. Typical algorithms include Watershed [1], Normalized Cuts [2], Mean Shift [3], Felzenszwalb and Huttenlocher's graph-based method (FH) [4], Multi-Layer Spectral Segmentation (MLSS) [5], and Segmentation by Aggregating Superpixels (SAS) [6]. However, these methods neglect the obvious discontinuities between two regions that lead to segmentation boundaries. To resolve this problem, contour-based methods, like gPb-owt-ucm [7], were presented to find connected regions blocked by contours.



(a) Contour Cue

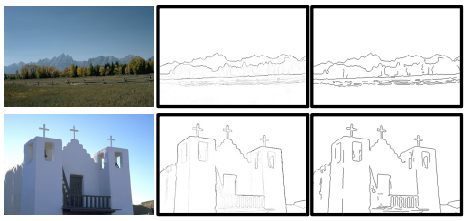
(b) Surface Cue

(c) Depth Cue

Fig. 1. Original images and their corresponding cue maps. The cue is useful for segmentation of the left image, but is harmful to segment the right image. (a) 1D contour cue. (b) 2D surface cue. (c) 3D depth cue.

To overcome the limitations of region-based and contour-based methods, researchers tried to combine both surface and contour properties for better segmentation. They either took the contour cue as post-processing to correct the results of region-based segmentation, or adopted the contours as barriers in the graph [8, 9]. In spite of the combination of 1D contour and 2D surface cues, the algorithms still fail in two categories of challenging cases. One is to separate object and environment that are too involved to find the boundaries for two adjacent regions using surface properties, especially when in the shadow (leakage problem). The other is to combine textured regions, which have high contrast inside and no clear contour outside the region (over-segmentation problem). These challenges motivate us to leverage 3D depth cue of an image to alleviate these two problems in image segmentation.

According to the study of human visual perception [10, 11], people are more concerned with dissimilarities or high contrast between two regions, and group similar regions in



(a) (b) (c)

Fig. 2. Illustration of 1D contour cue. (a) Original image. (b) Structured Edge Detection [12]. (c) 1D contour cue.

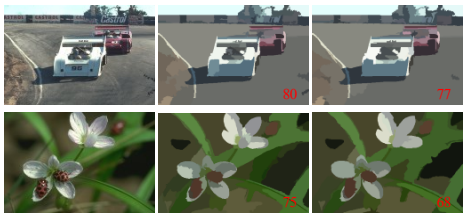
appearance and depth layer, corresponding to what we called 1D contour, 2D surface, and 3D depth cues. The contour cue describes discontinuities between two adjacent regions; the surface cue illustrates similarities inside one region; and the depth cue makes use of blur estimation to indicate the image layout even if some regions are apparently similar.

In this paper, we study how to fully utilize 1D contour, 2D surface, and 3D depth cues for automatic image segmentation. It is clear that although each cue might be sufficient to segment some images, each of them has its own scope and limitations in that the results are not consistently reliable. As shown in Fig. 1, the corresponding cue will be helpful for the left images to achieve good segmentation, while it will cause problems for the right ones. We found that the contour cue is more reliable if the contour is longer and more closed, but it cannot find the discontinuities between two regions if the boundary is blurred, in low contrast, or in smooth transition. The surface cue is unable to simplify the representation of the regions with complex textures and large variance, which may lead to over-segmentation. The depth cue is helpful to clean the regions, especially for the textured ones, but unreliable if there is no edge details within these regions.

To maximize the positive effects but suppress the negative parts of these cues, we propose a novel image segmentation solution. First, three elementary segmentation modules are developed for these three cues respectively. For 1D contour cue, we adopt the structured edge detection [12] to find the map of long contours. For 2D surface cue, two CCPs [9] are applied to build the segmentation module. For 3D depth cue, we propose a depth estimation solution based on the low cost robust blur estimator [13]. The proposed depth estimation approach can handle low contrast regions and reduce the noises inside the regions, which is helpful to segment different textured regions even if they share similar appearance.

Then, to aggregate contour, surface, and depth cues, we propose a content-dependent spectral (CDS) graph for layered spectral segmentation, which is designed to build more reliable connection between surface nodes. If two neighboring surface nodes are disconnected by the long contour, or not in the identical depth layer, they will be disconnected in the graph; otherwise, the edge weight will be determined by their chroma affinities.

Extensive experiments on the Berkeley Segmentation Database [14] not only show the superior performance of the proposed algorithm over state-of-the-art methods, but also verify the necessities of the three cues in image segmentation. To the best of our knowledge, this is the first work to apply depth estimation for automatic image segmentation.



(a) (b) (c)

Fig. 3. Illustration of 2D surface Cue. (a) Original image. (b) CCP with $h_r = 5$. (c) CCP with $h_r = 6$. (b)-(c) are 2D surface cues. The red number on the bottom-right of each surface map is the number of regions for that surface cue.

2. THREE ELEMENTARY CUES AND SEGMENTATION MODULES

In this section, we introduce the three elementary cues and corresponding segmentation modules, which are the basic components of the proposed content-dependent spectral (CDS) segmentation.

2.1. 1D Contour Cue

Contour detection indicates the boundaries between two adjacent regions for image segmentation using the luma and chroma discontinuities. The longer the contour is, the more reliable the separation is to avoid noisy textured edges. In our design, we adopt the structured edge detection [12], followed by binarization to link the points into lists of coordinates pairs as long as possible [15]. Some results after edge linking are shown in Fig. 2 where our 1D contour cue are illustrated. In the contour map, black pixels indicate contours, and white pixels indicate non-contours.

2.2. 2D Surface Cue

Region-based segmentation, which utilizes surface properties to group similar pixels in appearance together, is quite successful in the last decade. The recent contour-guided color palette (CCP) [9] was demonstrated to outperform most of superpixel-based algorithms to generate robust surface layer. To increase the diversity of the surface layers, we create two CCP maps with $h_r = 5$ and $h_r = 6$ as our 2D surface cue, as shown in Fig. 3.

2.3. 3D Depth Cue

Nevertheless, the surface cue fails to group the textured regions where high contrast appears, and fails to find the accurate boundaries of those regions. These limitations will be alleviated by the depth cue.

Depth estimation from a still image globally separates the scenery into several layers, which helps to simplify the representation of the regions for image segmentation. The intuitive idea for depth estimation is that when taking a photo, the objects close to the focal plane are in focus, while the objects far from the focal plane are out of focus [16].

Hu and Haan's blur estimator [13] is a useful approach to generate a sparse defocus map around the object boundaries, as shown in Fig. 5(b), where darker pixels indicate regions in-focus and brighter pixels suggest regions out-of-focus. However, the map is not dense and smooth enough, which cannot be adopted for the depth estimation directly. In addition, for some dark areas, the contrast is too low to estimate the blur.

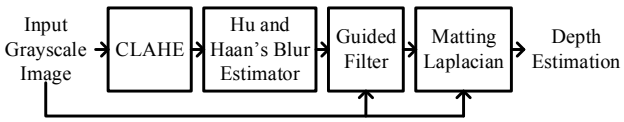
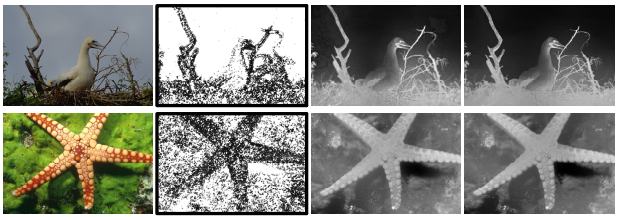


Fig. 4. Block diagram of the proposed depth estimation.



(a) (b) (c) (d)

Fig. 5. Illustration of 3D depth cue. (a) Original image. (b) Sparse defocus map by Hu and Haan [13]. (c) After guided filter [18]. (d) Our depth cue, where darker pixels indicate regions out-of-focus, and brighter pixels suggest regions in-focus. We can see the branches regions and the starfish regions are in the identical layer by our depth cue.

To overcome these problems, we propose the depth estimation algorithm as shown in Fig. 4 as our 3D depth cue. The depth estimation has four components, including CLAHE [17], Hu and Haan's blur estimator, guided filter [18], and matting Laplacian [19].

We apply contrast limited adaptive histogram equalization (CLAHE) as a preprocessing step to improve the local contrast, and bring out more details for the dark areas. At the same time, it restrains the noise in the relatively homogeneous regions. After the sparse blur estimation, some black pixels appear inside some regions, which come from textured regions or noise. To attenuate the depth variance inside the regions, we adopt guided filter to clean the textures and noises but also keep the edges (guided by the original color image), which helps to group the object regions in the same depth layer. Finally, the matting Laplacian is utilized to propagate the sparse defocus map to the entire image. The optimization function is formulated in Eqn. 1.

$$E(d) = d^T L d + \lambda \left(d - \hat{d}\right)^T D \left(d - \hat{d}\right) \qquad (1)$$

where $\hat{d}$ and d are the vector forms of the sparse defocus map and the target full defocus map. L is the matting Laplacian matrix and D is a diagonal matrix. λ is the scalar to balance between fidelity to the sparse depth map and smoothness of interpolation. There are two assumptions for this function. One is that the pixels nearby with similar appearance should have similar depth estimation. The other is that the depth estimation for the sparse pixels should be close to the sparse defocus map as much as possible.

Fig. 5 illustrates the 3D depth cue. We can see the depth cue can separate object and environment even if they share the similar appearances, like the bird and the sky regions. On the other hand, it also can well group the textured regions, like the branches and the starfish regions, which cannot be realized by the contour or the surface cue.

However, when there are no edges, the blur estimation on those regions will be meaningless. For this kind of cases, we have to rely on the surface cue to improve the unreliable part of the depth cue. That is the reason we have to utilize all of the three cues for segmentation.

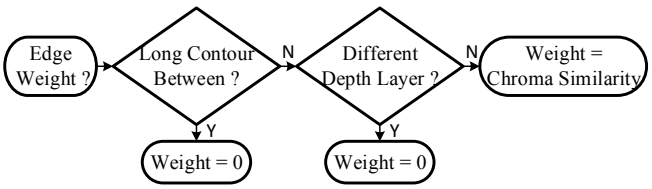


Fig. 6. Flowchart of how to measure the edge weight between two neighboring surface nodes.

3. CONTENT-DEPENDENT SPECTRAL SEGMENTATION

Layered affinity models have attracted more attentions in recent years [5, 6, 9]. They make full use of the surface/superpixel layers to connect pixels far from each other, so as to reduce the over-segmentation problem. However, it is still questionable to simply describe the affinities between neighboring surface nodes by their average similarities in Lab color space [20] all the time as [9] did. It will be too hard to differentiate two adjacent regions with apparently high similarity (especially when in the shaddow), even if their depth levels or texture patterns are so different.

To overcome the problem and achieve fine-scale image segmentation, we propose the content-dependent spectral (CDS) graph based on the advantages of three elementary and complementary cues, instead of the unreliable low-level features directly. In our design, the reliable parts of the contour and the depth cues inferred by luma have higher priority than the surface cue to determine the edge weight between two neighboring surface nodes, whose flowchart is shown in Fig. 6. At the first step, when a long contour is straddled by two neighboring surface nodes in the surface cue, the connection between them is ignored; later on, when the average depth estimation between them is large, the connection is also neglected; otherwise, the connection will be measured by their chroma similarities. The chroma component is leveraged here because it is more luminance invariant than the luma one, and could reduce over-segmentation problem in the shadow and the textured areas. For other connections between pixel node and surface node, we follow the same measure as the layered affinity models [6].

It will be the same as CCP-LAS[9] if we remove the contour and the depth cues and measure the edge weight between two surface nodes as Lab similarities.

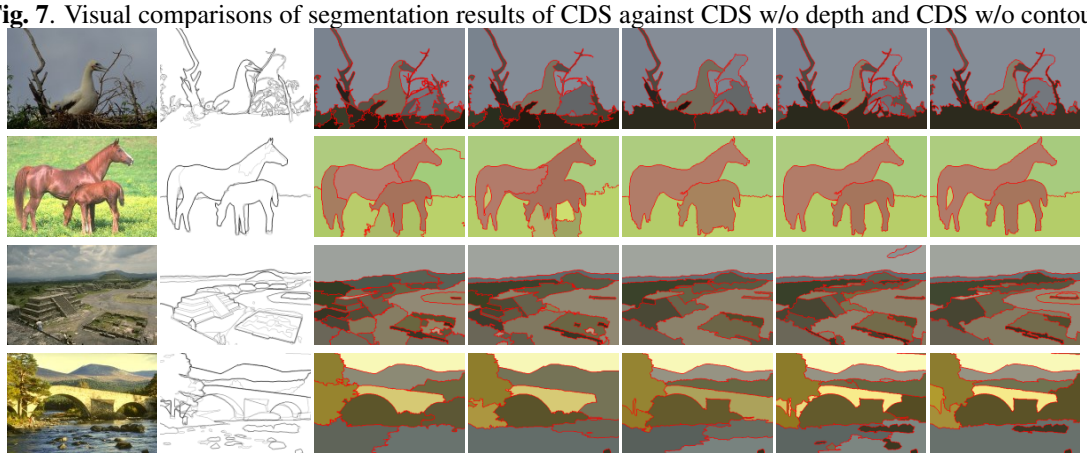
4. EXPERIMENTAL RESULTS

We evaluate the proposed segmentation method on the Berkeley Segmentation Dataset [14] and compare with state-of-the-art methods. Five common criteria [5, 6, 9] are used: 1) Segmentation Covering (Cov) [7]; 2) Probabilistic Rand Index (PRI) [21]; 3) Variation of Information (VoI) [22]; 4) Global Consistency Error (GCE) [14]; and 5) Boundary Displacement Error (BDE) [23]. When Cov and PRI are larger or the other three are smaller, the segmentation performs better.

The contour cue of CDS contains all the edges with length over 10 pixels. The surface cue of CDS is generated by CCP [9] with $h_r = 5$ and $h_r = 6$. For the CDS graph in SAS, the

(a) Original (b) Ground Truths (c) 1D Contour Cue (d) 2D Surface Cue (e) 3D Depth Cue (f) CDS w/o Depth (g) CDS w/o Contour (h) CDS

Fig. 7. Visual comparisons of segmentation results of CDS against CDS w/o depth and CDS w/o contour.



(a) Original (b) Ground Truths (c) MLSS (d) SAS (e) CCP-LAM (f) CCP-LAS (g) CDS

Fig. 8. Visual comparisons of segmentation results of CDS against four state-of-the-art methods: MLSS, SAS, CCP-LAM and CCP-LAS.

smoothness factor c, scale factor in surface affinity σ_{cue} and affinity between pixels and superpixels τ are empirically set to $c = 10^{-5}$, $\sigma_{\text{cue}} = 20$, and $\tau = 10^{-3}$. The edge weight between neighboring surface nodes will be 0, when the proportion of long contours on the common boundary is over 0.5, or the average depth estimation is over normalized scale 0.1.

In Table 1, we report the results of ten other popular segmentation methods: NCut [2], JSEG [24], MeanShift [3], FH [4], MNCut [25], NTP [26], MLSS [5], SAS [6], CCP-LAM and CCP-LAS [9]. In comparison, we list the results of the proposed content-dependent spectral method (CDS) with all the three elementary cues, without depth cue, and without contour cue, respectively[1].

From Table 1 we can see that, CDS outperforms all the other methods in terms of all the five metrics. Besides 2D surface cue that is actually the footstone of the segmentation, 1D contour cue and 3D depth cue have comparable contribution to the final segmentation performance. Fig. 7 shows one visual comparison of CDS segmentation against CDS w/o depth and CDS w/o contour. In the segmentation result, without depth, the water regions are over-segmented, and without contour, the bear regions cannot be separated correctly. Therefore, these three cues are indispensable and complementary for a better segmentation. Fig. 8 shows some example segmentations using all the three cues with four reference methods: MLSS, SAS, CCP-LAM, and CCP-LAS. We can observe that, owning to the depth cue, CDS is helpful to alleviate the over-segmentation problem, such as the branches, the grass, the historical sites, and the river areas in Fig. 8. It is noted that CDS only applies two surface layers, while the other four algorithms use at least three surface layers.

[1]Note that we do not have the algorithm CDS w/o surface cue, for the whole CDS algorithm is designed based on the 2D surface cue.

Table 1. Performance comparison on the BSDS300 Dataset. The best result are highlighted in bold.

Algorithm	**Cov↑**	**PRI↑**	**VoI↓**	**GCE↓**	**BDE↓**
NCut [2]	0.44	0.7242	2.9061	0.2232	17.15
JSEG [24]	N/A	0.7756	2.3217	0.1989	14.40
MeanShift [3]	0.54	0.7958	1.9725	0.1888	14.41
FH [4]	0.51	0.7139	3.3949	0.1746	16.67
MNCut [25]	0.44	0.7559	2.4701	0.1925	15.10
NTP [26]	N/A	0.7521	2.4954	0.2373	16.30
SDTV [27]	0.57	0.7758	1.8165	0.1768	16.24
RIS-HL [28]	0.59	0.8137	1.8232	0.1805	13.07
MLSS [5]	0.53	0.8146	1.8545	0.1809	12.21
SAS [6]	0.62	0.8319	1.6849	0.1779	11.29
CCP-LAM [9]	**0.68**	0.8404	1.5715	0.1635	10.20
CCP-LAS [9]	**0.68**	0.8442	1.5871	0.1582	10.46
CDS	**0.68**	**0.8539**	**1.5712**	**0.1572**	**10.18**
CDS (w/o depth)	0.65	0.8449	1.6293	0.1580	10.48
CDS (w/o contour)	0.64	0.8426	1.6185	0.1597	10.51

5. CONCLUSION

In this work, we proposed to leverage 1D contour cue, 2D surface cue, and 3D depth cue for automatic image segmentation. The contour cue describes discontinuities between two regions. The surface cue illustrates similarities inside one region. The depth cue makes use of blur estimation to indicate the image layout even if some regions are apparently similar, and merge the similar textures as well. We presented the three elementary cues and corresponding segmentation modules, based on which the content-dependent spectral (CDS) segmentation framework was proposed. The new scheme of edge weight between two neighboring surface nodes makes use of the advantages of the three cues and overcome the unreliable low-level measures. Experimental results on the Berkeley Segmentation Dataset have shown the necessity of all the three elementary cues for better segmentation and the superiority of the proposed method over state-of-the-art methods.

6. REFERENCES

[1] Jos B.T.M. Roerdink and Arnold Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundam. Inf.*, vol. 41, no. 1,2, pp. 187–228, Apr. 2000.

[2] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[3] Dorin Comaniciu and Peter Meer, "Mean Shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[4] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sept. 2004.

[5] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee, "Learning full pairwise affinities for spectral segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1690–1703, July 2013.

[6] Zhenguo Li, Xiao-Ming Wu, and Shih-Fu Chang, "Segmentation Using Superpixels: A bipartite graph partitioning approach," in *CVPR'12*, 2012, pp. 789–796.

[7] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[8] Xavier Muñoz, Jordi Freixenet, Xavier Cufí, and Joan Martí, "Strategies for image segmentation combining region and boundary information," *Pattern Recogn. Lett.*, vol. 24, no. 1-3, pp. 375–392, Jan. 2003.

[9] Xiang Fu, Chien-Yi Wang, Chen Chen, Changhu Wang, and C.-C. Jay Kuo, "Robust image segmentation using contour-guided color palettes," in *ICCV'15*, 2015, pp. 1618 – 1625.

[10] James J Gibson, *The perception of the visual world.*, Houghton Mifflin, 1950.

[11] Vicki Bruce, Mark A. Georgeson, and Patrick R. Green, *Visual Perception: Physiology, Psychology and Ecology*, Psychology Press, 4th edition, 2003.

[12] Piotr Dollr and C. Lawrence Zitnick, "Structured forests for fast edge detection," in *ICCV'13*, 2013, pp. 1841–1848.

[13] Hao Hu and Gerard de Haan, "Low cost robust blur estimator," in *ICIP'06*, 2006, pp. 617–620.

[14] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV'01*, 2001, pp. 416–423.

[15] Peter Kovesi, "MATLAB and Octave functions for computer vision and image processing," Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia, 2000, Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[16] Shaojie Zhuo and Terence Sim, "Defocus map estimation from a single image," *Pattern Recogn.*, vol. 44, no. 9, pp. 1852–1858, Sept. 2011.

[17] Karel Zuiderveld, *Graphics Gems IV*, Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[18] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, June 2013.

[19] Anat Levin, Dani Lischinski, and Yair Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008.

[20] Wikipedia, "Lab color space — wikipedia, the free encyclopedia," 2014, <http://en.wikipedia.org/wiki/Lab_color_space>.

[21] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, June 2007.

[22] Marina Meilă, "Comparing Clusterings: An axiomatic view," in *ICML'05*, 2005, pp. 577–584.

[23] Jordi Freixenet, Xavier Muñoz, David Raba, Joan Martí, and Xavier Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *ECCV'02*, 2002, pp. 408–422.

[24] Yining Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[25] Timothee Cour, Florence Benezit, and Jianbo Shi, "Spectral segmentation with multiscale graph decomposition," in *CVPR'05*, 2005, pp. 1124–1131.

[26] Jingdong Wang, Yangqing Jia, Xian-Sheng Hua, and Changshui Zhang, "Normalized tree partitioning for image segmentation," in *CVPR'08*, 2008, pp. 1–8.

[27] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *ICCV'09*, 2009, pp. 817–824.

[28] Jiajun Wu, Junyan Zhu, and Zhuowen Tu, "Reverse image segmentation: A high-level solution to a low-level task," in *BMVC'14*. 2014, BMVA Press.