Single Image Segmentation with Estimated Depth

Ryo Yonetani[1]
yonetani@vision.kuee.kyoto-u.ac.jp
Akisato Kimura[2]
akisato@ieee.org
Hitoshi Sakano[2]
sakano.hitoshi@lab.ntt.co.jp
Ken Fukuchi[3]
k2.fukuchi@jaist.ac.jp

[1] Kyoto University
Japan
[2] NTT Communication Science Labs.
Japan
[3] Japan Advanced Institute of Science and Technology
Japan

Object segmentation is a fundamental problem in computer vision. Although many segmentation methods have been proposed, most of them still rely on the appearances of images (i.e., colors or textures) [1, 2, 3, 4, 6, 8]. Consequently, they have a difficulty in distinguishing an object from the background with a similar appearance to the object.

To overcome this difficulty, **we employ a depth map of an input image as an additional cue to the object segmentation**. The main contribution of this work is to introduce a novel segmentation framework that utilizes the depth map combined with a color image to describe the features of objects and backgrounds, where the depth map is estimated from the color image. While a depth map has great potential for use in segmentation, finding a way of integrating two completely different physical quantities, namely the color and depth, has remained unclear. We introduce an integration of the color and depth likelihood on objectness and backgroundness, which simply and effectively extends a traditional segmentation framework based on the Markov random fields (MRF) [2]. By refining the likelihood with the depth information, our proposed method can suppress the incorrect detection of misleading backgrounds.

A single image is expressed by K, where K includes color information $\mathcal{C} = \{C_x \in \mathbb{R}^3\}_{x\in\Omega}$, and in our case, depth information $\mathcal{Z} = \{Z_x \in \mathbb{R}\}_{x\in\Omega}$ (x is a position in the image domain $\Omega \subset \mathbb{N}^2$). Object segmentation is the problem of assigning the label $\mathcal{A} = \{A_x\}_{x\in\Omega}$, which gives a label $A_x = \{0,1\}$ to each pixel, where the labels 1 and 0 at x respectively correspond to the object and background. The statistical relationship between K and $\mathcal{A}$ can be described by an MRF, and the appropriate configuration of the labels can be derived by minimizing the following energy function E:

$$E = \sum_{x\in\Omega} \left\{ \phi_{\mathrm{D}}(K \mid A_x) + \xi_{\mathrm{D}}(A_x) + \sum_{y\in N_x} \left(\phi_{\mathrm{S}}(K \mid A_x, A_y) + \xi_{\mathrm{S}}(A_x, A_y) \right) \right\},$$

where N_x is a 4-neighborhood system of the position x. The data prior term $\xi_{\mathrm{D}}(A_x)$ evaluates how likely to an object the position is for all the pixels in an image, and the smoothness prior term $\xi_{\mathrm{S}}(A_x, A_y)$ is given by the Kronecker delta to ensure the spatial continuity of labels. The data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$ is modeled by the negative log likelihood of the data value conditioned by the labels: traditionally $\phi_{\mathrm{D}}(K \mid A_x) \propto -\log p(C_x \mid A_x)$. On the other hand, the smoothness likelihood term gives the difference of intensities where labels are spatially discontinuous.

An integration of the color and depth cue is a central part of this study. For introducing depth information into the segmentation framework, we here present a key observation of a structural difference of depth maps from color images. As shown in Figure 1, depth-map structures are quite different from those of color images. In particular, the spatial discontinuities between the pixel values of objects and backgrounds in depth maps do not always agree with those in color images (e.g., an object and the floor on which the object is placed). As a result, **a consideration of depth continuities prevents us from distinguishing objects from backgrounds, which implies that depth information is inappropriate to the smoothness term** ϕ_{S}. On the other hand, Figure 1 also demonstrates that the averages in depth distributions appear at different values between the object and background, while the corresponding intensity distributions look like each other. From these observation, **we decide to introduce depth information into the data term** ϕ_{D}. Specifically, we take particular note of "foregroundness" (nearness) Z_x besides a color value C_x, and fuse them as a weighted sum of likelihood values to derive the data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$. Consequently, $\phi_{\mathrm{D}}(K \mid A_x)$ is modified as follows:

$$\phi_{\mathrm{D}}(K \mid A_x = i) \propto -\log p(C_x \mid A_x = i) - \alpha_i \log p(Z_x \mid A_x = i) \ (i = 0, 1),$$

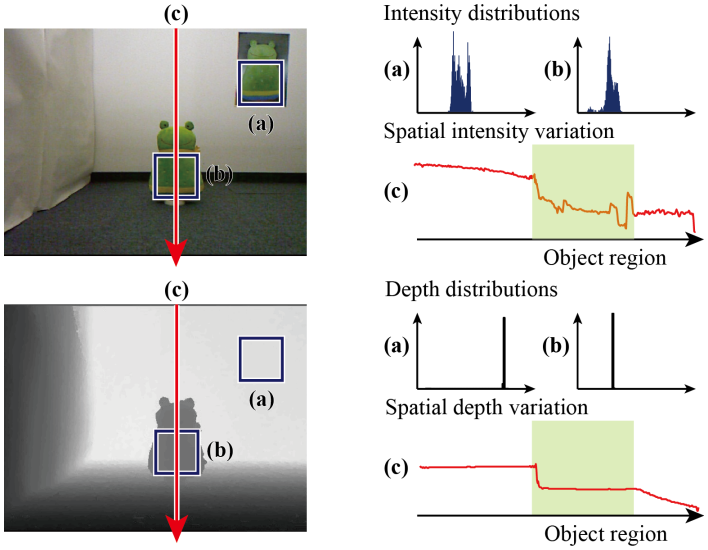

Figure 1: Distributions (blue) and spatial variations (red) of the data in color images and depth maps. The green rectangles describe the region on which the object is displayed.

where α_i is a scale factor of the depth likelihood individually set for both $A_x = 1$ and $A_x = 0$. Note that depth and color distributions may take a different variation because of the difference in the possible range of themselves. We determine α_i by cross validation in the experiments.

Our implementation of the proposed framework comprises not only the integration of colors and depths but automatic computation of the prior term $\xi_{\mathrm{D}}(A_x)$ using a visual attention models [1, 5], and single-image depth estimation via supervised learning [7]. That is, **the proposed method performs automatic object segmentation from a single image, which requires no actual depth maps corresponding to input images.**

[1] K. Akamine, K. Fukuchi, A. Kimura, and S. Takagi. Fully automatic extraction of salient objects from videos in near real time. *The Computer Journal*, 55(1):3–14, 2012.

[2] Y. Boykov and G. Funka-Lea. Graph Cuts and Efficient N-D Image Segmentation. *IJCV*, 70(2):109–131, 2006.

[3] K. Fukuda, T. Takiguchi, and Y. Ariki. Graph Cuts by Using Local Texture Features of Wavelet Coefficient for Image Segmentation. In *ICME*, 2008.

[4] Z. Kato and T. Pong. A Markov Random Field Image Segmentation Model for Color Textured Images. *Image and Vision Computing*, 24 (10):1103–1114, 2006.

[5] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network. In *ICME*, 2008.

[6] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. on Graphics*, 23(3):309–314, 2004.

[7] A. Saxena, S. Chung, and A. Ng. Learning Depth from Single Monocular Images. In *NIPS*, 2006.

[8] S. Vicente, V. Kolmogorov, and C. Rother. Joint Optimization of Segmentation and Appearance Models. In *ICCV*, 2009.