# 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

**Abstract**

A novel framework for automatic object segmentation is proposed that exploits depth information estimated from a single image as an additional cue. For example, suppose that we have an image containing an object and a background with a similar color or texture to the object. The proposed framework enables us to automatically extract the object from the image while eliminating the misleading background. Although our segmentation framework takes a form of a traditional formulation based on Markov random fields, the proposed method provides a novel scheme to integrate depth and color information, which derives objectness/backgroundness likelihood. We also employ depth estimation via supervised learning so that the proposed method can work even if it has only a single input image with no actual depth information. Experimental results with a dataset originally collected for the evaluation demonstrate the effectiveness of the proposed method against the baseline method and several existing methods for salient region detection.

## 1 Introduction

This paper presents a novel framework for automatic object segmentation utilizing a depth map estimated from an input image, which can distinguish an object from a background with a similar color or texture (see Figure 1). The proposed method can work with only a single image, in an automatic manner, and it can even run without any actual depth map corresponding to the input image.

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 [5, 11, 18, 20, 24, 30]. Therefore, those methods have a difficulty in distinguishing an object from the background with a similar appearance to the object. For instance, suppose that an input image contains a target object and that there is a *mimic* that looks similar to the object as shown in Figure 1(a). In this case, it is inevitable that the appearance-based methods incorrectly detect the mimic such as in Figure 1(b).


http://dx.doi.org/10.5244/C.26.28

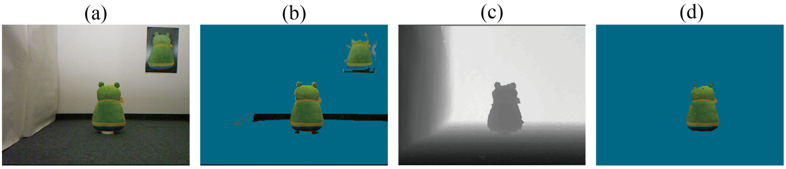


Figure 1: Concept figure of our approach. (a) Image containing an object (middle) and a mimic (top right) with a similar appearance. (b) Segmentation result using only colors. (c) Depth map corresponding to the image. (d) Segmentation result using colors and depths.

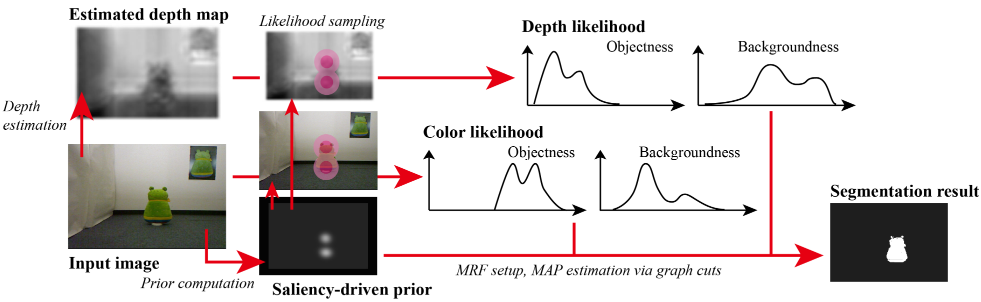


Figure 2: Overview of proposed method.

On the other hand, humans can correctly distinguish the object from many types of background since we subconsciously take account of the depth information of images (Figure 1(c)). The depth difference between the object and background plays a crucial role to eliminate any incorrect detection of the background such as in Figure 1(d).

In this paper, we employ a depth map of an input image as an additional cue to the segmentation. The main contribution of this work is to introduce a novel segmentation framework that utilizes an estimated depth map to describe the features of the object and background. 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. With an observation of depth-map structures, we introduce an integration of color and depth likelihood on objectness and backgroundness, which simply and effectively extends a traditional segmentation framework based on the Markov random fields (MRF). By refining the likelihood with depth information, our proposed method can suppress the incorrect detection of mimics and other backgrounds.

Figure 2 shows an overview of the proposed method. The proposed method basically takes a form of the traditional framework proposed in [5] (see Section 2.2); it builds likelihood distributions of objectness and backgroundness by sampling pixel values of an input image based on a given prior (an initial seed for the segmentation). Specifically, the proposed method selects samples from both input image and depth map to individually estimate the likelihood distribution, and integrates likelihood values from the two distributions as a form of a weighted sum. Several discussions on the integration and related work are presented in Section 2.3. Automatic segmentation can be achieved with the help of a saliency-driven prior computed from a single image (see Section 2.4). On utilizing the depth, our method includes depth estimation so that it works even if the input image has no actual depth information. Specifically, we estimate depth maps via supervised learning (see Section 3).

# 2 Proposed method

## 2.1 Problem definition

Let us first define the problem formulation and mathematical notations. A single image is expressed by $K = \{\mathcal{I}, \mathcal{C}\}$, where $\mathcal{I} = \{I_x \in \mathbb{R}\}_{x\in\Omega}$ denotes an intensity image consisting of a pixel value $I_x$ at position $x \in \Omega$; ($\Omega \subset \mathbb{N}^2$ denotes the image domain) and $\mathcal{C} = \{C_x \in \mathbb{R}^3\}_{x\in\Omega}$ denotes a color image with a value $C_x$. The image has an object region $\mathcal{O} \subset \Omega$ to be segmented. In its background region $\mathcal{B} = \Omega - \mathcal{O}$, there may be regions with a similar color or texture to the object, which we refer to as mimics. 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 $x \in \mathcal{O}$ and background $x \in \mathcal{B}$.

Specifically, the problem in this paper is automatic object segmentation from a single image. Automatic segmentation does not allow the manual avoidance of false segmentation whereas manual segmentation approaches allow users to give an explicit background label to misleading regions. Furthermore, single image segmentation cannot employ motion or occlusion information, which has potential to specify backgrounds.

## 2.2 MRF-based object segmentation

The statistical relationship between an input image $K$ and assigned labels $\mathcal{A}$ can be described by an MRF, and the appropriate configuration of the labels, $\hat{\mathcal{A}}$, can be derived via a maximum-a-posteriori (MAP) estimation, i.e., $\hat{\mathcal{A}} = \arg\max_{\mathcal{A}} p(\mathcal{A} \mid K)$. This estimation is equivalent to an energy minimization problem where the energy function can be represented as the negative log likelihood of the joint posterior density distribution of the MRF, $E(\mathcal{A} \mid K) = -\log p(\mathcal{A} \mid K)$. The energy function is defined as follows:

$$E(\mathcal{A} \mid K) = \sum_{x\in\Omega} \Big\{ \phi_{\mathrm{D}}(K \mid A_x) + \xi_{\mathrm{D}}(A_x) + \sum_{y\in N_x} \big( \phi_{\mathrm{S}}(K \mid A_x, A_y) + \xi_{\mathrm{S}}(A_x, A_y) \big) \Big\}, \tag{1}$$

where $N_x$ is a 4-neighborhood system of the position $x$. $\phi_{\mathrm{D}}(K \mid \cdot)$ and $\phi_{\mathrm{S}}(K \mid \cdot)$ are a likelihood term, and $\xi_{\mathrm{D}}(\cdot)$ and $\xi_{\mathrm{S}}(\cdot)$ are a prior term, where the terms specified by the subscription D and S describe data and smoothness terms, respectively. 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. It is formulated as $\xi_{\mathrm{D}}(A_x) = -\log p(A_x)$, where the density $p(A_x)$ is given as $p(A_x = 1) = 1$ or $p(A_x = 1) = 0$ if the label 1 (object) or 0 (background) is given at the point $x$ respectively, $p(A_x = 1) = 0.5$ if no label is provided at the point, and $p(A_x = 0)$ is defined as $p(A_x = 0) = 1 - p(A_x = 1)$. In addition, the smoothness prior term $\xi_{\mathrm{S}}(A_x, A_y)$ is given by the Kronecker delta.

Whereas the prior stands for the objectness or backgroundness in the image domain, the likelihood evaluates them in the feature domain. The data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$ is the negative log likelihood that imposes pixel-wise penalties for assigning the label $A_x$ to pixel $x$. This likelihood basically employs the color information as $\phi_{\mathrm{D}}(K \mid A_x) \propto -\log p(C_x \mid A_x)$. The likelihood distribution $p(C_x \mid A_x)$ can be modeled as a Gaussian mixture model (GMM). The RGB values $C_x$ are first sampled based on the prior $p(A_x = 1)$ and $p(A_x = 0)$ individually, and then GMMs $p(C_x \mid A_x = 1)$ and $p(C_x \mid A_x = 0)$ are estimated from the samples based on the expectation-maximization (EM) algorithm. The smoothness likelihood term gives the difference of intensities for reducing the cost of two adjacent labels as follows:

$$\phi_{\mathrm{S}}(K \mid A_x, A_y) \propto -\exp\left\{-\frac{(I_x - I_y)^2}{2\sigma^2}\right\} \cdot \frac{1}{\|x-y\|} \text{ if } A_x \neq A_y. \tag{2}$$

Given the energy function shown above, it can be minimized by finding the minimum cut on a graph equivalent to the MRF (see [6, 25] for details of the algorithm).

## 2.3 Introducing estimated depth

We now propose a novel formulation of segmentation, which introduces depth information. The likelihood term evaluates features extracted from an input image, and thus it is expected to involve features from a depth map.

**Introduction of depth feature and its extraction** As shown in Figure 3, depth-map structures are quite different from those of color images. In particular, the spatial discontinuities of pixel values between 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). Therefore, a consideration of depth continuities prevents us from distinguishing objects from backgrounds, which implies that depth information is inappropriate to the smoothness term $\phi_{\mathrm{S}}$.

Instead, we introduce depth information into the data term $\phi_{\mathrm{D}}$. The problem is which kind of depth features is promising. Here we note that we can utilize only the estimated depth map that is not always accurate. Therefore, a convexity and a surface normal, which are often utilized as depth features such as [8, 13, 22], are hard to introduce. Instead we take particular note of "foregroundness" (nearness) of regions. Figure 3 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. The above consideration indicates that the absolute depth value is expected to be a simple but effective feature.

**Integration of color and depth information** The color and depth, or the features extracted from them, are essentially difficult to fuse directly since they represent completely different physical quantities. We therefore introduce a fusion of likelihood; specifically, a weighted sum of color and depth likelihood values is introduced for the data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$. With the depth map denoted as $\mathcal{Z} = \{Z_x\}_{x\in\Omega}$, $\phi_{\mathrm{D}}(K \mid A_x)$ is modified as follows:

$$\phi_{\mathrm{D}}(K \mid A_x = i) \propto -\log p(C_x \mid A_x = i) - \alpha_i \log p(Z_x \mid A_x = i) \ (i = 0, 1), \tag{3}$$

where $\alpha_i$ denotes a scale factor of the depth likelihood, which is individually set for both $A_x = 1$ and $A_x = 0$. Note that distributions of depths $\mathcal{Z}$ and colors $\mathcal{C}$ may take a different variation because of the difference in the possible range of $\mathcal{Z}$ and $\mathcal{C}$. We determine $\alpha_i$ by cross validation in the experiments.

Depth likelihood distributions $p(Z_x|A_x = 1)$ and $p(Z_x|A_x = 0)$ are modeled by the GMM as well as color likelihood distributions $p(C_x|A_x = 1)$ and $p(C_x|A_x = 0)$. The parameters of the GMM are estimated based on the EM algorithm, via the sampling of depth values based on the prior $p(A_x = 1)$ and $p(A_x = 0)$.

**Related work on multi-cue integration** When the objective of integration is object detection or classification, several studies evaluate how much each cue contributes to the tasks for the integration [9, 13]. For image segmentation, cues are often integrated in specific terms in

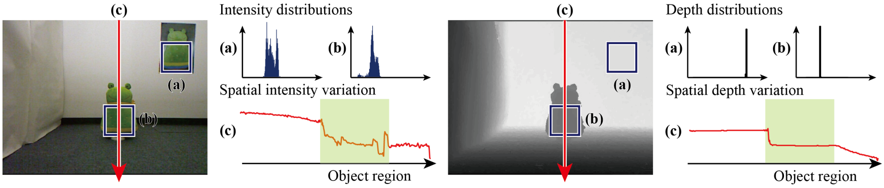


Figure 3: Distributions (blue) and spatial variations (red) of the data in color images and depth maps. The green rectangles describe the region on which objects are displayed.

the MRF [11] or by combining MRFs [17, 18]. Here, our proposed method takes a similar approach to [11, 17, 18] since it integrates color and depth information based on the MRF. In [17, 18], they basically regard multiple features (color and texture in [18], and color and motion in [17]) as equals, and introduce the other feature (texture or motion) to the smoothness term. Our approach employs depth information together with color, and thus it is difficult to apply the above approaches because of the characteristics of depth information as shown in Section 2.3. On the other hand, [11] introduces color and texture information for the data term, and color information for the smoothness term. The proposed method can be regarded as the same group as the method above.

Several studies have reported the joint utilization of appearance and depth information. Depth estimation and segmentation are solved simultaneously using stereo images [4, 32]. Given depth maps, color and depth information is integrated for indoor scene segmentation [29] or classification [3]. From the point of utilizing the given depth maps, the proposed method is especially close to [3, 29]. However, the proposed method is different from those methods since it requires no actual depth information thanks to the depth estimation.

## 2.4 Automatic computation of prior

To avoid the manual labeling of the prior $\xi_{\mathrm{D}}(A_x)$, several automatic approaches have been proposed [2, 10, 12, 19, 31]. As described in Section 2.2, $\xi_{\mathrm{D}}(A_x)$ plays a role to give objectness and backgroundness to positions in an image domain as well as to yield feature samples for the estimation of likelihood distributions $\phi_{\mathrm{D}}(K \mid A_x)$. Considering the former role, the regions with high objectness prior are expected to specify object positions roughly. However, the latter role requests the objectness prior to be high only within object regions.

To meet these requirements, we employ the eye focusing density map (EFDM) [23] that indicates where humans tend to fixate in an image. While the EFDM is based on a typical saliency map [15], it introduces the specifications of several high salient positions via stochastic computation. The obtained EFDM is applied to the prior density $p(A_x = 1)$. Exceptionally, the prior density at the edge of the image is assumed to be $p(A_x = 1) = 0$ since some of the edges are expected to be background. The effectiveness of EFDM for segmentation has been confirmed in [2, 10].

In terms of utilizing saliency maps for segmentation, salient-region detection (SRD) techniques including [1, 7, 16] have a similar concept. They are essentially quite different approaches; their computation of saliency maps also includes detection of object-region boundaries, which results in the segmentation based on thresholding techniques. The comparison between the proposed method and SRD-based methods is provided in Section 4.

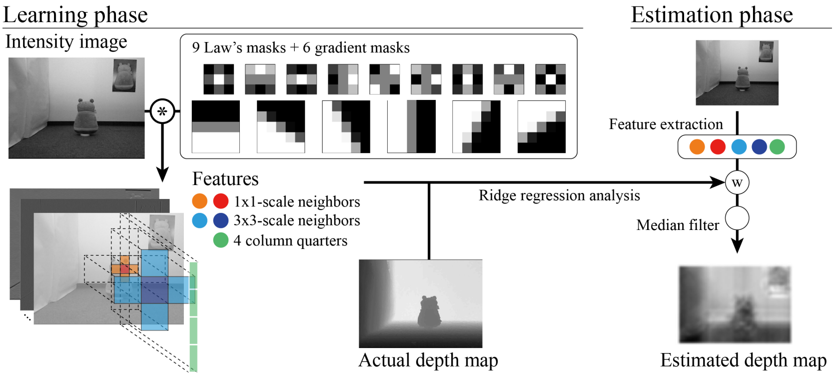


Figure 4: Overview of depth estimation.

# 3 Depth estimation

Depth estimation has several typical approaches including binocular stereo vision, structure from motion (SFM), and shape from X. This paper focuses on single image segmentation, and thus requires a depth estimation from a single image such as [14, 21, 26, 27, 28]. We here introduce a supervised approach which simplifies the method proposed in [27] to reduce computational efforts in the learning phase while keeping the concepts of the method. The original work [27] introduces a formulation based on the MRF which involves spatial image-feature relationships. The MAP estimation of the MRF is eventually equivalent to the regression which maintains the robustness of model and the smoothness of output. Instead, we introduce a pixel-wise regression of depth values with $\ell_2$-norm regularization of model parameters. Moreover, we apply a median filter to the depth map to obtain smoothness.

We employ several features suggested in [27, 28] (see Figure 4). Specifically, the primitive features are first computed for each pixel by applying Law's masks and texture gradient masks to the intensity image $\mathcal{I}$. We then adopt 3 types of patches considering spatial relationships: $1\times1$ neighbors, $3\times3$ neighbors, and column quarter partitions. Features for estimating depth maps are finally derived by aggregating the primitive features in those patches.

# 4 Experiments

## 4.1 Experimental setup

To verify the proposed method, we prepared a dataset consisting of color images and depth maps as a first step. The dataset was collected carefully under the controlled environment including 4 instances of colored objects, under the 4 kinds of backgrounds, which included several mimics with a similar color or texture to the objects. Here, for simplicity, we assume that the image includes only one object [1]. The mimics included single-colored rectangular papers or posters in which the objects were printed.

Pairs of aligned color images and depth maps were captured with a structured light sensor. 16 different scenes (4 objects $\times$ 4 backgrounds) constitute the obtained dataset, and

[1]Since the proposed method takes the form of traditional MRF-based frameworks, it can be easily applied to multiple object segmentation via $\alpha$-expansion or other approaches.

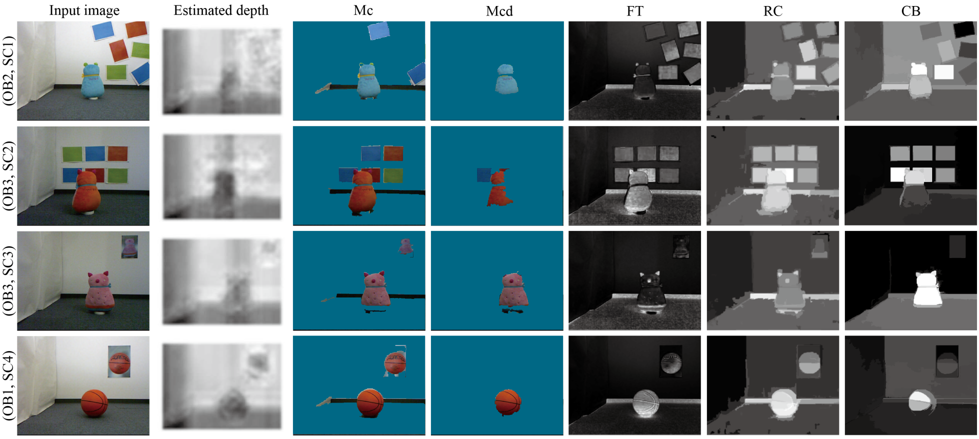


Figure 5: Example of segmentation results using $M_c$: color-based method and $M_{cd}$: proposed method, and saliency computation using FT: [1], RC: [7], and CB: [16].

each of the scenes has 70 pairs of color images and depth maps, which include variations of object poses, positions, and lighting environments. Consequently, 1120 pairs were captured. Depth maps obtained with a structured light sensor are known to have deficiencies, which are called infra-red (IR) occlusions. This phenomenon often occurs when the IR laser light is occluded by some obstacles. Such IR occlusions in depth maps were interpolated with a pre-processing technique by using the nearest-neighbor depth values.

848 randomly chosen pairs were employed as a training dataset for depth estimation, and the remaining 272 pairs were used as a test dataset. We manually annotated the ground-truth region of objects for the test images, and applied a pixel-wise evaluation to all the segmentation results. The recall, precision, and F-measure score were adopted as evaluation measures. The scale constants $\alpha_i$ presented in Section 2.3 were estimated via 4-fold cross validation based on the F-measure.

Color-based segmentation was employed as the baseline approach (referred to as $M_c$). Note that both $M_c$ and the proposed method ($M_{cd}$) utilize the same prior to keep the same experimental conditions. Additionally, several SRD-based methods, FT [1], RC [7], CB [16], were examined here. These approaches originally have unique thresholding techniques to obtain segmentation results from their saliency maps. However, in the experiments we determined a threshold based on the cross validation to achieve fairness between those methods.

## 4.2 Results and discussions

**Results** Table 1 shows the scores averaged over each measure obtained with each method. In addition, Figure 5 shows some examples of segmentation results. With regard to Figure 5, (SC1) and (SC2) have a background with single-colored rectangular papers. In (SC1), the object does not occlude the papers whereas (SC2) includes objects that occlude some of the papers. On the other hand, (SC3) and (SC4) have a background consisting of a poster on which the object is displayed. In particular, (SC4) includes a larger poster than (SC3). These results demonstrate that $M_{cd}$ can work well with regard to the F-measure score thanks to the improvement of precision, while it shows a comparable recall to the other methods.

Table 1: Scores averaged over each measure (controlled images)

| | Recall | Precision | F-measure |
|---|---|---|---|
| $M_c$ | **0.92** | 0.47 | 0.61 |
| $\mathbf{M_{cd}}$ | 0.88 | **0.76** | **0.80** |
| FT | 0.58 | 0.42 | 0.46 |
| RC | 0.58 | 0.69 | 0.62 |
| CB | 0.67 | 0.60 | 0.60 |

Table 2: Scores averaged over each measure (complex images)

| | Recall | Precision | F-measure |
|---|---|---|---|
| $M_c$ | **0.91** | 0.44 | 0.57 |
| $\mathbf{M_{cd}}$ | 0.85 | 0.70 | **0.74** |
| FT | 0.48 | 0.59 | 0.51 |
| RC | 0.64 | 0.71 | 0.65 |
| CB | 0.80 | **0.79** | **0.74** |

**Improvement of precision** $M_{cd}$ can suppress the incorrect detection of backgrounds despite of the use of the same prior as the baseline $M_c$. Generally, background regions are incorrectly detected when the objectness prior captures not only object regions but also parts of the backgrounds. However, since depth values in object and background regions were different from each other and their variations were quite small in both regions as shown in Figure 3, the incorrect detection of background was suppressed even if the prior was unstable. Only the case that $M_{cd}$ fails segmentation is when the background occupies most of the regions with high objectness prior such as Figure 6.

**Limitation on recall** $M_{cd}$ sometimes brings down over-segmentation at object boundaries as shown in Figure 5, which results in the decrease of recall scores. As described in Figure 7, the depth estimation implemented in $M_{cd}$ tends to regard objects as convex regions. Hence, the depth values of the object boundaries tend to be similar to those of backgrounds, and it eventually brings down the over-segmentation. Together with the improvement of depth estimation, it will be effective to adaptively emphasize colors than depths for integration at the regions with inaccurate depths.

**Comparison with salient-region detection** As mentioned in Section 2.4, the SRD-based methods, FT, RC and CB, take different approaches from MRF-based approaches including the proposed method. One of characteristics of SRD-based methods is that they do not specify what the object is, and thus sometimes detect mimics incorrectly or at the worst fail to capture objects (see the 1st and 3rd row of the column FT) as shown in Figure 5. On the other hand, SRD-based methods can obtain accurate region boundaries at the stage of their saliency computation, while the proposed method often tends to bring down the over-segmentation. Obviously the SRD-based methods can work well for images that contain large-size objects and require no specification of the objects as shown in [1, 7, 16]. Eventually, it is effective to choose appropriate methods depending on the types of scenes.

**Performance for images with complex backgrounds** Finally, Figure 8 and Table 2 show some results of segmentation for 86 pairs of images with complex backgrounds consisting of several office scenes. Since the dataset contains small number of images against a large variation of backgrounds, the parameters for depth estimation, $\alpha_i$ and the threshold for binalization in the SRD-based methods were manually adapted with regard to each of the scenes. On these complex images, the baseline method $M_c$ often detect backgrounds incorrectly

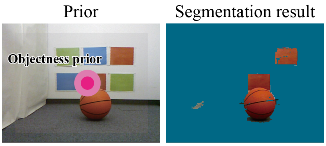


Figure 6: Failure case due to the objectness prior.

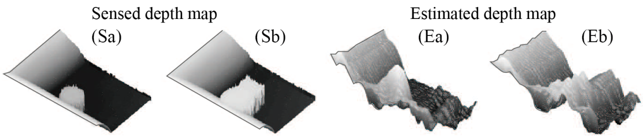


Figure 7: Examples of depth estimation. (Ea) and (Eb) correspond to sensed depth maps (Sa) and (Sb), respectively.

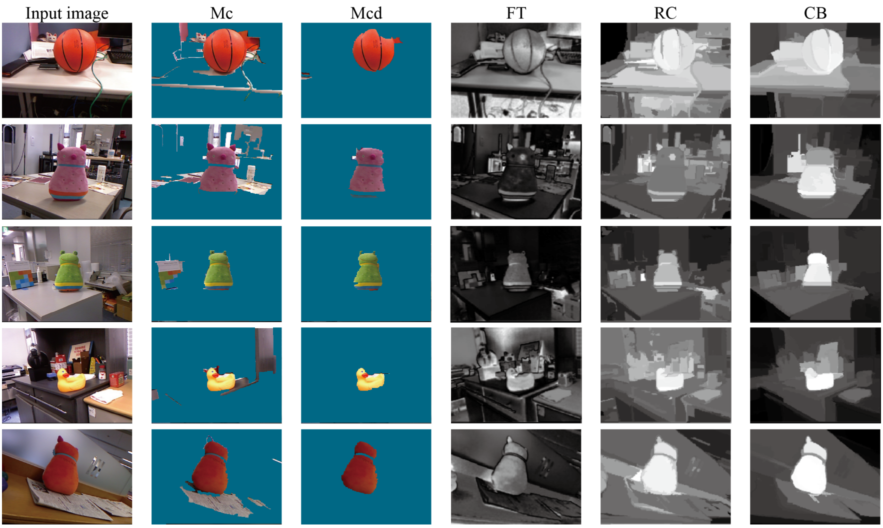


Figure 8: Examples of segmentation results and saliency computation for images with complex backgrounds.

when the prior behaves unstable. On the other hand, the proposed method $M_{cd}$ generally eliminates the incorrect background detection thanks to the introduction of depth information though it often brings down over-segmentation. Regarding to the SRD-based methods, some results show that they completely fail to specify objects, and others show the incorrect detection of backgrounds around the objects. Still, the method CB can accurately capture the objects, and it obtains comparable F-measure scores with the proposed method.

# 5 Conclusions

We proposed a novel framework for automatic object segmentation from a single image. Since the method employs depth information combined with color information, the false detection of backgrounds containing a similar appearance to the objects is greatly decreased.

The future work will include not only an improvement of depth estimation and adaptive determination of the importance weights of colors and depths in integration, but the build of a dataset consisting of color images and depth maps with complex backgrounds to introduce a large-scale depth estimation, which will realize an open test of the proposed method.

**Acknowledgements** The authors thank Dr. Naonori Ueda, Dr. Eisaku Maeda, Dr. Futoshi Naya, Dr. Hiromi Nakaiwa, and Dr. Hiroshi Sawada of NTT Communication Science Laboratories for their help.

# References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-Tuned Salient Region Detection. In *CVPR*, 2009.

[2] 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.

[3] A. Bar-hillel, D. Hanukaev, and D. Levi. Fusing Visual and Range Imaging for Object Class Recognition. In *ICCV*, 2011.

[4] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. Object Stereo — Joint Stereo Matching and Object Segmentation. In *CVPR*, 2011.

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

[6] Y. Boykov and V. Kolmogorov. Basic Graph Cut Algorithms. In A. Blake, P. Kohli, and C. Rother, editors, *Markov Random Fields for Vision and Image Processing*, chapter 2. The MIT Press, 2011.

[7] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[8] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model Globally, Match Locally: Efficient and Robust 3D Object Recognition. In *CVPR*, 2010.

[9] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila. Multi-Cue Pedestrian Classification with Partial Occlusion Handling. In *CVPR*, 2010.

[10] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato. Saliency-Based Video Segmentation with Graph Cuts and Sequentially Updated Priors. In *ICME*, 2009.

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

[12] K. Fukuda, T. Takiguchi, and Y. Ariki. Automatic Segmentation of Object Region Using Graph Cuts Based on Saliency Maps and AdaBoost. In *ISCE*, 2009.

[13] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In *ICCV*, 2011.

[14] D. Hoiem, A. Efros, and M. Hebert. Automatic Photo Pop-up. *ACM Trans. on Graphics*, 24(3), 2005.

[15] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259, 1998.

[16] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng. Automatic salient object segmentation based on context and shape prior. In *BMVC*, pages 110.1–110.12, 2011.

[17] P. Jodoin, M. Mignotte, and C. Rosenberger. Segmentation Framework Based on Label Field Fusion. *IEEE Trans. on IP*, 16(10):2535–2550, 2007.

[18] 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.

[19] Y. Lee, J. Kim, and K. Grauman. Key-Segments for Video Object Segmentation. In *ICCV*, 2011.

[20] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image Segmentation with a Bounding Box Prior. In *ICCV*, 2009.

[21] B. Liu, S. Gould, and D. Koller. Single Image Depth Estimation From Predicted Semantic Labels. In *CVPR*, 2010.

[22] A. Mian, M. Bennamoun, and R. Owens. Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes. *IEEE Trans. on PAMI*, 28(10): 1584–1601, 2006.

[23] 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.

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

[25] C. Rother, V. Kolmogorov, Y. Boykov, and A. Blake. Interactive Foreground Extraction: Using Graph Cut. In A. Blake, P. Kohli, and C. Rother, editors, *Markov Random Fields for Vision and Image Processing*, chapter 7. The MIT Press, 2011.

[26] D. Rother and G. Sapiro. Seeing 3d objects in a single 2d image. In *ICCV*, pages 1819–1826, 2009.

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

[28] A. Saxena, M. Sun, and A. Ng. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. on PAMI*, 31(5):824–840, 2009.

[29] N. Silberman and R. Fergus. Indoor Scene Segmentation using a Structured Light Sensor. In *ICCV Workshop on 3DRR*, 2011.

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

[31] F. Yu, C. Jian, L. Zhenglong, and L. Hanqing. Saliency Cuts: An Automatic Approach to Object Segmentation. In *ICPR*, 2008.

[32] G. Zhang, J. Jia, and H. Bao. Simultaneous Multi-Body Stereo and Segmentation. In *ICCV*, 2011.