

# DEPTH MAP GENERATION USING LOCAL DEPTH HYPOTHESIS FOR 2D-TO-3D CONVERSION

Na-Eun Yang<sup>1</sup>, Ji Won Lee<sup>2</sup>, and Rae-Hong Park<sup>3</sup>

<sup>1,2,3</sup>Department of Electronic Engineering, Sogang University, Seoul, Korea

<sup>1</sup>naeun@sogang.ac.kr, <sup>2</sup>nkmission@sogang.ac.kr, <sup>3</sup>rhpark@sogang.ac.kr

## ABSTRACT

*A single two dimensional (2D) image does not contain depth information. An infinite number of points in the three dimensional (3D) space are projected to the same point in the image plane. But a single 2D image has some monocular depth cues, by which we can make a hypothesis of depth variation in the image to generate a depth map. This paper proposes an interactive method of depth map generation from a single image for 2D-to-3D conversion. Using a hypothesis of depth variation can reduce the human effort to generate a depth map. The only thing required from a user is to mark some salient regions to be distinguished with respect to depth variation. The proposed algorithm makes hypothesis of each salient region and generates a depth map of an input image. Experimental results show that the proposed method gives natural depth map in terms of human perception.*

## KEYWORDS

*Depth Estimation, Depth Map Generation, 2D-to-3D Conversion, Pseudo 3D*

## 1. INTRODUCTION

Thanks to the growth of three dimensional (3D) market including stereo vision, 3D graphics, 3D visualization and so on, various 3D technologies have been realized. Lots of 3D devices, such as television, mobile phone, and projector, have been developed. Accordingly, the demand for 3D contents has been increased. Compared to making two dimensional (2D) contents, making 3D contents requires special equipment like Rig, experts to produce 3D contents, and extra time to editing 3D contents. Therefore, production of 3D contents is time-consuming and expensive process. Instead of making 3D contents directly, 2D-to-3D conversion is used as an alternative to meet user's demand for 3D contents.

For successful 2D-to-3D conversion, depth information is needed. Based on a depth map, we can generate stereoscopic image or another view image from a single image [1]-[3]. The depth information is not included in 2D images. But human perceives a sense of depth from various heuristic, monocular depth cues: focus/defocus [4], [5], relative height/size and texture gradient [6], structural feature from occlusion [7], geometry and texture [8], [9], and so on. These monocular depth cues make human perceive depth from a single-view image. Based on these cues, many studies [4]-[9] have been done. Even though an image has various monocular cues, it is difficult for these studies to estimate depth map from monocular cues alone. Lai et al. [4] estimated depth from defocus. Zhuo and Sim [5] generated a defocus map by estimating defocus blur at edge location and propagated the blur amount to get a dense defocus map. The defocus map reflects depth information well but it is assumed that there exists some blur depending on depth in the input image. In other words, the image should be taken from a camera with the shallow depth of field of a lens and a large aperture. Jung et al. [6] made depth map using relative height cue, in which a closer object in scene is shown in the lower part of the projected image. They detected edge information and traced strong line with rules of the relative height cue. This method is good for usual scenery images but the performance of depth estimation highly depends on the composition of input image. Dimiccoli and Salembier [7] segmented an input image using

depth ordering and T-junctions. T-junction is a structural feature from occlusion, which indicates that an object exists partly in front of another object. This method is good for an image with simple objects but the performance of depth ordering is degraded with the increase of ambiguous T-junction features in outdoor images. Cheng et al. [8] produced depth map using depth hypothesis. They analyzed a geometrical perspective of input image to make a depth hypothesis. Han and Hong [9] constructed a hypothesis for depth map by using vanishing point detection. It is not easy to form a hypothesis that matches with an input image.

There are some methods with prior information. Saxena et al. [10] applied supervised learning to predict a depth map. They collected a training set of monocular image and ground truth depth map of unstructured outdoor images. This method provided satisfactory depth map for outdoor images but it required a lot of training data. Liu et al. [11] presented a learning based method in which seven semantic labels were used to estimate depth map.

These automatic methods are still limited to get enough performance from arbitrary input image. Still in many parts of the 2D-to-3D conversion, progress proceeds with human interaction. Recently, Ward et al. showed that if the estimated depth gives a shape similar to that of an object even though the depth estimation is not accurate, then it is enough to generate 3D effect [12]. They proposed an interactive system for adding depth information to movies, in which depth templates were used to form a 3D shape. This pseudo-3D technique is useful for reducing human intervention.

We propose a depth map generation method for 2D-to-3D conversion using local depth hypothesis. The proposed method is a semi-automatic and simple method with a little intervention from a user.

The rest of the paper is organized as follows. Section 2 describes the proposed method of depth map generation. Experimental results of the proposed depth map generation method are given and discussed in Section 3. Section 4 describes the proposed extensions. Finally, in Section 5 conclusions and future research directions are given.

## 2. PROPOSED DEPTH MAP GENERATION METHOD

We propose an interactive depth map generation method using local depth hypothesis. Figure 1 shows a block diagram of the proposed depth map generation method. It consists of four parts: scene grouping, local depth hypothesis generation, depth assignment, and depth map refinement. Let  $I$  be an input image and  $M$  denote a user input that indicates how to segment  $I$  into several salient regions  $S$ .  $H_{local}$  is the local depth hypothesis and  $G$  represents the grouped image using a graph-based segmentation algorithm [13].  $D_{init}$  signifies the initial depth map whereas  $D_{final}$  denotes a refined final depth map. Description of each part is given in the following.

### 2.1. Scene Grouping

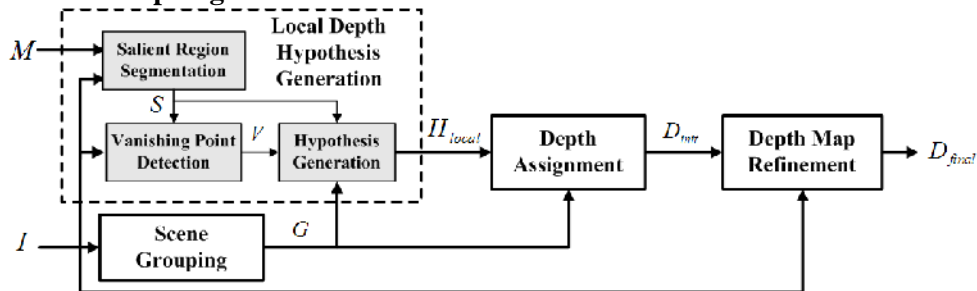


Figure 1. Block diagram of the proposed depth map generation method.

The local depth hypothesis  $H_{local}$  represents general depth transition across salient regions. However, it is not enough to show details of depth variation inside the salient region; depth discontinuities can exist between objects in each salient region. So we need a segmentation to represent detailed depth variations inside the salient region. It is assumed that regions of similar intensity are likely to have similar depth. We use a graph-based segmentation algorithm [13] in grouping similar regions in order to improve salient segmentation and assign the same depth value to the segmented region at the next stage. Figure 2(b) shows the scene grouping result  $G$  of the input image in figure 2(a), in which the same group is represented by the same color. This result can distinguish detailed depth variations in the scene.



Figure 2. Segmentation results of scene grouping. (a) original image (450×340), (b) scene grouping result.

## 2.2. Local Depth Hypothesis Generation

### 2.2.1. Salient region segmentation

In a scene, depth varies gradually. Due to this fact, we make a local depth hypothesis. In depth hypothesis generation block in figure 1, depth hypothesis is generated with structural information of the input image and user interaction. To divide regions of different depth variation, an input image is segmented into two or more salient regions using an interactive graph-cut algorithm [14]. A user defines salient regions of an input image to distinguish depth discontinuity between objects as shown in figure 3(a). As shown in figure 3(b), one of segmented regions (gray region) represents a main object whereas the others (white region and black region) are background. This step segments the input image with a much smaller number of regions compared to figure 2(b). The result of scene grouping and salient segmentation can be different at boundaries. Because the result of scene grouping reflects detail structure, we refine the salient segmentation using scene grouping result. figure 3(c) shows the refined result of salient segmentation.

### 2.2.2. Salient region segmentation

Using structural information, we obtain a cue to build local depth hypotheses  $H_{local}$ . First, we extract lines using Hough transform from an edge map of the input image and detect the vanishing point of each salient region,  $V$  [15]. Edge map can be obtained using Canny edge detector [16]. Lines corresponding to each salient region are used to determine a vanishing point. Vanishing point is determined as a point that minimizes the sum of distances of detected lines [15], which is expressed as

$$\min_{x,y} \sum_{i=1}^N \frac{v_i}{M} (\rho_i - x \cos \theta_i - y \sin \theta_i) \quad (1)$$

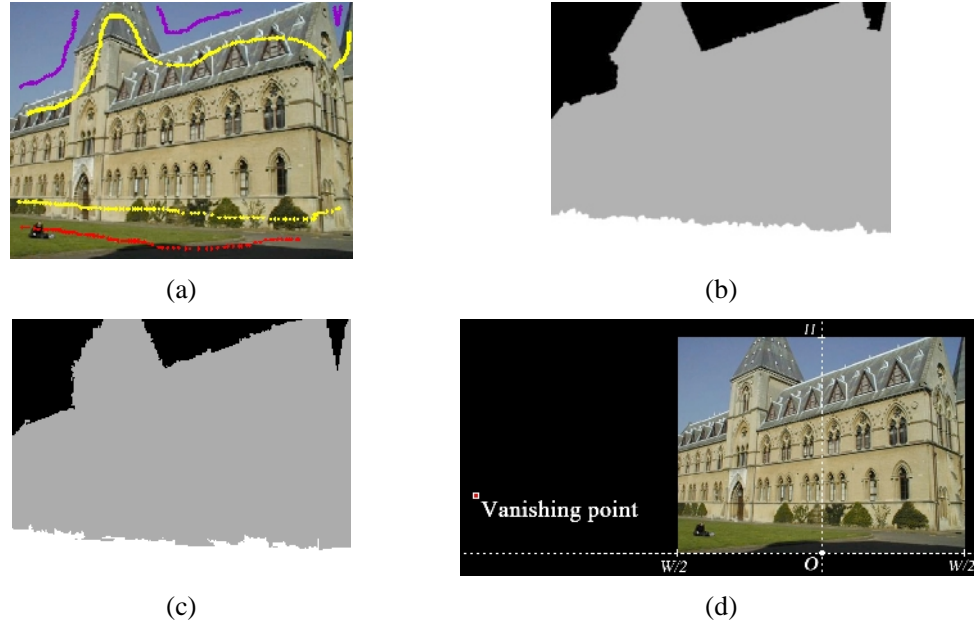


Figure 3. Segmentation of depth salient regions and vanishing point detection. (a) original image (450×340) with user input strikes, (b) salient region segmented image, (c) refined salient region segmented image, (d) detected lines and the vanishing point of gray region in (c).

where  $N$  is the number of detected lines,  $(l_i, l_i)$  represents a set of detected lines,  $v_i$  denotes the number of times that  $i$ -th line pair is observed, and  $M$  signifies the total number of votes in polar coordinate system. Figure 3(d) shows a detection result of lines and the vanishing point of the gray region in figure 3(c).

Before we determine hypothesis of each region, depth information in depth map is expressed in grayscale and the origin is assumed to be at the bottom center of the input image. The brighter the region is, the closer it is located to a camera. The detected vanishing point represents the farthest point in an image. The closer a point to the vanishing point, the farther the point locates from a viewer or a camera. This property is represented in terms of the distance from the vanishing point.

### 2.2.3. Hypothesis generation

Based on the detected vanishing point, local depth hypothesis is formed according to gradual depth variation. The depth hypothesis of each salient region will be determined by the Euclidean distance [17] and relative height depth cue [9]. Euclidean distance is simple and easy to generalize. It has isotropic directional characteristic unlike chessboard distance or city block distance. The Euclidean distance hypothesis of  $k$ -th salient region,  $H_E^k(x, y)$  is determined from the vanishing point as

$$H_E^k(x, y) = \sqrt{(x - x_{vp}^k)^2 + (y - y_{vp}^k)^2} / \max(H_E^k) \quad (2)$$

where  $(x_{vp}^k, y_{vp}^k)$  represents the position of the vanishing point of  $k$ -th salient region.

Relative height depth cue represents that the closer a point in 3D world coordinate to a camera is, the lower the point is projected in a 2D image plane [6]. Natural scene images are generally composed of ground and sky. The ground is shown in lower part of the image whereas the sky in the upper part. Therefore, relative height depth cue hypothesis  $H_R(x, y)$  is determined by  $y$  coordinate alone (independent of the detected vanishing point) as

$$H_R(x, y) = \frac{H - y}{H} \quad (3)$$

where  $H$  is the height of the input image. This hypothesis reflects the gradual variation in depth with the  $y$  coordinate value, that is, gray scale for depth representation changes gradually along the  $y$ -axis.

Our proposed method combines these two depth hypotheses to determine the depth hypothesis of  $k$ -th region  $H^k(x, y)$  as

$$H^k(x, y) = w_E^k H_E^k(x, y) + w_R^k H_R(x, y) \quad (4)$$

where  $w_E^k$  and  $w_R^k$  denote weight coefficients for two hypotheses: Euclidean distance hypothesis and relative height depth cue, respectively. If some salient regions have no detected line or no vanishing point, the depth hypothesis of each region will be determined by relative height depth cue only, i.e.,  $H^k(x, y) = H_R(x, y)$ . Otherwise, depending on where  $k$ -th vanishing point,  $w_R^k$  is determined as

$$w_R^k = \begin{cases} 0, & \text{if } y_{VP}^k \leq 0 \\ y_{VP}^k / 2H, & \text{if } 0 < y_{VP}^k < H \\ \min\{y_{VP}^k / 2H, 1\}, & \text{otherwise} \end{cases} \quad (5)$$

and  $w_E^k$  is computed as  $w_E^k = 1 - w_R^k$ . The larger  $y$  coordinate of vanishing point  $y_{VP}^k$  is located inside the image, the higher the contribution of relative height depth cue is. Using a combination of two basic hypotheses  $H_E$  and  $H_R$  we can make a hypothesis that reflects both depth variation and the location of the detected vanishing point.

Figure 4 shows how the depth hypothesis is determined. Figure 4(a) is Euclidean distance hypothesis  $H_E$  of gray salient region in figure 3(c). The vanishing point of the region is located in the left outside the input image therefore the depth is deeper from right to left. Figure 4(b) shows relative height depth cue hypothesis  $H_R$  and figure 4(c) shows combined depth hypothesis  $H$ . The top left part of figure 4(c) is darker than that of figure 4(a) because the relative height depth cue in figure 4(b) affects depth hypothesis.

Figure 5 shows some examples of combined depth hypothesis  $H$  with respect to location of vanishing point  $V$ . Depending on the location of vanishing point, the weighting factor and the shape of Euclidean distance hypothesis are determined. If the detected vanishing point is far outside of the input image along the positive  $y$ -axis, then the hypothesis is determined by the relative high depth cue alone as shown in Figure 5(a). Figures 5(b)-5(f) show the combined hypotheses which are determined from both basic hypotheses. Figures 5(g)-5(i) show hypotheses that are determined by Euclidean distance alone because the  $y$ -coordinate of the detected vanishing point is lower outside of the input image.

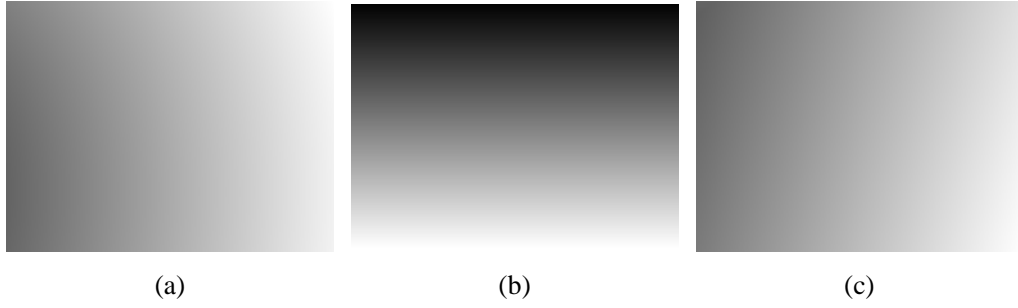


Figure 4. Determination of depth hypothesis. (a) Euclidean distance hypothesis  $H_E$ , (b) relative height depth cue hypothesis  $H_R$ , (c) combined depth hypothesis  $H$ .

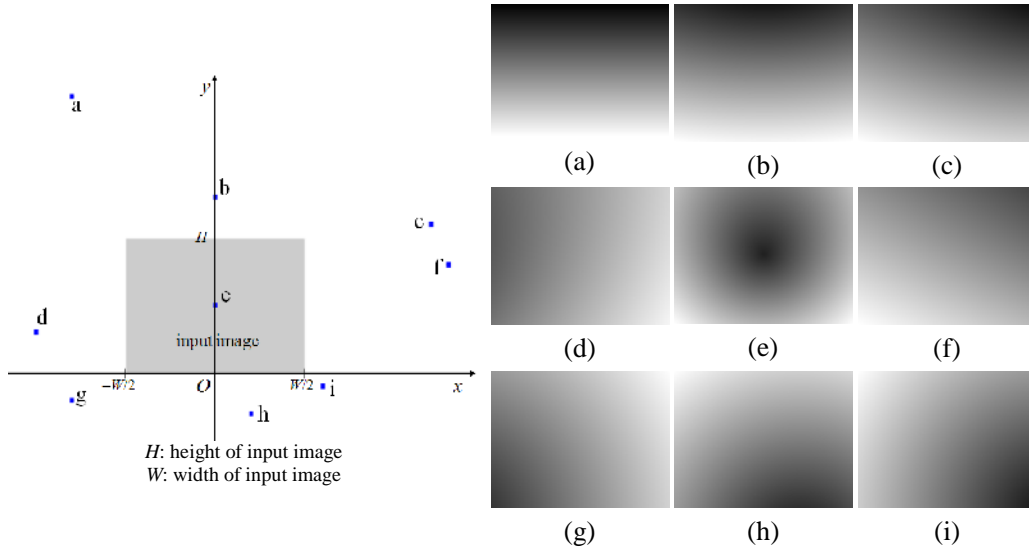


Figure 5. Combination of depth hypotheses depending on the location of detected vanishing point, where the origin is at the bottom center of input image. Left image shows relative position of detected vanishing point. (a)  $(-0.8W, 2.1H)$ , (b)  $(0, 1.3H)$ , (c)  $(1.2W, 1.1H)$ , (d)  $(-W, 0.3H)$ , (e)  $(0, 0.5H)$ , (f)  $(1.3W, 0.8H)$ , (g)  $(-0.8W, -0.2H)$ , (h)  $(0.2W, -0.3H)$ , (i)  $(0.6W, -0.1H)$ .

With the salient region segmented image (figure 3(c)), we combine two types of depth hypotheses (figures 4(a) and 4(b)) to reflect the general tendency of depth transition with salient features preserved. Its result, i.e., the local depth hypothesis  $H_{local}$  is shown in figure 6(a). The local depth hypothesis of other region is made by depth height cue alone because vanishing point is not detected in that region.

### 2.3. Depth Assignment

Using the results obtained in previous steps, we generate initial depth map  $D_{init}$  with  $H_{local}$ . We assign a depth value to each segment group using the local depth hypothesis. The initial depth value at given point  $D_{init}(x, y)$  is assigned by local depth hypothesis  $H_{local}$  and the average depth value in the scene group  $G$ ,  $\overline{H_{local}^{G(x,y)}}$ . The average depth value of the scene group  $G$ ,  $\overline{H_{local}^{G(x,y)}}$  is computed as

$$\overline{H}_{local}^{G(x,y)} = \frac{1}{N_{S(x,y)}} \sum_{(p,q) \in G(x,y)} H_{local}(p,q) \quad (6)$$

where  $N_{S(x,y)}$  represents the number of pixels in a scene group  $S(x,y)$ . In scene grouping process, we assume that regions of similar intensity are likely to have similar depth values. This assumption does not hold if depth variation is large within the same region. Usually, scene group with a large number of pixels has large depth variation, where depth variation is measured in terms of the depth difference between the maximum and minimum depth values. Therefore, we check the size (number of pixels) of the region to classify scene groups. Initial depth map  $D_{init}$  is determined as

$$D_{init}(x,y) = \begin{cases} \overline{H}_{local}^{G(x,y)}, & \text{if } N_{S(x,y)} < N_{th} \\ \frac{1}{2}(H_{local}(x,y) + \overline{H}_{local}^{G(x,y)}), & \text{otherwise} \end{cases} \quad (7)$$

where  $N_{th}$  denotes threshold to detect regions that have large difference between minimum and maximum depth values within the same region. If the scene group is small enough to have the uniform depth, the depth value of the scene group is determined only by the average value of the local depth hypothesis. Otherwise, the depth value of scene group is determined using both the average value of the local depth hypothesis of the scene group and the local depth hypothesis.

Figure 6(b) shows initial depth map  $D_{init}$  of input image. The initial depth map,  $D_{init}$  represents depth information with regard to depth hypothesis while preserving detail information. A group with wide depth range, as indicated by a box in figure 6(b), is not assigned to a constant depth value. A depth value of point A is 153 while that of point B is 204. This result reflects that wide variation of depth hypothesis is also considered as well.

## 2.4. Depth Refinement

In the initial depth map  $D_{init}$ , each region can have a depth value different from those of neighboring pixels though they have similar depth values that belong to the same object in the original image. If one region with the same depth in a real scene is divided into several sub-regions with each different depth value, it can produce unnatural artifacts. So, using cross-bilateral filter [8], [18], the proposed method refines the initial depth map as

$$D_{final}(\mathbf{p}) = \frac{1}{W_p} \sum_{\mathbf{q} \in \Omega(\mathbf{p})} G_{\sigma_s}(\|\mathbf{p} - \mathbf{q}\|) G_{\sigma_r}(\|I(\mathbf{p}) - I(\mathbf{q})\|) D_{init}(\mathbf{q}), \quad (8)$$

where  $\mathbf{p}$  denotes coordinate  $(x, y)$  and  $\mathbf{q}$  is the coordinate of its neighboring pixels, and the normalization constant  $W_p$  is expressed as



Figure 6. Depth hypothesis generation. (a) local depth hypothesis  $H_{local}$ , (b) initial depth map  $D_{init}$ .

$$W_{\mathbf{p}} = \sum_{\mathbf{q} \in \Omega(\mathbf{p})} G_{\sigma_s}(\|\mathbf{p} - \mathbf{q}\|) G_{\sigma_r}(\|I(\mathbf{p}) - I(\mathbf{q})\|). \quad (9)$$

where  $D_{final}$  represents refined depth map,  $G_{\sigma}$  is the Gaussian function with mean  $\mathbf{p}$  and scale  $\sigma$ ,  $\Omega(\mathbf{p})$  denotes neighboring pixels of  $\mathbf{p}$ . The input image is used as a reference image for cross-bilateral filtering. This process can preserve discontinuities in depth, while smoothing regions of similar intensity at the same time.

### 3. EXPERIMENTAL RESULT AND DISCUSSIONS

Experimental results give the intermediate result of the proposed method to show the performance of the proposed method and its application.

Figure 7 shows experimental results with regard to vanishing point that lies inside the image. Figure 7(a) is *Hallmonitor* sequence image used as an input image and figure 7(b) shows a scene grouped image  $G$  and figure 7(c) shows refined salient segmentation result. Figure 7(d) shows local depth hypothesis  $H_{local}$ . This result is obtained with two regions of salient segmentation, people and background. The vanishing point of the background is located in the upper middle part of the image. Figure 7(e) shows depth assignment result  $D_{init}$  from local depth hypothesis (figure 7(d)). This depth map shows that the distance at the end of the corridor is the largest. The ceiling is farther than floor from viewer. This result coincides with human perception.

Figure 8 shows experimental result of influence of salient region segmentation. Figure 8(a) is *Akko&Kayo* sequence image used as an input image and figures 8(b) and 8(c) show a scene grouped result  $G$  and refined salient segmentation result, respectively. In this test image, vanishing point of the white salient region is not detected, thus the local depth hypothesis  $H_{local}$  is determined by relative height depth cue  $H_R$  alone as shown in figure 8(d). The initial depth map in figure 8(e) shows that even the local depth hypothesis  $H_{local}$  is determined by relative height depth cue  $H_R$  alone, the initial depth map can distinguish object and background since the salient segmentation result  $S$  is powerful information from user.

The visual quality of the proposed method depends on result of scene grouping  $G$ . In depth assignment step, the depth is assigned according to scene grouping result. Because detail structures of the depth map follow the result of scene grouping, the scene grouping should segment an image to faithfully represent distinct detail structures.



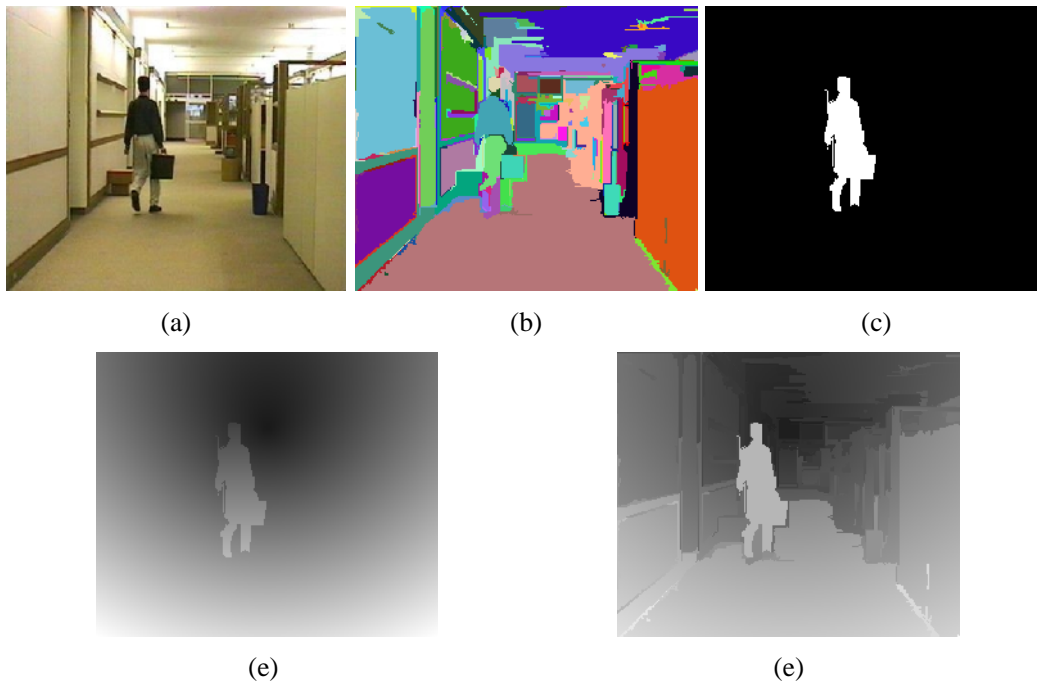


Figure 7. Experimental results. (a) original input image (*Hallmonitor*,  $352 \times 288$ ), (b) scene grouped image  $G$ , (c) refined salient segmentation, (d) local depth hypothesis  $H_{local}$ , (e) initial depth map  $D_{init}$ .

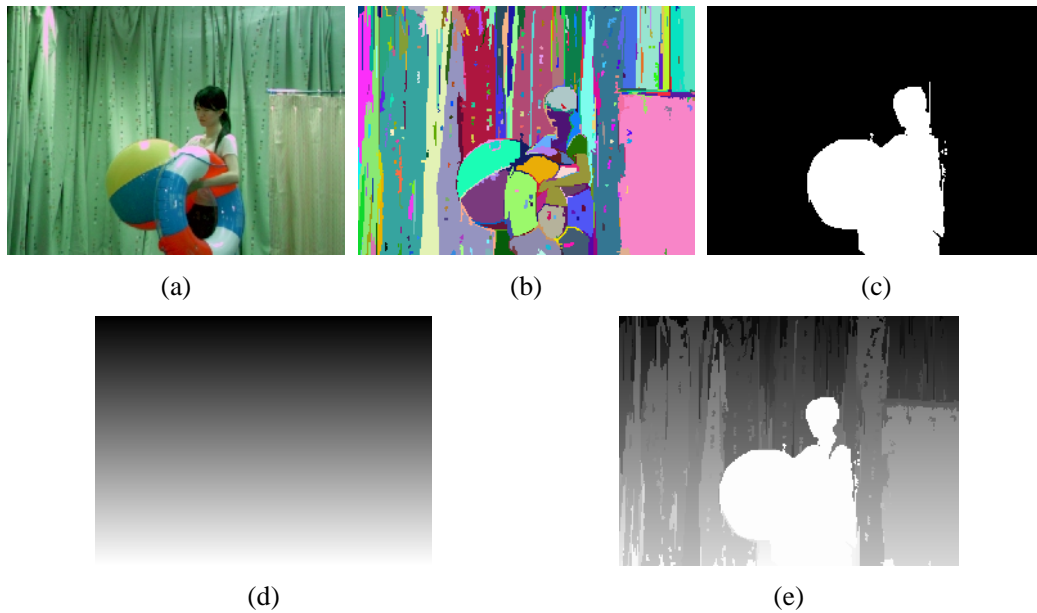


Figure 8. Experimental results. (a) original input image (*Akko&Kayo*,  $640 \times 480$ ), (b) scene grouped image  $G$ , (c) refined salient segmentation, (d) local depth hypothesis  $H_{local}$ , (e) initial depth map  $D_{init}$ .

Figure 9 shows simulation results of the proposed method compared with two existing 2D-to-3D conversion methods [8], [9]. First row shows the depth maps in grayscale whereas second row the depth maps represented by 3D plot. Figures 9(a) and 9(d) show the depth map by Cheng et al.'s algorithm [8] that uses a single (right to left) depth hypothesis. Figures 9(b) and 9(e) show the depth map by Han and Hong's method [9], in which a depth hypothesis is estimated from Gaussian distribution with a vanishing point and height depth cue. Because both methods are global methods with a single depth hypothesis, the results cannot accurately reflect the local depth discontinuity. In the result of Cheng et al.'s algorithm, the right part of the sky appears closer than building. In figure 9(d), the right roof and middle wall is too pop-up rather than neighboring region of wall. In the result of Han and Hong's method, the building is closer along lower part of the building. And two existing methods give a big difference in the right side of outer wall of the building and the left side. On the other hand, as shown in figure 9(c), the proposed method can effectively reflect the local depth transition by the salient segmentation. And figure 9(f) shows the proposed method reflects human perception well. Therefore, the sky appears farther than building and the depth changes gradually in the outer wall. With simple user interaction, we can generate a depth map that preserves both the global and local transitions of depth.

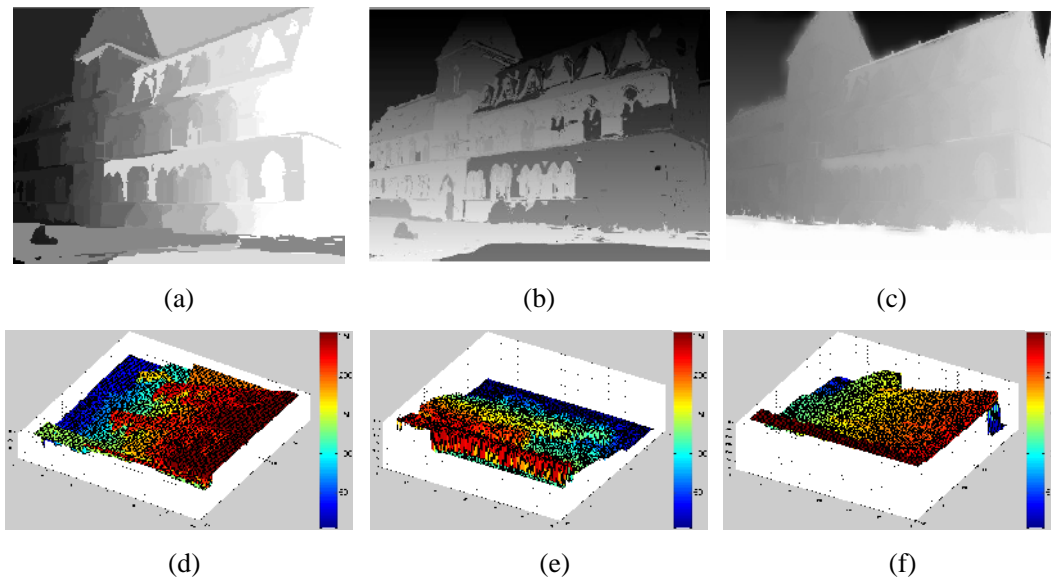


Figure 9. Final depth map. From left to right, Cheng *et al.*'s method [8], Han and Hong's method [9], and proposed method. (a)-(c) depth map represented by grayscale, (d)-(e) depth map represented by 3D plot.

Depth map is used to depth image based rendering. Figure 10 shows anaglyph image that is generated by using input image and depth image. Anaglyph image provides a 3D effect when viewed with red-cyan glasses. To generate an anaglyph image, a single input image and its depth map are used. Using depth information, we can generate stereo pair images from the input image.

The proposed method generates depth map with simple user interaction. The proposed system takes more computation than other methods because our method needs additional salient segmentation process. However, these days cloud computing service is being started. Cloud computing services can reduce the load of hardware implementation. It delivers computing data via Internet to make server compute the process. Therefore, the proposed method can be applied to such consumer devices or applications.

The proposed method can be applied to key frame interpolation for 2D-to-3D video conversion [19]. It generates frame. The depth sequences of non-key frames are interpolated using depth maps of key frames.



Figure 10. Stereo view depth image based rendering result. (a) *Building*, (b) *Hallmonitor*.

#### 4. EXTENSIONS

The proposed method uses user interactions only for distinct depth discontinuities. Based on the proposed method, we can extend the proposed method with more active use of user interaction. For extensions of basic structure of the proposed methods, we consider using user information more explicitly or obtaining more information for depth hypothesis generation. If we use more active user interaction, some processes can be replaced or eliminated. Then, the use of user interaction can be done more efficiently.

First extension is to use the user interaction that can replace the vanishing point detection and hypothesis generation process. With no additional user input, we can consider defining some basic hypotheses. Many scenery scenes can be categorized by several hypotheses. We can consider some hypotheses as basic linear hypotheses that are modified version of relative height depth cue with the direction of gradual depth variation is changed. In [8], five basic hypotheses are proposed. In addition to these basic hypotheses, uniform hypothesis for planar region and Euclidean hypotheses can be used. Figure 11 shows these basic hypotheses. Figures 11(a)-11(c) show three types of linear hypotheses. Figures 11(d) and 11(e) are two types of Euclidean distance hypotheses: convex type and concave, respectively. Figure 11(f) shows uniform hypothesis for background that can be used for the case in which relative emphasis is on an object in the background [20]. When user inputs seeds, user determines which basic hypothesis is matched to the stroked region. This can be easily done in user stroke step with no additional input of different color strokes. This extension directly obtains 3D information from user. Therefore, it prevents errors due to incorrect estimation from vanishing point detection or hypothesis generation.

Or, we can consider taking the vanishing point with additional user input. User can easily provide position of the vanishing point by pointing a position or input a coordinate. This can eliminate the vanishing point detection step.

Figure 12 shows experimental results of the proposed methods: basic structure (figures 12(a)-12(d)), extension 1 (figures 12(e)- 12(h)), and extension 2 (figures 12(i)-12(l)). In extension 1, user selects bottom to top (figures 11(a)), right to left (figures 11(b)), and uniform (figures 11(f)) hypothesis for Building image whereas concave Euclidean and uniform hypothesis for *Akko&Kayo* sequence image. In extension 2, user gives location of vanishing point that is left outside of the image for gray salient region of building image. For *Akko&Kayo* sequence image,

user selects location of vanishing point of whiter region which is inside the white salient region. As we use user interaction actively, the performance of the proposed method is enhanced with the

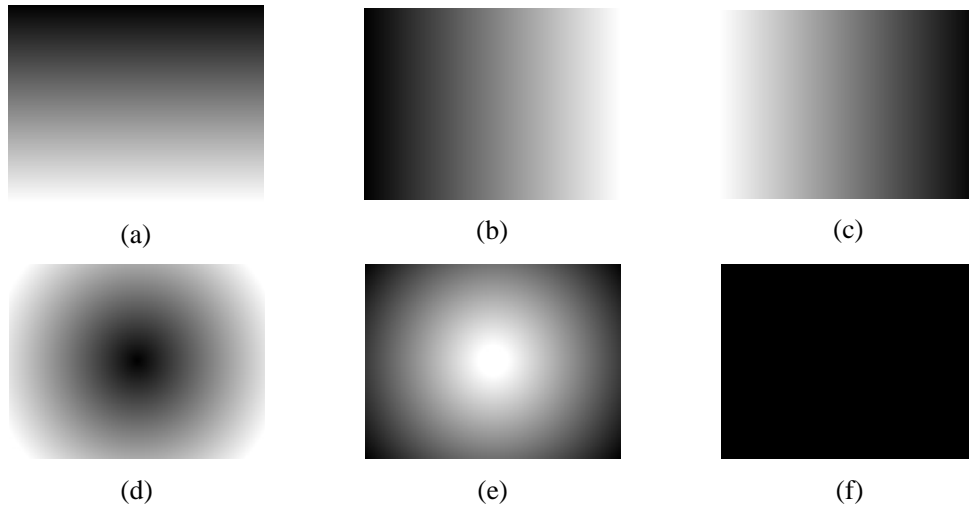


Figure 11. Basic hypotheses. (a) bottom to top linear hypothesis (same as relative height depth cue), (b) right to left linear hypothesis, (c) left to right linear hypothesis, (d) convex Euclidean hypothesis, (e) concave Euclidean hypothesis, (f) uniform hypothesis for background.


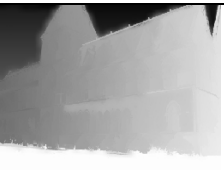
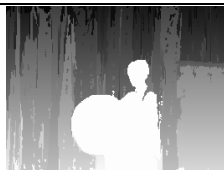


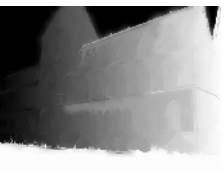
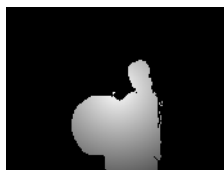
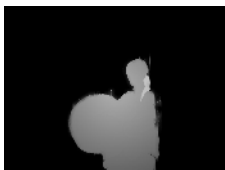

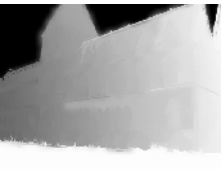
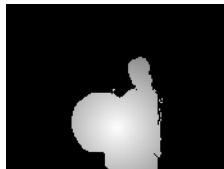
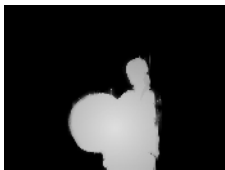
Methods	<i>Building</i>		<i>Akko&amp;Kayo</i>	
	$H_{local}$	$D_{final}$	$H_{local}$	$D_{final}$
Basic structure				
	(a)	(b)	(c)	(d)
Extension 1				
	(e)	(f)	(g)	(h)
Extension 2				
	(i)	(j)	(k)	(l)

Figure 12. Comparison of experimental results of basic structure and extensions. viewpoint of human perception.

Table 1 shows a comparison of the proposed method (basic structure) and its extensions. Extension 1 gets hypotheses information from user whereas extension 2 acquires vanishing point information from user. To get prior information directly from user can reduce some process to determine hypothesis. It significantly improves the accuracy of depth hypothesis without unduly increasing user interaction.

Table 1. Comparison of Basic Structure and Extensions.

Methods	Type of Basic Hypotheses	Information Provided by User Interaction	Necessity of Vanishing Point Detection
Basic structure	1 relative height depth cue 1 adaptive Euclidean	Depth discontinuity only	O
Extension 1	5 basic linear [8] 2 Euclidean 1 uniform background	Depth discontinuity and depth hypothesis	O
Extension 2	2 adaptive Euclidean 1 uniform background	Depth discontinuity and determination of adaptive depth hypothesis	X

## 5. CONCLUSIONS

This paper proposes a depth map generation method from a single image for 2D-to-3D conversion with user interaction. The proposed method combines depth hypotheses with the salient segmented image, and refines the initial depth map using a cross-bilateral filter. The proposed depth map maintains salient depth values and local transition of depth. It can generate natural depth map from the viewpoint of human perception and be easily applied to video interpolation for 2D-to-3D conversion. With additional considerations of the use of user interaction, the proposed method can be extended in many ways. Future research will focus on reducing human intervention, so that ultimately, the proposed depth map generation method can be automated.

## ACKNOWLEDGEMENTS

This work was supported in part by the Second Brain Korea 21 Project.

## REFERENCES

- [1] Kauff, P., Atzpadin, N., Fehn, C., Müller, M., Schreer, O., Smolic, A., & Tanger, R., (2007) "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability", *Signal Process.: Image Commun.*, Vol. 22, No. 2, pp217-234.
- [2] Shin, H.-C., Kim, Y.-J., Park, H., & Park, J.-I., (2008) "Fast view synthesis using GPU for 3D display", *IEEE Trans. Consumer Electronics*, Vol. 54, No. 4, pp2068-2076.
- [3] Jung, C. & Jiao, L. C., (2011) "Disparity-map-based rendering for mobile 3D TVs", *IEEE Trans. Consumer Electronics*, Vol. 57, No. 3, pp1171-1175.
- [4] Lai, S.-H., Fu, C.-W., & Chang, S., (1992) "A generalized depth estimation algorithm with a single image", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 14, No. 4, pp405-411.
- [5] Shuo, S. & Sim, T., (2011) "Defocus map estimation from a single image", *Pattern Recognition*, Vol. 44, No. 9, pp1852-1858.

- [6] Jung, Y. J., Baik, A., Kim, J., & Park, D., (2009) "A novel 2D-to-3D conversion technique based on relative height depth cue", in *Proc. Stereoscopic Displays & Applications XX*, Vol. 7237, doi: 10.1117/12.806058.
- [7] Dimiccoli, M. & Salembier, P., (2009) "Exploiting T-junctions for depth segregation in single images", in *Proc. IEEE Conf. Acoustics, Speech, & Signal Processing*, pp1229-1232, Taipei, Taiwan.
- [8] Cheng, C.-C., Li, C.-T., & Chen, L.-G., (2010) "A novel 2D-to-3D conversion system using edge information", *IEEE Trans. Consumer Electronics*, Vol. 56, No. 3, pp1739-1745.
- [9] Han, K. & Hong, K., (2011) "Geometric and texture cue based depth-map estimation for 2D to 3D image conversion", in *Proc. IEEE Int. Conf. Consumer Electronics*, pp651-652, Las Vegas, NV, USA.
- [10] Saxena, A., Chung, S. H., & Ng, A. Y., (2006) "Learning depth from single monocular images", *Advances in Neural Information Processing Systems 18*, Y. Weiss & B. Sch. Ed. Cambridge, MIT Press, pp161-1168.
- [11] Liu, B., Gould, S., & Koller, D., (2010) "Single image depth estimation from predicted semantic labels," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp1253-1260, San Francisco, CA, USA.
- [12] Ward, B., Kang, S. B., & Bennett, E. P., (2011) "Depth director: A system for adding depth to movies", *IEEE Computer Graphics and Applications*, Vol. 31, No. 1, pp36-48.
- [13] Felzenszwalb, P. F., & Huttenlocher, D. P., (2004) "Efficient graph-based image segmentation", *Int. J. Computer Vision*, Vol. 59, No. 2, pp167-181.
- [14] Li, Y., Sun, J., Tang, C.-K., & Shum, H.-Y., (2004) "Lazy snapping", *ACM Trans. Graphics*, Vol. 23, No. 3, pp303-308.
- [15] Cantoni, V., Lombardi, L., Porta, M., & Sicard, N., (2001) "Vanishing point detection: Representation analysis and new approaches", in *Proc. 11th Int. Conf. Image Anal. and Process*, pp90-94, Palermo, Italy.
- [16] Canny, J., (1986) "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 8, No. 6, pp679-698.
- [17] Gonzalez, R. C. & Woods, R. E., (2010) *Digital Image Processing, Third edition*, Upper Saddle River, Pearson Education.
- [18] Jachalsky, M., Schlosser, & Gandolph, D., (2010) "Confidence evaluation for robust, fast-converging disparity map refinement", in *Proc. IEEE Conf. Multimedia and Expo*, pp1399-1404, Gold Coast, Australia.
- [19] Lie, W.-N, Chen, C.-Y., & Chen, W.-C, (2011) "2D to 3D video conversion with key-frame depth propagation and trilateral filtering," *Electron. Lett*, Vol. 47, No. 5, pp319-321.
- [20] Kim, J., Choe, Y., & Kim, Y., (2011) "High-quality 2D to 3D video conversion based on robust MRF-based object tracking and reliable graph-cut-based contour refinement", in *Proc. Int. Conf. Information and Communication Technology Convergence*, pp360-365, Seoul, Korea.

**Authors**

**Na-Eun Yang** received the B.S. degree in electronic engineering from Sogang University in 2011. Currently she is working toward the M.S. degree in electronic engineering from Sogang University. Her current research interests are image processing and image enhancement.



**Ji Won Lee** received the B.S. degree in physics from Sookmyung Women's University in 1999. She received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sogang University in 2004, 2006, and 2011, respectively. In 2011, she joined LG Electronics Co., Ltd. Her current research interests are image processing and resolution enhancement.



**Rae-Hong Park** was born in Seoul, Korea, in 1954. He received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, Korea, in 1976 and 1979, respectively, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1981 and 1984, respectively. In 1984, he joined the faculty of the Department of Electronic Engineering, School of Engineering, Sogang University, Seoul, Korea, where he is currently a Professor. In 1990, he spent his sabbatical year as a Visiting Associate Professor with the Computer Vision Laboratory, Center for Automation Research, University of Maryland at College Park. In 2001 and 2004, he spent sabbatical semesters at Digital Media Research and Development Center, Samsung Electronics Co., Ltd. (DTV image/video enhancement). His current research interests are computer vision, pattern recognition, and video communication. He served as Editor for the Korea Institute of Telematics and Electronics (KITE) Journal of Electronics Engineering from 1995 to 1996. Dr. Park was the recipient of a 1990 Post-Doctoral Fellowship presented by the Korea Science and Engineering Foundation (KOSEF), the 1987 Academic Award presented by the KITE, the 2000 Haedong Paper Award presented by the Institute of Electronics Engineers of Korea (IEEK), the 1997 First Sogang Academic Award, and the 1999 Professor Achievement Excellence Award presented by Sogang University.

