2D-to-3D Conversion System using Depth Map Enhancement

Ju-Chin Chen and Meng-yuan Huang
Department of Computer Science and Information Engineering
National Kaohsiung University of Applied Sciences, Kaohsiung, Kaohsiung, Taiwan, ROC
[e-mail: jc.chen@cc.kuas.edu.tw]
*Corresponding author: Ju-Chin Chen

Received June 15, 2015; revised December 19, 2015; accepted January 10, 2016; published March 31, 2016

Abstract

This study introduces an image-based 2D-to-3D conversion system that provides significant stereoscopic visual effects for humans. The linear and atmospheric perspective cues that compensate each other are employed to estimate depth information. Rather than retrieving a precise depth value for pixels from the depth cues, a direction angle of the image is estimated and then the depth gradient, in accordance with the direction angle, is integrated with superpixels to obtain the depth map. However, stereoscopic effects of synthesized views obtained from this depth map are limited and dissatisfy viewers. To obtain impressive visual effects, the viewer's main focus is considered, and thus salient object detection is performed to explore the significance region for visual attention. Then, the depth map is refined by locally modifying the depth values within the significance region. The refinement process not only maintains global depth consistency by correcting non-uniform depth values but also enhances the visual stereoscopic effect. Experimental results show that in subjective evaluation, the subjectively evaluated degree of satisfaction with the proposed method is approximately 7% greater than both existing commercial conversion software and state-of-the-art approach.

Keywords: 2D-to-3D conversion system, significant region detection, visual attention

http://dx.doi.org/10.3837/tiis.2016.03.012 **ISSN : 1976-7277**

1. Introduction

Recently, 3D signal processing has attracted considerable attention in the field of computer vision. Compared with 2D displays, 3D displays provide a more realistic visual experience. Hence 3D display is regarded as the next revolution for many applications including the entertainment industry, multimedia systems, and broadcasting. For instance, since 2005, starting with Chicken Little, 3D movies have become popular, and have brought substantial business benefits. Box office receipts for Avatar amounted to over 25 billion dollars worldwide. Moreover, since 2010, many TV brands such as SONY, Panasonic, and Samsung have begun manufacturing 3D TVs for home theater systems.

For 3D display, several approaches have been employed to obtain 3D content, including active depth sensing [1] and stereo vision recording [2]. Active depth sensing applies sensors such as structured light and time-of-flight sensors [3] to estimate depth information [1]. Stereo vision recording relies on multiple cameras to capture multiple views and estimate depth information using stereo matching [2]. However, these methods require additional complex devices and the 3D visual effect is limited to a home 3D TV because of minor disparities [4]. In addition, these methods are only suitable for new production. Reproducing 3D visual effects for existing 2D image/video data is difficult because manual editing of the depth information is time-consuming. The lack of high-quality 3D content has become a bottleneck for the growth in the 3D industry [4], [5]. Therefore, 2D-to-3D image/video conversion algorithms have been proposed in recent years that can be categorized into semi-automatic methods, which need user interactive operations [6-8] and fully-automatic methods, which output 3D contents without any user interactions [9-10]. In most semi-automatic 2D-to-3D conversion frameworks, a small number of video frames, i.e., key frames, of the video sequence are annotated with depth information by users, and the rest of the video frames, i.e., non-key frames, are converted to 3D automatically [5], [7-8]. In accordance with cognitive studies showing that the human visual system is sensitive to foreground objects, in [8], an interactive step was added for foreground depth refinement. Compared with fully-manual conversion systems, semi-automatic conversion methods can provide more reliable results than fully-automatic methods. However, human participation is impractical in many scenarios [5]. Therefore, fully-automatic 2D-to-3D conversion methods need to be developed.

According to human depth perception mechanisms, several 2D-to-3D conversion techniques [1-2], [4], [9-14] have been proposed which mainly focus on recovering the depth map. Humans can integrate various depth cues to generate depth perception, including monocular depth cues such as motion, focus/defocus cues, and relative height/size cues, to perceive the relative distance of objects within a real scene and binocular depth cues from a visual system to realize the 3D location of an object [2], [14]. A survey of automatic 2D-to-3D systems and depth map generation can be found in [9]. In addition, there are a few software packages that can retrieve depth maps and provide automatic 2D-to-3D conversion, such as DDD's TriDef 3D [15] and ArcSoft's Media Converter [16]. However, the stereoscopic visual effect produced by these tools is not obvious because of limited information [10]. Recently, learning-based methods either based on inferring depth maps from extra images (data-driven approaches) [11], [13], [17] or estimating such maps with model learning [18-21] have attracted much attention. Given a large dataset consisting of either stereopairs [11] or image + depth pairs [13], [17], data-driven approaches infer the depth map of a 2D query by finding the most similar images in the dataset and fusing (or warping) corresponding depth maps for final

depth estimation. Note that instead of a general case, a domain-specific conversion system was developed in [11] to provide better results. Moreover, a supervised learning strategy, i.e., Markov random field [18-19] or deep learning [20-21], was used to learn the relation between 3D structures and 2D image features. Rather than extracting low-level features such as texture and color as in [19], deep learning approaches require neither hand-crafted features nor assumptions about the semantic information of a scene.

Overall, 2D-to-3D conversion problems face some challenges to generate pleasing 3D effects. First, depth values inside the same object need to be uniform [2]. Hence, pre-processing has to be performed by considering both color similarity and spatial distance to segment an input image into subregions [2]. The more complete (or detailed) the pre-processing results, the more uniform the depth values inside the object. Second, the depth relationships between all objects are considered [2]. In [14] and [22], the authors use motion parallax as the depth cue and integrate the depth information with the object grouping results. Third, the resulting 3D visual effect is often not sufficiently impressive for viewers because of minor disparities. Most conversion systems focus on generating precise depth values based on human depth perception, however, visual attention is ignored and it is an important factor for pleasing perception. Recently, visual attention models have been discussed in related research areas, such as photo quality assessment [23], region of interest [24-24], and saliency object detection [25-26], which all describe the elements of a visual scene that are likely to the attract attention of humans. For example, Sun et al. [23] proposed one visual attention model for constructing a face-sensitive saliency map and a rate of focused attention measurement was proposed for quality assessment. Based on the observation that humans pay more attention to those image regions that have contrast with their surroundings, Cheng et al. [25] proposed a regional contrast based saliency extraction algorithm, which can simultaneously consider global contrast differences and spatial coherence. Moreover, not only static information is explored, a dynamic attention model is proposed as well. Zhang et al. [26] learned the shifting path of human gaze, called the active graphlet path, to mimic the process of humans looking at one photo. According to the semantic significance the subregion of a photo is selected for photo cropping.

In this study, we propose a single-image-based 2D-to-3D conversion system to solve the aforementioned problems. The main contribution to modification of the depth map comes from a fusing depth gradient estimation and salient object detection. These give an accurate depth gradient while enhancing visual stereoscopic effect. Instead of estimating depth values from the depth cues directly, the perspective model and the atmospheric scattering model are first employed to estimate the captured direction from five hypotheses [1]. To avoid inconsistent depth assignment for the same object, the minimum-spanning-tree (MST) is applied to group pixels having similar colors and spatial locality, to facilitate object segmentation [1]. Hence, the initial depth map can be generated by integrating the depth gradient in accordance with the hypothesis and segmentation results. Furthermore, to improve the visual effects, salient region detection based on graph-cut regions is used to generate a binary salient map. Then the depth map for the salient region is locally modified within this region. Finally, depth-image-based rendering (DIBR) uses backward mapping and bilinear interpolation to generate a left and right stereoscopic image pair. By refining the depth map, we can enhance the 3D visual perception for humans on the generated stereoscopic image pair.

The remainder of this paper is organized as follows. In Section 2 we review the current 2D-to-3D conversion system and depth estimation approaches. The overview of the proposed system is introduced in Section 3, and the detail of each system module is presented in Section 4. Section 5 describes a subjective perception study performed to evaluate the stereoscopic

effect of the proposed method and compares it with those of state-of-the-art methods. We draw conclusions and present discussion in Section 6.

2. Related Work

Generation of 3D image/video from 2D image/video has been studied for many years [1-2], [4], [9-14], [28]. These approaches focus on estimating depth information either from a single image or multiple images and filling holes in the synthesis process of stereoscopic views. Tam et al. [13] transferred one input image into a YCbCr color space and then the values of the Cr channel were used as a depth map. Han et al. [29] estimated a depth map based on both geometric and texture cues from a single image. By detecting line features, the geometric cue, i.e., the vanishing point, was used to generate the initial depth map. Then the texture cue obtained from the segmentation results was used to refine the depth map for an accurate result. However, the uniformity of the depth value within the same object is not guaranteed with the obtained texture cue. Jung et al. [30] used gradient and linear perspective cues for depth map estimation. Rather than estimate a depth value from the depth cue. Cheng et al. [2] proposed a five depth gradient hypotheses for depth assignment. Before assigning depth values, the MST is applied to group pixels based on their color and spatial locality. Then according to a depth hypothesis, a relative depth value is assigned to each region. To remove blocky artifacts, cross bilateral filtering is applied to enhance the visual comfort. Note that this work is the first to considering both monocular and binocular cues. Finally, the depth map is fed into DIBR to synthesize the stereoscopic image pair. On the other hand, to obtain more reliable results, Guttmann et al. [7] proposed a semi-automatic approach with user interaction. Although the system can achieve good results, many complex processes are needed to obtain the final depth map. Thus, to reduce the computation complexity, in [31], a hybrid paradigm is proposed that random walks [32] and graph cuts [33] are used to generate a final cohesive depth map. Via user-defined strokes, which are seen as a rough estimation of the depth values, the proposed system can estimate the depth values for the rest of the image. Compared with one single image input, a video can provide more depth cues for depth map generation. Lin et al. [34] proposed a 2D-to-3D video conversion scheme for MPEG videos. Because the memory size for video is massive, the motion cue is extracted directly from the MPEG bit stream to reduce computation complexity. The other depth cues, such as atmospheric perspective, texture gradient, linear perspective, and relative height, are obtained from decoded frames. Kim et al. [35] proposed an accurate depth map generation scheme. After performing MRF-based contour tracking, the graph-cut segmentation is applied to refine the contour to repair tracking errors in the complex background. However, most video conversion systems are off-line processes due to manual interactions or complex computation. Tsai et al. [1] proposed a real-time 2D-to-3D video conversion system that is implemented on both software and hardware for optimization. Using unified streaming dataflow, multi-thread schedule synchronization, and CUDA acceleration, a 1920 × 1080p at 30 fps video conversion is achieved.

Among these 2D-to-3D conversion studies, depth estimation is a key technique that can affect the quality of synthesized virtual images. However, depth estimation from a single image is an ill-posed problem, since the true 3D structure is ambiguous in that a given image might be generated from an infinite number of 3D objects [18]. Without prior knowledge of the scene, depth estimation cannot be carried out. However, this task is not difficult for human beings who can infer a 3D structure via a stereo vision system and prior knowledge. Saxena et al. [18-19] summarized the human visual cues for 3D scene understanding into four categories:

monocular, stereo, motion parallax, and focus cues. Therefore, depth estimation algorithms are developed based on these visual cues for both single-image-based and multiple-image-based methods. The former methods estimate the depth map based on monocular cues that can be extracted from one image such as image classification results [36], geometric perspective [37], texture gradient [34], atmospheric perspective [38], and relative heights [39]. The latter methods, developed based on stereo cues, motion parallax, and focus cues [40], require more than one image that can be acquired from multiple cameras. The stereo cue takes the disparity variance to estimate an object's distance that is inversely proportional to the distance of the object from cameras. In other words, a distant object gets smaller disparities than a close object. Motion parallax [41] is based on the fact that the observed motion difference of a close object is larger than that of a distant object if they travel with the same velocity, and thus one can estimate the relative distances in a scene. However, if the object or camera is static, depth cannot be estimated. Additionally, motion information and object segmentation are assumed to be known, and hence, they are suitable only for images with simple backgrounds. In this study, we focus on visual perception enhancement for single images; more in-depth discussion concerning multiple-image-based methods has been discussed elsewhere [2], [41].

Assuming objects with uniform color or texture, algorithms of shape from shading [42] and shape from texture [43], are developed. However, the algorithms are not capable of handling complex images [18]. Rather than relying on the image size of specific objects, by studying the Fourier spectrum for different scenes, Torralba and Oliva [44] estimates absolute mean depth for the scene by recognizing the structure properties in the image. In [2], the camera direction is estimated according to the location of the vanishing point. The limitation of this method is that the depth estimation fails if lines are not detected in the image. Similar to the linear perspective method, the texture gradient estimates depth according to the cue that distant objects look smaller and more compact [34]. However, these methods are only suitable for scenes with regular or similar objects such as flowers. Jung et al. [39] used the relative height of objects to assign depth maps. For scenes with similar objects, the objects observed in the upper parts of the images are relatively distant. Subsequently, the distance between the objects and line boundaries can be used to estimate the depth for each pixel. However, for this estimation, more complicated preprocessing is required, such as object segmentation or salient object detection. Su et al. [38] used motion difference to first segment the image into foreground and background, and later, linear perspective, atmospheric perspective, and relative horizontal height results were fused to estimate depth information. In [37], the texture gradient was meticulously applied to estimate scene complexity, and the combination weights were estimated according to the least square error. Then, the depth information derived from motion difference, atmospheric perspective, and texture gradient results were appropriately fused to obtain reliable results.

On the other hand, instead of developing algorithms from heuristic assumptions, Saxena et al. [18] proposed a learning-based approach to deduce the 3D structure from a training set of monocular images that consisted of unstructured indoor and outdoor views with ground-truth depth maps. The authors observed that local features are not sufficient for depth estimation, and thus a hierarchical, multi-scale Markov random field (MRF) was used to model not only local features but the relationship between different parts of the image. Furthermore, instead of dividing the image into small rectangular patches, the authors modified their work by applying image segmentation to obtain superpixels to satisfy the planar assumption [19] for reliable depth estimation**.** By observing the significant progress in image classification [45] and object detection [46] brought by deep learning algorithms, learning-based approaches based on deep

convolutional neural networks (CNNs) for single-image depth estimation have rapidly developed in recent years. In 2014, Tian et al. [47] proposed a depth inference model relying on a CNN that contained several convolutional and pooling layers as the basic architecture as well as a linear regressor as the last layer for depth value inference. Compared with graphical-model-based methods [18-19], the proposed method requires neither engineered features nor assumptions about the semantic information of a scene. It can provide results that are competitive with [19] in terms of low computational complexity in a test time. In addition, Eigen et al. [20] trained two CNNs, i.e., the coarse- and fine-scale networks, for depth map prediction. The coarse-scale network first estimates the global structure of a scene, and the fine-scale network edits coarse prediction results to align with local details. Note that the depth map is directly estimated by the CNN and a large amount of labeled data must be collected so that networks can be trained with all possible layouts [48]. Unlike the method of directly estimating depth values presented in [20], [47], Liu et al. [21], [48] proposed a deep convolutional neural field model to formulate the depth estimation as a deep continuous conditional random field (CRF) learning problem in which CRF is explicitly used to model the relations of neighboring superpixles, and potential functions are learned in a unified CNN framework. This is a prior work to explore CNN for structured learning problem with a graphical model. Because the translation invariance is preserved, no superpixel coordinate needs to be encoded, and hence, compared with [20], the methods in [21], [48] can train a network using a standard dataset to obtain competitive performance without additional training data or any geometric prior. The depth of a new test image can be estimated via the MAP inference with a closed-form solution. According to the experimental results, deep-learning based methods [21], [48] can outperform state-of-the-art results [17], [20], [49] for both indoor and outdoor scene datasets.

Stereo image synthesis is another important issue for 2D-to-3D conversion. In the field of computer vision, virtual view synthesis can be roughly categorized into model-based rendering (MBR) and image-based rendering (IBR) [50]. MBR needs to construct a 3D model to render a virtual view [51], whereas IBR does not require 3D details of a scene in the generation process. IBR aims at synthesizing virtual views from images even when no geometric information is given. Famous techniques such as light field [52] and lumigraph [53] have been proposed; the former interprets inputted images as 2D slices of a 4D function to characterize the flow of light in a static scene, whereas the latter uses a subset of plenoptic functions to describe the flow of light in all directions. In recent years, some warping-based methods that do not require depth maps have formulated the render process as an optimization problem [54]. In addition, the depth-image-based rendering (DIBR) proposed in advanced three-dimensional television system technologies [55] can generate two views using a single 2D image as well as its corresponding depth image, which gives a depth value for each pixel. Because of the attractive features of DBR, such as its efficient computation ability, most 2D-to-3D conversion systems apply DIBR to synthesize disparate images. DIBR mainly consists of three processes: disparity computation, pixel shifting, and hole filling. However, two problems occurred in the synthesized image, which are termed occlusion and disocclusion [4], [56]. Occlusion means that two different pixels have warped to the same location in the synthesized images, and this can be solved by generating the image using pixels closer to the camera [4]. However, disocclusion is more difficult to solve because it is caused by occlusion in the original input video. As a result, no information can be provided to generate these pixels in the synthesized images, referred as a hole. Thus, a hole-filling process [4], [57-58] such as bilinear interpolation or preprocessing the depth map is required to reduce the net area of holes in the image. For example, Wang et al. [56] proposed an asymmetric edge adaptive filter

(AEAF), inspired by the bilateral filter, for depth map generation and hole filling. Via asymmetric smoothing of depth maps, AEAF can fill the area of holes in synthesized images to reduce artifacts and distortions and preserve object edges simultaneously.

3. Overview of the proposed 2D-to-3D conversion system

Generally, the main goal of 2D-to-3D conversion systems is to generate a stereoscopic image pair of a given 2D image based on an estimated depth map. **Fig. 1** presents a flowchart of the proposed system. As shown in the figure, to estimate the scene depth from a single image, linear and atmospheric perspective cues are retrieved from a given input image via line detection and blurred degree estimation, respectively. In the process, the capturing direction is defined based on the five hypothetical directions [1]. To cope with inconsistent depth assignment for the same object, the MST is applied for object segmentation [1]. Hence, the initial depth map can be generated by fusing the depth gradient and segmentation results. For most 3D videos, the stereoscopic effect for human perception is limited. To enhance visual perception, the salient region, which represents the main visual attention of the viewer, is further evaluated. Then the depth assignment can then be refined for the salient region and its surrounding pixels. Note that incorporation of the salient region can not only enhance the stereoscopic degree but also correct the depth consistency of an object. Finally, using the input image and the refined depth map, a stereoscopic image pair can be synthesized via DIBR and a hole filling algorithm performed to refine the results.

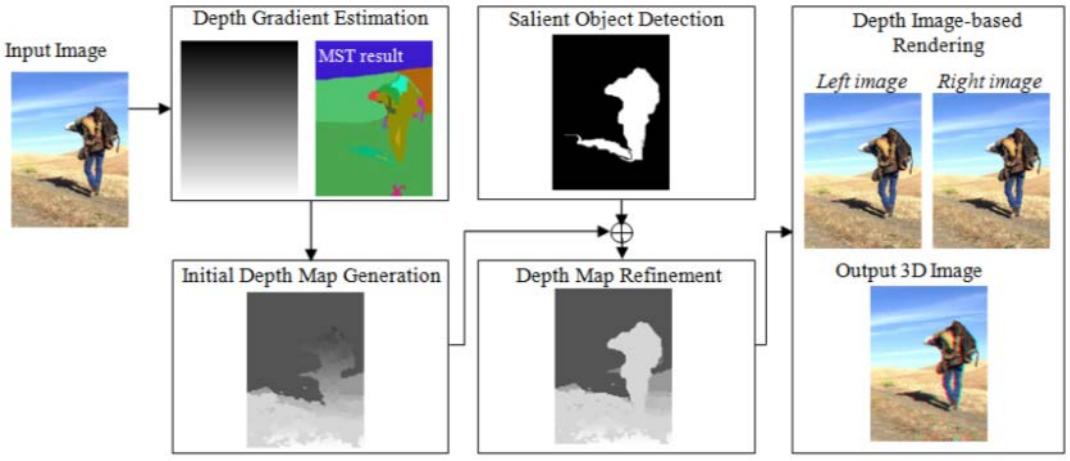


Fig. 1. System flowchart.

4. 2D-to-3D Conversion System with Depth Map Enhancement

As shown in **Fig. 1**, the proposed system commences with depth gradient estimation. After estimating the depth gradient and segmenting objects, the initial depth map is generated. By integrating the salient region, the depth map is further refined to enhance the stereoscopic effect. The approach employed in our study involves not only enhancement of visual perception but also correction of the depth map for consistency within objects. The details of the system are discussed in the following subsections.

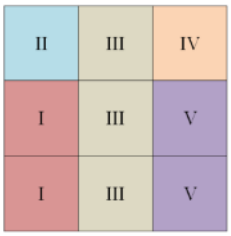


Fig. 2. Divided regions of an image that correspond to the five hypothetical directions representative of the capturing direction.

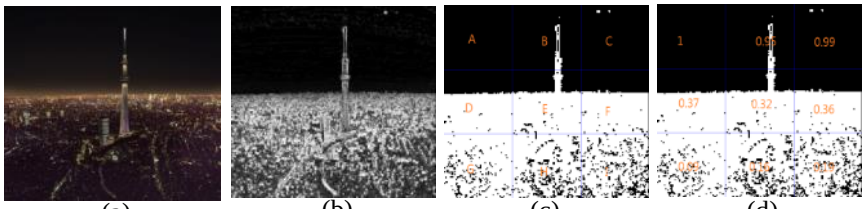
Hypothesis	Type 1	Type 2	Type 3	Type 4	Type 5
Location of Vanishing Point	I	II	III	IV	V
Capturing Direction	←	↖	↑	↗	→
Direction Angle θ	0°	45°	90°	135°	180°
Depth Gradient					
Image Example					

Fig. 3. Five hypothetical directions and their corresponding depth gradients.

4.1 Depth Gradient Estimation

To recover depth information from an input image, two cues are used to estimate the scene depth: linear and atmospheric perspective cues. The former is based on the perspective model of human visual perception, whereas the latter is based on the atmospheric scattering model wherein far objects are blurred by particles in the atmosphere.

For evaluating static or moving scenes, five directions representing the viewer's position relative to a scene can be roughly classified as left to right, right to left, bottom to top, left-bottom to right-top, and right-bottom to left-top. According to the perspective property, the Hough transformation is first applied to detect lines in the image. Then five hypothetical directions are defined according to the position of the vanishing point to represent the capturing direction [1] with a direction angle based on the perspective cue θ_p. **Fig. 2** illustrates the divided regions of an image and the corresponding hypothetical directions. For example, if the vanishing point is in region IV, the hypothesis is set to Type 4, the direction angle is estimated as right to left, and $\theta_p = 135°$. We note here that if the location of the vanishing point is inaccurate because of inconsistent line detection results or no line detection, the default hypothesis is Type 3 in which is the most likely capturing direction. The five hypothetical directions and their corresponding depth gradients are summarized in **Fig. 3**.



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

Fig. 4. Depth information from the atmospheric cue: (a) input image, (b) contrast value C(u,v) (white regions have a larger contrast value), (c) binary result, and (d) ratio of the number of black pixels relative to the number of pixels in each block.

The perspective cue works well when the scene contains objects with parallel lines, for example, roads and buildings. To generate a convincing depth map, the atmospheric cue [34] is also applied. The light reflected from a far object is scattered by particles in the atmosphere; thus, distant objects appear blurred, and close objects appear sharper and with higher contrast. Hence, by dividing the input image into non-overlapping 3×3 blocks, the contrast value can be defined as [34]

$$C(u,v)=\frac{a(k)-b(k)}{a(k)+b(k)}, \tag{1}$$

where $C(u,v)$ is the contrast value for each pixel (u,v) in the k-th block, and $a(k)$ and $b(k)$ are the maximum and minimum gray values, respectively, in the k-th block. Larger $C(u,v)$ values represent scenes with higher contrast, i.e., an object within this block is closer to the viewer. **Fig. 4** shows an example of this method; the darker pixels represent the region distant from the viewer. Then a binary result is obtained by setting a threshold r and the number of black pixels $M(k)$ for the k-th block is calculated as follows:

$$M(k)=\sum_{(u,v)\in k}h(u,v)$$
$$where\ h(u,v)=\begin{cases}1, & if\ C(u,v)<\gamma\\ 0 & otherwise\end{cases}. \tag{2}$$

According to the defined hypothetical directions, the direction angle based on the atmospheric cue θ_a can be estimated from this result. As shown in **Fig. 4(d)**, regions A, B, and C have more black pixels, and thus, the hypothesis is inferred as being Type 3. To obtain more a precise depth gradient, the direction angles θ_p and θ_a estimated based on the perspective and atmospheric cues, respectively, are fused. However, three relationships between these values can be obtained:

A. The two estimated angles are consistent, $\theta_p=\theta_a$;

B. The difference between the two angles is smaller than 90°, i.e., $|\theta_p-\theta_a|\le 90°$;

C. The difference between the two angles is larger than 90°, i.e., the estimated results are contradictory.

Therefore, the final depth gradient is obtained as

$$\theta_d=\begin{cases}\theta_p & if\quad \theta_p=\theta_a\\ (\theta_p+\theta_a)/2 & if\quad |\theta_p-\theta_a|\le 90°\\ 90° & otherwise\end{cases} \tag{3}$$

4.2 Initial Depth Map Generation

Note that the depth gradient is estimated for a given image while ignoring image content, and this results in the possibility that one planar object can be assigned with various depth values. For example, in **Fig. 1**, although the depth structure of a person is not planar, for the viewer it makes sense that the person with a distance to camera would be represented with disparity values that correspond to a planar structure. Hence, to generate a consistent depth map, an input image is segmented by MST [1]. An image is represented by a graph containing vertices and edge links, where a vertex of the graph is composed of 4 by 4 pixels and each edge link between vertices is measured by the difference of mean values of neighboring blocks.

After obtaining the image depth gradient and segmented regions, the initial depth map can be generated by assigning the depth value for each segmented region (object) R by [1]:

$$G(R) = 128 + 255\left\{\sum_{I(x,y)\in R} \alpha \frac{x - W/2}{W} + \beta \frac{y - H/2}{H}\right\} / size(R) \tag{4}$$

where W and H are the image width and height, respectively, x and y are pixel coordinates, size(R) is the number of pixels in region R, α is the left-to-right weight, defined as $\cos(\theta_d)/c$, β is the bottom-to-top weight, defined as $\sin(\theta_d)/c$ and c is $\cos(\theta_d) + \sin(\theta_d)$. Note that each segmented object R is assigned the same depth value and the larger value of *Depth*(R) indicates that the object is closer to the viewer.

4.3 Salient Object Detection

To enhance stereoscopic effect, the image region serving as the viewer's primary focus must be evaluated. According to studies in neuroscience and psychology, high contrast and moving objects easily attract a viewer's focus. In other words, viewers will most readily focus on pixels with a color different from their surrounding pixels or those belonging to moving objects. This region or object of focus has been named the salient region or the salient object, respectively, in the computer vision field [24-25]. In this study, we briefly discuss the previously described algorithm [25] which we apply for salient object detection.

For a given input image, the image is first segmented into regions [60]. Then, for each region A, a color histogram is built in RGB color space. To reduce computational complexity, each color channel is quantized into 12 bins, and thus, the size of the histogram is 12^3. After obtaining the statistical information for each region, the salient value for region A_p can be defined as [25]

$$S(A_p) = \sum_{A_p \neq A_q} w(A_q) D(A_p, A_q), \tag{5}$$

where $w(A_q)$ is the number of pixels in region q, and the color distance $D(A_p, A_q)$ between A_p and A_q is calculated by, $D(A_p, A_q) = \sum_{i=1}^{n_p} \sum_{j=1}^{n_q} T(h_{p,i}) T(h_{q,j}) E(h_{p,i}, h_{q,j})$ where n_p and n_q are the respective number of bins, $T(h_{p,i})$ is the probability of obtaining color i in region p, $T(h_{q,j})$ is the probability of obtaining color j in region q, and $E(h_{p,i}, h_{q,j})$ is the Euclidean distance of color in $L^*a^*b^*$ color space. Note that the salient value represents contrast with the neighboring region.

Moreover, high contrast between neighboring regions will be more readily noticed for regions closer to the viewer. The spatial information is therefore incorporated and the salient value is reformulated as

(a) (b) (c)

Fig. 5. Example of salient object detection: (a) input image, (b) salient map, and (c) segmented object by GrabCut [60].

(a) (b) (c)

Fig. 6. Examples of (a) type 3 depth gradient, (b) initial depth map, and (c) refined depth map.

$$S(A_p) = \sum_{A_p \neq A_q} \exp(-\frac{D_S(A_p, A_q)}{\sigma_S^2}) w(A_q) D(A_p, A_q), \tag{6}$$

where $D_S(A_p, A_q)$ is the Euclidean distance between regions A_p and A_q, and σ_S controls the weight of the spatial information. **Fig. 5(b)** shows the salient map and **Fig. 5(c)** shows the detected salient object segmented by GrabCut [25], [60].

4.4 Depth Map Refinement for Stereoscopic Effect Enhancement

After obtaining the salient object, the next consecutive process is to refine the depth map. According to the binary result of salient object detection, the pixels of the initial depth map can be classified into non-salient and salient regions. If the pixel belongs to the non-salient region, its depth value (Eq. (4)) is not modified. On the other hand, if it belongs to the salient region, the depth value is modified and the consistency with other pixels within the same segmented object must be preserved. Thus, all the pixels within the salient region will be reset to the maximum depth value within the salient region. Therefore, the modified depth value $F(x,y)$ for each pixel can be formulated as

$$\begin{aligned} &F(x,y) = G_R(x,y)[1 - S(x,y)] + V(x,y)S(x,y), \\ &where \;\; V(x,y) = \max_{S(x,y)=1} G(x,y) \end{aligned} \tag{7}$$

and $G_R(x,y)$ is the assigned depth value for the pixel (x,y) in the corresponding region R (Eq. (4)), $S(x,y) \in \{0,1\}$ is the binary salient map from GrabCut, and $V(x,y)$ is the maximum depth gradient value assigned within the salient region. **Fig. 6** shows the refinement results based on **Fig. 5**. Object segmentation creates depth values that are consistent with the region containing the woman's body; however, differences between the results obtained from the salient map are observed for the face region. After modifying the depth values, the depth values within the woman's overall image are more consistent.

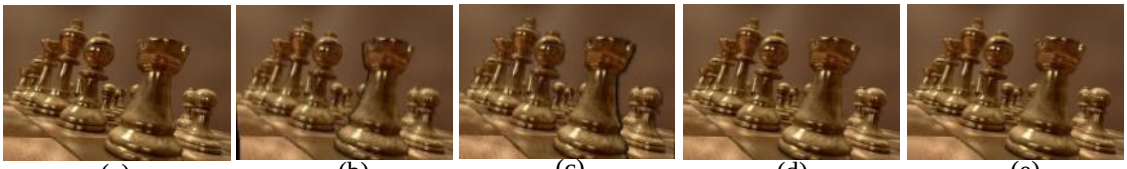
4.5 Stereoscopic Images Synthesis

Based on the refined depth map, DIBR [55] is applied to generate a left and right stereoscopic image pair, by

$$x_{left} = x_c + \frac{t_x}{2}\frac{f}{Z}$$
$$x_{right} = x_c - \frac{t_x}{2}\frac{f}{Z}, \tag{8}$$

where x_{left}, x_{right}, and x_c are the pixel coordinates in the left, right, and input image, respectively, t_x is the disparity between two eyes (generally 6.5 cm), f is the focus length, and Z is the depth value. Forward mapping is a method to synthesize images by setting the pixel value of x_{left} and x_{right}, respectively, and **Fig. 7** shows the results of forward mapping. It can be observed that large holes (black regions) surrounding the pixels near the frontal chesses. Note the presence of holes at the left areas of chesses in the left image (**Fig. 7(b)**) and at the right areas of chesses in the right image (**Fig. 7(c)**). This condition is referred to as the disocclusion problem [4]. In order to fill the holes, backward mapping with bilinear interpolation is applied. Each pixel coordinate in the left and right image is mapped back to the input image by Eq. (8) to get the reference pixel x'_{cl} and x'_{cr}, respectively. **Fig. 7(d)** and **Fig. 7(e)** shows the hole filling results.

$$x'_{cl} = x_{left} - \frac{t_x}{2}\frac{f}{Z}$$
$$x'_{cr} = x_{right} + \frac{t_x}{2}\frac{f}{Z}. \tag{9}$$



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

Fig. 7. (a) Input images; (b) and (c) are the left and right images synthesized by forward mapping, respectively; (d) and (e) are left and right images synthesized by backward mapping using bilinear interpolation, respectively.

4.6 Computational Complexity Analysis

The proposed system consists of four main steps: depth gradient estimation, salient object detection, depth map refinement, and stereoscopic image synthesis by DIBR. Depth gradient estimation consists of retrieving linear and atmospheric perspective cues. In addition, MST process is performed before generating the initial depth map. The complexity of line detection for the linear perspective cue is $O(n^2)$, where n is the number of edge points in the image. The complexity of retrieving an atmospheric cue is $O(m)$, where $m = W \times H$, W and H are the width and height of the image, respectively, and the complexity for MST is $O(e^{5E-5N})$ [2]. Here, N and E are the numbers of vertices and edges, respectively. In our study, N is $\frac{m}{16}$

and E is $\frac{2m-W-H}{4}$ when a 4×4 pixel block is used for MST. In addition, according to [25], the time complexity of salient object detection is $O(m\log m)$ if image segmentation [59] is applied to obtain a better saliency map. Contrarily, if the image segmentation is discarded, the time complexity reduces to $O(m)$. For the proposed method, the complexity of the last two steps, i.e., depth map refinement and stereoscopic image synthesis by DIBR, is $O(m)$. Overall, the computational complexity of the proposed method is $O(e^m + m\log m)$. In addition, we measure the running time on images with 400×300 pixels. On average, time costs for two main processes, i.e., depth gradient estimation and salient object detection, are 1.2s and 1.5s, respectively; when inputting a 2D image, the system takes 2.84s to obtain stereoscopic views. The experiments are performed on a PC with Intel CPU i7-4470 at 3.4 GHz with 4 cores and 16 GB memory.

Table 1. Six types of images with simple or complicated backgrounds and the type of presence of the main object.

		Background Complexity	
		Simple	Complicated
Presence of main object	Single main object		
	Multiple objects		
	No main object		

5. Experimental Results

In this section, the test database and evaluation mechanism is introduced. The results of depth gradient estimation by the proposed method are then analyzed, and the refined depth maps are compared with existing approaches. To demonstrate the improved stereoscopic effect in our study, the proposed method were compared with other synthesized image pairs generated using the depth map estimated by MST segmentation alone, denoted as *DG_MST*, and the binary depth map of the salient region, denoted as *DG_SALIENT*. In addition, a comparison of the results produced by commercial software and those by the state-of-the-art method is provided.

5.1 Dataset Collection

To analyze the stereoscopic effects of the proposed method, 90 test images were collected from the Internet consisting of images having two levels of background complexity and three

types of main object presence. Definitions of image types and corresponding examples are listed in **Table 1**, and the corresponding number of images for each type is listed in **Table 2**.

Table 2. Number of images for each image type listed in Table 1.

		Background Complexity	
		Simple	Complicated
Presence of main object	Single main object	28	16
	Multiple objects	19	7
	No main object	10	10

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

Fig. 8. Examples of linear perspective error: (a) and (c) are input images whereas (b) and (d) show the detected lines displayed in blue, which resulted in misclassification of the vanishing point.

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

Fig. 9. Examples of atmospheric perspective error: (a) and (c) are input images; (b) and (d) show the binary results of contrast values within each block.

5.2 Analysis of Depth Gradient Estimation

The linear perspective depends on the presence of lines in an image and the resulting estimation of the vanishing point. However, the absence of a line or inconsistent line detection will result in an incorrect location of the vanishing point and an inaccurate capturing angle. **Fig. 8** shows examples of such error results, where the detected lines are given in blue. The correct direction angle is 90°. However, because multiple, inconsistent lines were detected, the direction angles of **Fig. 8(a)** and **Fig. 8(b)** were estimated as 0° and 180°, respectively. The atmospheric scattering cue is based on the contrast values within each block and is used to estimate the depth gradient. However, when multiple objects exist in an image, the method will result in an incorrect depth gradient, as shown in **Fig. 9**. **Fig. 10** shows examples of depth maps of the collected images. These were generated by data transfer (DT) [17] using state-of-the-art data-driven 2D-to-3D conversion [11], MST, and the proposed method. It is seen that [17] can generate smoother depth maps than the other two methods, while MST and the proposed method provide more scene detail. More depth levels can be produced by MST and the proposed method. Note that the results obtained by [17] and MST were under- and over-segmented, respectively, and that the proposed method can produce a balance between these to provide consistent depth values within the object.

5.3 Stereoscopic Images Synthesis

For a given test image, the proposed work will synthesize left and right image pairs. The red–

cyan image will then be synthesized by [55]. To measure the stereoscopic effect, subjective evaluation was performed. The test interface is shown in **Fig. 11**. Seven generated results were randomly laid out on the interface, which consists of two results produced by commercial software (TriDef [15] and Media [16]) and five results produced by *DG_MST*, *DG_SALIENT*, Defocus map [61], DT [17], and the proposed method from their corresponding depth maps. Note that *DG_MST* used the depth gradient estimated by MST only and *DG_SALIENT* used binary depth map estimation of the salient region. Although study [61] did not aim at 2D-to-3D conversion, the generated defocus map can also provide depth values. Eleven volunteers then voted for the results that they deemed to offer a relatively better stereoscopic visual and comfortable perception. Then eleven volunteers who voted for the preferred results, which were deemed to offer the better stereoscopic visual and comfortable perception. There are 20 sampled test images with no salient content, e.g., general scenery. Hence, under these conditions, depth map refinement is discarded in the proposed method and the depth gradient is estimated only by MST. To classify whether salient content exits, the salient map is binarized by a threshold, as discussed previously. The ratio defined as the number of white pixels relative to the black ones is then used to classify the presence of salient content. Note that this ratio is set to 3.0 in our study. **Fig.12** shows images with no salient content. It can be observed that the size of the white region is larger than that of the black region, which indicates that no salient object is the subject of focus.

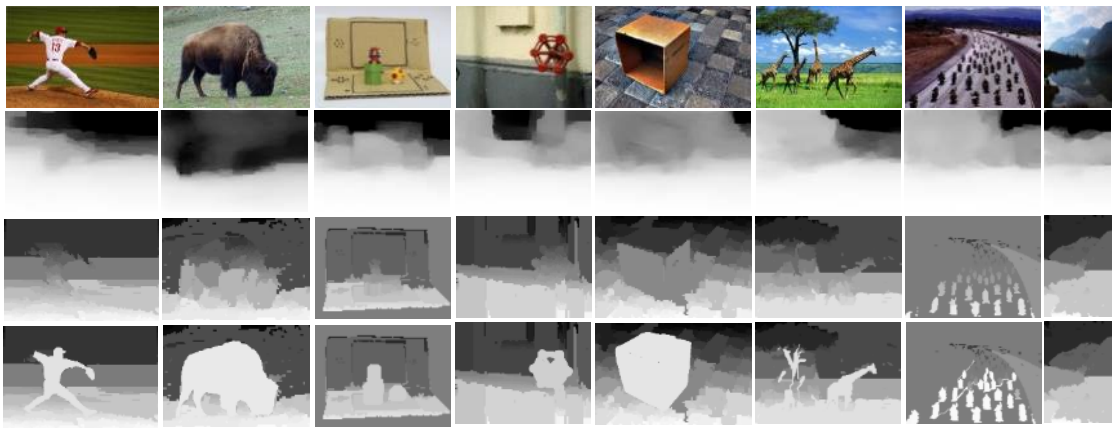


Fig. 10. Examples of depth maps. The first row shows the original image. Depth maps shown in the second, third, and fourth rows were generated by [17], MST, and the proposed method, respectively.

Fig. 11. Test Interface for subjective evaluation. Test interface for subjective evaluation. Red–cyan images generated by TriDef [15], Media [16], the proposed method, *DG_MST* (which produces a depth gradient estimate using MST segmentation only), *DG_SALIENT* (which uses a binary depth map based on the salient region detection results [25]) , Defocus map (which uses a defocus map [61]), and DT (which uses a depth map [17]) were randomly displayed. Each volunteer selected the image(s) of the seven rendered results that offered the best stereoscopic visual and comfortable perception.

Table 3. Summary of degrees of satisfaction from the subjective evaluation of the stereoscopic effects of 90 test images with various image contents produced by commercial software (TriDef [15] and Media [16]) and depth maps generated by MST (*DG_MST*), salient region (*DG_SALIENT*), Defocus map [61], DT [17], and the proposed method. Blue font indicates best performance, and green italic font indicates second best.

	TriDef	Media	*DG_MST*	*DG_SALIENT*	Defocus map	DT	Our method
Single main object + Simple background (Fig.12)	0.30	0.45	*0.51*	0.20	0.42	0.47	**0.57**
Multiple objects + Simple background (Fig. 13)	0.32	0.46	*0.50*	0.24	0.43	0.43	**0.53**
No main object + Simple background (Fig. 14)	0.25	0.41	0.39	0.37	**0.45**	*0.42*	0.39
Single main object + Complicated background (Fig. 15)	0.26	0.47	0.45	0.24	*0.48*	0.35	**0.60**
Multiple objects + Complicated background (Fig. 16)	0.26	0.40	**0.51**	0.34	0.46	*0.47*	0.42
No main object + Complicated background (Fig. 17)	0.36	0.59	**0.69**	0.60	*0.61*	0.55	0.70

Fig. 12. Examples of scenery images to the left with their corresponding salient maps in the middle, and their binarized salient maps to the right. The ratio defined as the number of white pixels relative to the black ones is used to classify the presence of salient content. In the right figures, the size of the white region is larger than that of the black region, which indicates that no salient object is detected.

Table 3 summarizes the degree of satisfaction as the subjective evaluation result. The analysis is performed for six categories, and each category comprises images with one level of background complexity and one type of main object present (**Table 1**). Note that for each test image, more than one result can be selected. Then, the degree of satisfaction is defined as the ratio of the number of votes to the number of test images in each category multiplied the number of volunteers. In the first category, i.e., single main object with a simple background, the proposed method produced the most satisfactory results, with degrees of satisfaction 6% and 10% higher than those achieved with *DG_MST* and DT, respectively. More than half of the volunteers were pleased with the results of the proposed method. In the fourth category, the degree of satisfaction with the proposed method was 12% and 13% higher than those achieved with Defocus map and Media, respectively. Thus, the proposed method can provide an impressive perception of images with one salient object even when the background is complex. When an image is composed of multiple objects with a simple background, the proposed method still provided the highest degree of satisfaction, but the improvement was not very obvious (3% greater than that achieved with *DG_MST*). In the fifth category, for images with

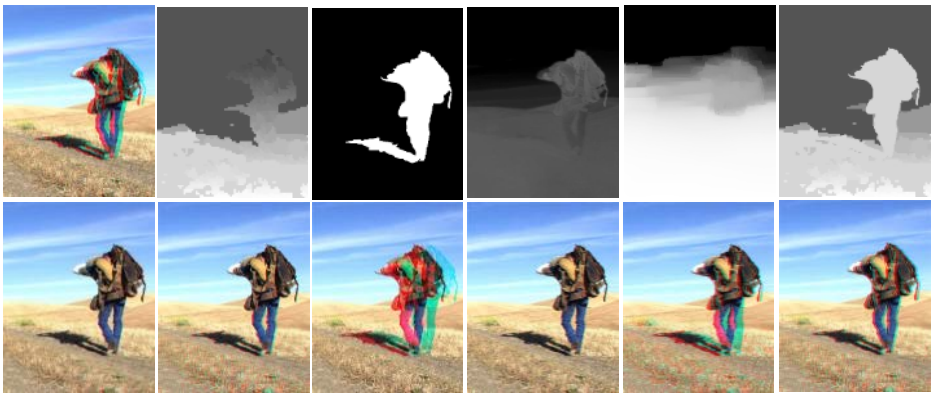


Fig. 13. Example of single main object with a simple background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, fourth, fifth, and sixth columns show the results of *DG_MST*, *DG_SALIENT*, Defocus map, DT, and the proposed method, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps.

Fig. 14. Example of multiple objects with a simple background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, fourth, fifth, and sixth columns show the results of *DG_MST*, *DG_SALIENT*, Defocus map, DT, and the proposed method, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps.

complicated backgrounds, *DG_MST* and DT provided relatively better results. In this case, the results might be attributable to the proposed method improperly combining the depth values of the salient object with those of the background region; depth maps that do not enhance the stereoscopic effect, such as *DG_MST*, DT, or Defocus map, can provide better results in this case. In the sixth category, *DG_MST* provided relatively better results than Defocus map and DT; it is possible that when there is no main object in a scene, the stereoscopic effect can be perceived through the level of depth change, and thus, more detail in the depth map leads to a higher degree of satisfaction. For the six categories, the average degrees of stratification produced by TriDef, Media, DG_MST, *DG_SALIENT*, Defocus map, DT, and the proposed method were 0.29, 0.46, 0.51, 0.33, 0.48, 0.45, and 0.54, respectively. Overall, the commercial software TriDef produced the worst results. It is also observed from **Table 3** that the results of *DG_SALIENT* were not satisfactory because the method tends to overemphasize the salient object and neglect its positional relation relative to the background, resulting in an incorrect depth gradient. Although the stereoscopic perception of the salient object is enhanced, the method still cannot provide satisfactory results. On the other hand, the salient object content is not emphasized by *DG_MST*, and the method can give better results than *DG_SALIENT* because of the more accurate depth gradient. From a comparison of the above results produced by the proposed method with those of other methods, it is apparent that our framework can provide more satisfactory results when a main salient object is present because the proposed framework estimates depth gradients more precisely when compared with *DG_SALIENCY*

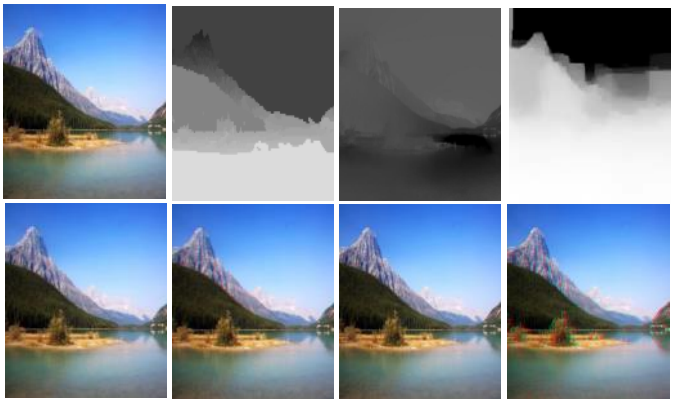


Fig. 15. Example of no salient object with a simple background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, and fourth columns show the results of *DG_MST*, Defocus map, and DT, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps. Note that results of the proposed method and *DG_SALIENT* are equivalent to those of *DG_MST*.

and DT and emphasizes the stereoscopic effect surrounding the salient object region better than Defocus map. Note, as mentioned above, if no salient object is detected in the test image, the process of depth map refinement will be discarded and the synthesized results will be the same as *DG_MST*. However, it can be observed that the results of test images containing no main object with a simple background and no main object with a complicated background for *DG_MST*, *DG_SALIENT*, and the proposed study are not consistent (**Table 2**). These results were obtainable because some reviewers voted for a single result. More synthesized results of various image content are shown in **Fig. 13**, **Fig. 14**, **Fig. 15**, **Fig. 16**, **Fig. 17**, and **Fig. 18** (it is suggested that the reader wear red–cyan glasses before evaluating the results for better visual effect). Furthermore, TriDef and Media required 0.3s and 1.2 s, respectively, to generate synthesis image pairs for one image with 400×300 pixels. For the same image, Defocus map [61] and DT [17] required 64.5s and 76.7s, respectively, for depth map generation in MATLAB.

6. Conclusion

We proposed an image-based 2D-to-3D conversion system that provides significant stereoscopic visual effects for humans. To avoid situation where one depth cue is not present in the given image, two depth cues, linear perspective and atmospheric cues, are fused to estimate the depth information. Rather than a retrieving precise depth value for pixels from the depth cues, the observation direction angle of the image is estimated. Then the depth gradient in accordance with the direction angle is integrated with superpixels to obtain the initial depth map. To enhance the visual stereoscopic effect, the visual attention of humans is considered. Saliency object detection within the image is performed to explore a significance region and then the refinement of the depth map is conducted by enhancing the contrast depth value around the salient regions. Note that the depth consistency of all other regions is preserved. Thus, the refined depth map can not only maintain global depth consistency by correcting non-uniform depth values but visual stereoscopic effect is enhanced as well. According to our subjective evaluation study, our method produces the most satisfactory results; the degrees of satisfaction were 8% and 7% greater than those achieved using commercial 2D-to-3D conversion software and the state-of-the-art approach, respectively.

Fig. 16. Example of single main object with a complicated background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, fourth, fifth, and sixth columns show the results of *DG_MST*, *DG_SALIENT*, Defocus map, DT, and the proposed method, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps.

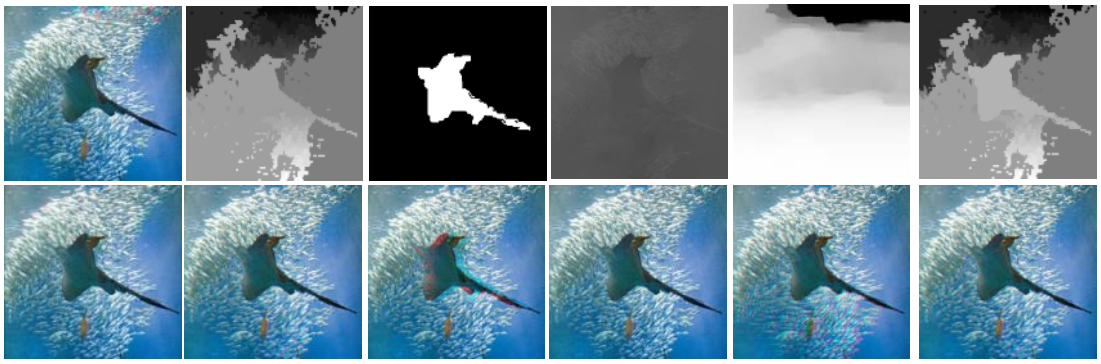


Fig. 17. Example of multiple objects with a complicated background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, fourth, fifth, and sixth columns show the results of *DG_MST*, *DG_SALIENT*, Defocus map, DT, and the proposed method, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps.

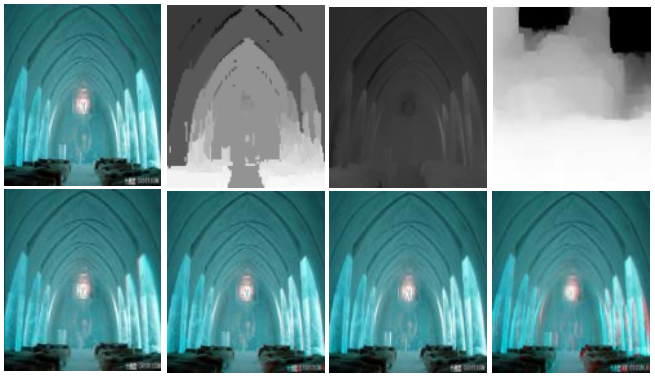


Fig. 18. Example of no salient object with a complicated background. Top and bottom images in the first column are synthesized views produced by TriDef and Media, respectively. The second, third, and fourth columns show the results of *DG_MST*, Defocus map, and DT, respectively. Depth maps are shown in the first row, and synthesized views are shown below their respective depth maps. Note that results of the proposed method and *DG_SALIENT* are equivalent to those of *DG_MST*.

Acknowledgment

This work is supported by National Science Council (NSC), Taiwan, under Contract of MOST 104-2221-E-151-028. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] S.F. Tsai, C.C. Cheng, C.T. Li, and L.G. Chen, "A Real-Time 1080p 2D-to-3D Video Conversion System," *IEEE Transactions on Consumer Electronics*, Vol. 57, No. 2, pp. 915-922, 2011. Article (CrossRef Link)

[2] C.C. Cheng, C.T. Li, and L.G. Chen, "A Novel 2D-to-3D Conversion System Using Edge Information," *IEEE Transactions on Consumer Electronics*, Vol. 56, No. 3, pp. 1739-1745, 2010. Article (CrossRef Link)

[3] S.B. Gokturk, H.Yalcin, and C. Bamji, "A Time-of-Flight Depth Sensor, System Description, Issues and Solutions," *IEEE Computer Vision and Pattern Recognition Workshop*, 2004. Article (CrossRef Link)

[4] L.M. Po, X. Xu, Y. Zhu, S. Zhang, and K.W. Cheung, "Automatic 2D-to-3D Video Conversion Technique based on Depth-from-Motion and Color Segmentation," *IEEE Conference on Signal Processing*, pp. 1000-1003, 2010. Article (CrossRef Link)

[5] X. Cao, Z. Li, and Q Dai, "Semi-Automatic 2D-to-3D Conversion Using Disparity Propagation," *IEEE Transactions on Broadcasting*, Vol. 57, No. 2, pp. 491-499, 2011. Article (CrossRef Link)

[6] R. Rzeszutek, R. Phan, and D. Androutsos, "Depth Estimation for Semi-automatic 2D to 3D Conversion," *ACM Conference on Multimedia*, pp. 817-820, 2012. Article (CrossRef Link)

[7] M. Guttmann, L. Wolf, and D. Cohen-or, "Semi-automatic Stereo Extraction from Video Footage," *IEEE International Conference on Computer Vision*, pp. 136-142, 2009. Article (CrossRef Link)

[8] Z. Zhang, C. Zhou, B. Xin, Y. Wang, and W. Gao, "An Interactive System of Stereoscopic Video Conversion," *ACM Conference on Multimedia*, pp.149-158, 2012. Article (CrossRef Link)

[9] L. Zhang, C. Vazquez, and S. Knorr, "3D-TV Content Creation: Automatic 2D-to-3D Video Conversion," *IEEE Transactions on Broadcasting*, Vol. 57, No. 2, pp. 372-383, 2011. Article (CrossRef Link)

[10] F. Guo, J. Tang, and H. Peng, "Automatic 2D-to-3D Image Conversion Based on Depth Map Estimation," *International Journal of Signal Process, Image Processing, and Pattern Recognition*, Vol. 8, No. 4, pp. 99-112, 2015. Article (CrossRef Link)

[11] K. Calagari, M. Elgharib, P. Didyk, A. Kaspar, W. Matusik, and M. Hefeeda, "Gradient-Based 2D-to-3D Conversion for Soccer Videos," in *Proc. of ACM Conference on Multimedia*, pp. 331-340, 2015. Article (CrossRef Link)

[12] W.J. Tam, C. Vazquez, and F. Speranza, "Three-dimensional TV: A Novel Method for Generating Surrogate Depth Maps using Colour Information," *SPIE 7237, Stereoscopic Displays and Applications*, 2009. Article (CrossRef Link)

[13] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-Based Automatic 2D-to-3D Image and Video Conversion," *IEEE Transactions on Image Processing*, Vol. 22, No. 9, pp. 3485-3496, 2013. Article (CrossRef Link)

[14] W.J. Tam and L. Zhang, "3D-TV Content Generation: 2D-to-3D Conversion," in *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 1869-1872, 2006. Article (CrossRef Link)

[15] DDD's TriDef 3D. Article (CrossRef Link)

[16] ArcSoft's Media Converter. Article (CrossRef Link)

[17] K. Karsch, C. Liu, and S. B. Kang, "Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 11, pp. 2144-2158, 2014. Article (CrossRef Link)

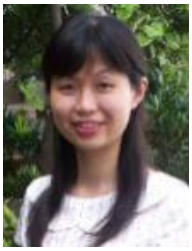
[18] A. Saxena, S.H. Chung, A.Y. Ng, "3-D Depth Reconstruction from a Single Still Image," *International Journal of Computer Vision*, Vol. 76, No. 1, pp. 53-69, 2007. Article (CrossRef Link)

[19] A. Saxena, M. Sun, A.Y. Ng, "Make3D: Learning 3-D Scene Structure from a Single Still Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 5, pp. 824-840, 2008. Article (CrossRef Link)

[20] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction From a Single Image Using a Multi-Scale Deep Network," *Advances in Neural Information Processing Systems*, pp. 2366-2374, 2014. Article (CrossRef Link)
[21] F. Liu, C. Shen, and G. Lin, and I. Rei, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PP, No. 99, 2015.
[22] D. Kim, D. Min, and K. Sohn, "A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis," *IEEE Transactions on Broadcasting*, Vol. 54, No 2, pp. 188-197, 2008. Article (CrossRef Link)
[23] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo Assessment based on Computational Visual Attention Model," *ACM International Conference on Multimedia*, pp. 541-544, 2009. Article (CrossRef Link)
[24] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007. Article (CrossRef Link)
[25] M.M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.M. Hu, "Global Contrast based Salient Region Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. Article (CrossRef Link)
[26] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li, "Actively Learning Human Gaze Shifting Paths for Semantics-Aware Photo Cropping," *IEEE Transactions on Image Processing*, Vol. 23, No. 5, pp. 2235-2245, 2014. Article (CrossRef Link)
[27] Y.L. Chang, C.Y. Fang, L.F. Ding, S.Y. Chen, and L.G. Chen, "Depth Map Generation for 2D-to-3D conversion by Short-Term Motion Assisted Color Segmentation," in *Proc. Of International Conference on Multimedia and Expo*, pp. 1958-1961, 2007. Article (CrossRef Link)
[28] J. Ko, M. Kim, and C. Kim, "2D-to-3D Stereoscopic Conversion: Depth-Map Estimation in a 2D Single-View Image," *SPIE Conference on Applications of Digital Image*, 2007. Article (CrossRef Link)
[29] K. Han and K. Hong, "Geometric and Texture Cue Based Depth-map Estimation for 2D to 3D Image Conversion," in *Proc. of IEEE International Conference on Consumer Electronics*, pp. 651-652, 2011. Article (CrossRef Link)
[30] J.I. Jung and Y.S. Ho, "Depth Map Estimation from Single-View Image using Object Classification based on Bayesian Learning," *IEEE 3DTV-Conference on Transmission and Display of 3D Video*, pp. 1-4, 2010. Article (CrossRef Link)
[31] R. Phan, R. Rzeszutek, D. Androutsos, "Semi-automatic 2D to 3D Image Conversion Using a Hybrid Random Walks and Graph Cuts based Approach," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 897-900, 2011. Article (CrossRef Link)
[32] L. Grady, "Random Walks for Image Segmentation," *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 28, No. 11, pp. 1768-1783, 2006. Article (CrossRef Link)
[33] Y. Boykov and G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," *International Journal of Computer Vision*, Vol. 2, No. 70, pp. 109-131, 2006. Article (CrossRef Link)
[34] G.S. Lin, C.Y. Yeh, W.C. Chen, and W.N. Lie, "A 2D to 3D Conversion Scheme Based On Depth Cues Analysis For MPEG Videos," *International Conference on Multimedia and Expo*, pp.1141-1145, 2010. Article (CrossRef Link)
[35] J. Kim, Y. Choe, and Y. G. Kim, "Robust MRF-based Object Tracking and Graph-cut-based Contour Refinement for High Quality 2D to 3D Video Conversion," *IEEE Pacific Rim Conference on Communications*, pp. 358-363, 2011. Article (CrossRef Link)
[36] J. Lee, S. Yoo, C. Kim, and B. Vasudev, "Estimating Scene-Oriented Pseudo Depth with Pictorial Depth Cues," *IEEE Transactions on Broadcasting*, Vol. 59, No. 2, pp. 238-250, 2013. Article (CrossRef Link)
[37] Y.M. Tsai, Y.L. Chang, and L.G. Chen, "Block-based Vanishing Line and Vanishing Point Detection for 3D Scene Reconstruction," in *Proc. of International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 586-589, 2006. Article (CrossRef Link)
[38] C.L. Su, K.N. Pang, T.M. Chen, G.S. Wu, C.L. Chiang, H.R. Wen, L.S. Huang, Y.H. Hsueh, and S.Y. Tseng, "A Real-time Full-HD 2D-to-3D Conversion System Using Multicore Technology," in *Proc. of International Conference on Multimedia and Ubiquitous Engineering*, pp. 273-276, 2011. Article (CrossRef Link)
[39] Y.J. Jung, A. Baik, J. Kim, and D. Park, "A Novel 2D-to-3D Conversion Technique based on Relative Height Depth Cue," in *Proc. of SPIE Conference on Stereoscopic Displays and Applications*, Vol. 7234, 2009. Article (CrossRef Link)

[40] A.E. Welchman, A. Deubelius, V. Conrad, H.H. Bülthoff, and Z. Kourtzi, "3D Shape Perception from Combined Depth Cues in Human Visual Cortex," *Nature Neuroscience*, Vol. 8, pp. 820-827, 2005. Article (CrossRef Link)
[41] Y. Lu, J. Zhang, Q. Wu, and Z. Li, "A Survey of Motion-parallax based 3-d Reconstruction Algorithms," *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 34, pp. 532-548, 2004. Article (CrossRef Link)
[42] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining Motion and Lighting for 3d Surface Reconstruction," *International Journal of Computer Vision*, pp. 75-90, 2002. Article (CrossRef Link)
[43] J. Malik and R. Rosenholtz, "Computing Local Surface Orientation and Shape from Texture for Curved Surfaces," *International Journal of Computer Vision*, Vol. 23, No. 2, pp. 149-168, 1997 Article (CrossRef Link)
[44] A. Torralba and A. Oliva, "Depth Estimation from Image Structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol .24, No. 9, pp. 1-13, 2002. Article (CrossRef Link)
[45] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, pp. 309-314, 2012. Article (CrossRef Link)
[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014. Article (CrossRef Link)
[47] H. Tian, B. Zhuang, Y. Hua, and A. Cai, "Depth Inference with Convolutional Neural Network," in *Proc. of IEEE Conference on Visual Communications and Image Processing*, pp. 169-172, 2014. Article (CrossRef Link)
[48] F. Liu, C. Shen, and G. Lin, "Deep Convolutional Neural Fields for Depth Estimation from a Single Image," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162-5170, 2015. Article (CrossRef Link)
[49] M. Liu, M. Salzmann, and X. He, "Discrete-Continuous Depth Estimation from a Single Image," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716-723, 2014. Article (CrossRef Link)
[50] M. Li, H. Chen, R. Li, and X. Chang, "An Improved Virtual View Rendering Method Based on Depth Image," in *Proc. of IEEE International Conference on Communication Technology*, pp. 381-384, 2011. Article (CrossRef Link)
[51] S. Knorr and T. Sikora, "An Image-Based Rendering (IBR) Approach for Realistic Stereo View Synthesis of TV Broadcast Based on Structure From Motion," in *Proc. of IEEE International Conference on Image Processing*, Vol. 6, pp. 572-575, 2007. Article (CrossRef Link)
[52] M. Levoy and P. Hanrahan, "Light Field Rendering," in *Proc. of ACM conference on Computer graphics and interactive techniques*, pp. 31-42, 1996. Article (CrossRef Link)
[53] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," in *Proc. of ACM conference on Computer graphics and interactive techniques* , pp. 43-54, 1996. Article (CrossRef Link)
[54] S.J. Luo, Y.T. Sun, I.C. Shen, and B.Y. Chen, "Geometrically Consistent Stereoscopic Image Editing Using Patch-Based Synthesis," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 21, No. 1, pp.56-67, 2014. Article (CrossRef Link)
[55] C. Fehn, "Depth-image-based rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," in *Proc. of SPIE Conference on Stereoscopic Image Processing and Renderin*, Vol. 5291, 2004. Article (CrossRef Link)
[56] L.H. Wang, X.J. Huang, M. Xi, D.X. Li, and M. Zhang, "An Asymmetric Edge Adaptive Filter for Depth Generation and Hole Filling in 3DTV," *IEEE Transactions on Broadcasting*, Vol. 56, No. 3, pp. 425-431, 2010. Article (CrossRef Link)
[57] R. Liu, H. Xie, F. Tian, Y. Wu, G. Tai, Y. Tan, W. Tan, B. Li, H. Chen, and L. Ge, "Hole-filling based on Disparity Map for DIBR," *KSII Transactions on Internet & Information Systems*, Vol. 6, No. 10, 2012. Article (CrossRef Link)
[58] C. Yao, T. Tillo, Y. Zhao, J. Xiao, H. Bai, and C. Lin, "Depth Map Driven Hole Filling Algorithm Exploring Temporal Correction Information," *IEEE Transactions on Broadcasting*, Vol. 60, No2, pp. 394-404, 2014. Article (CrossRef Link)
[59] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, Vol. 59, No. 2, pp. 167-181, 2004. Article (CrossRef Link)

[60] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM Transactions on Graphics*, Vol. 23, No. 3, pp. 309-314, 2004. Article (CrossRef Link)
[61] S. Zhuo and T. Sim, "Defoucs Map Estimation from a Single Image," *Pattern Recognition*, Vol. 44, No. 9, pp. 1852-1858, 2011. Article (CrossRef Link)



Ju-Chin Chen received her B.S., M.S. and Ph.D. degrees in computer science and information engineering from the National Cheng Kung University, Tainan, Taiwan, in 2002, 2004 and 2010, respectively. She is now an assistant professor in the Department of Computer Science and Information Engineering at the National Kaohsiung University of Applied Sciences, Taiwan. Her research interests lie in the fields of machine learning, pattern recognition, and image processing.

Meng-yuan Huang received his B.S. degree in computer science and information engineering from I-Shou University, Kaohsiung, Taiwan, in 2010. He received his M.S. degree in computer science and information engineering from the National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan, in 2012. His research interests lie in the fields of computer vision and 3d model reconstruction.