Brain-Inspired Framework for Fusion of Multiple Depth Cues

Chung-Te Li, Student Member, IEEE, Yen-Chieh Lai, Chien Wu, Student Member, IEEE, Sung-Fang Tsai, Student Member, IEEE, Tung-Chien Chen, Shao-Yi Chien Member, IEEE, and Liang-Gee Chen*, Fellow, IEEE, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, R.O.C

***Abstract*—2D-to-3D conversion is an important step for obtaining 3D videos, where a variety of monocular depth cues have been explored to generate 3D videos from 2D videos. As in a human brain, a fusion of these monocular depth cues can re-generate 3D data from 2D data. By mimicking how our brains generate depth perception, we propose a reliability-based fusion of multiple depth cues for an automatic 2D-to-3D video conversion. A series of comparisons between the proposed framework and the previous methods is also presented. It shows that significant improvement is achieved in both subjective and objective experimental results. From the subjective viewpoint, the brain-inspired framework outperforms earlier conversion methods by preserving more reliable depth cues. Moreover, an enhancement of 0.70-3.14 dB and 0.0059-0.1517 in the perceptual quality of the videos is realized in terms of the objective-modified peak signal-to-noise ratio and disparity distortion model, respectively.**

***Index Terms*—2D-to-3D conversion, brain-inspired fusion, depth generation, multiple depth cues**

I. Introduction

TECHNOLOGICAL revolutions in 3DTVs and displays have recently reshaped the way people live. 3D video processing has also become a trend in the field of video technology. However, a fundamental problem, lack of content, still exists. Therefore, the development of 2D-to-3D video conversion, which can convert all existing 2D videos into 3D video content, is urgently required. 2D-to-3D video conversion has therefore attracted many researchers to the field of video technology [1]–[12]. Fig. 1 shows the general flow of 2D-to-3D conversion. The depth information is estimated from a monocular video. Next, depth-image-based rendering (DIBR) is applied to generate stereoscopic views, which provide a 3D perception to viewers. In this system, depth information is a key component. However, the time-consuming semi-automatic generation of depth information [1]–[4] is a barrier to the mass-market promotion of depth information although these semi-automatic methods can provide high-quality depth information. Therefore, a cost-effective 2D-to-3D conversion system, which can automatically estimates the depth information from a monoscopic video, is demanded.

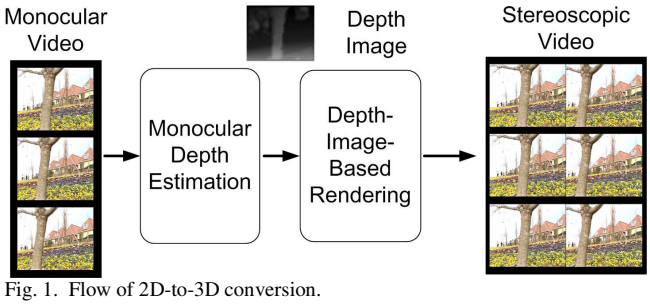

Fig. 1. Flow of 2D-to-3D conversion.

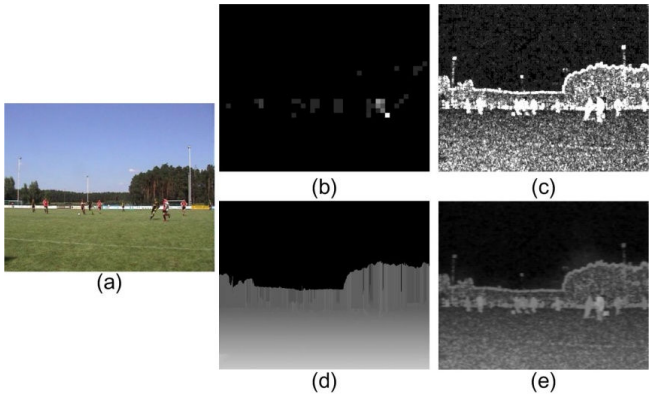

Fig. 2. Example of (a) sequence soccer, (b) depth from motion [5], (c) depth from edge [8], (d) depth from height in the visual field [19], and (e) integrated result [13] (linear combination of results obtained from [5], [8], and [19]).

In the recent years, several algorithms that automatically generate depth information for a 2D-to-3D conversion system have been developed to create more 3D content on the basis of a pre-selected depth cue. Examples of such approaches include depth from motion [5], defocus or focus [6], [7], edge [8], color [9], [10], and occlusion [11], [12]. The above-mentioned methods generate 3D images or videos with only one main depth cue. However, for complex scenes or regions, the pre-selected depth cue may not be sufficiently effective in certain specific cases.

Manuscript received April 1, 2012. This work was supported in part by the National Science Council, Taiwan, R.O.C. under Grant NSC-100-2221-E-002-248 and Himax Technologies, Inc. under Grant 101-S-C37. This paper was recommended by Associate Editor S. Battiato.

Chung-Te Li, Chien Wu, Sung-Fang Tsai, Tung-Chieh Chen, Shao-Yi Chien, and Liang-Gee Chen* are with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, R.O.C (corresponding author to provide phone: 886-2-3366-9739; fax: 886-2-3366-3718; e-mail: lgchen@ video.ee.ntu.edu.tw).

Yen-Chieh Lai was with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. He is now with the Mediatek Incorporation, Hsinchu, Taiwan, R.O.C

These automatic algorithms based on a single depth cue sometimes provide unreliable depth estimations for some complex regions, as shown in Figs. 2 (b), (c), and (d). Therefore, other depth cues need to be considered for an accurate depth generation. As shown in Fig. 2(b), in the case of depth from motion, only parts of the players on the soccer field appear to protrude. In the case of depth from edge, there are a large number of noises in the sky and ground areas, as shown in Fig. 2(c). In the case of depth from the height in the visual field, the players appear to vanish into the background, as shown in Fig. 2(d). The lack of some depth cues leads to a deficient depth estimation. Therefore, a fusion of the depth cues is required, particularly for complex scenes. The fusion of multiple depth cues can be traced back to the early 2000s. Battiaio et al. [13] [14] used a depth fusion method integrating the still image classification and geometry perspective to produce a smoother and more meaningful depth map. Several engineering works attempt to fuse depth cues [15]–[17] with simple linear combination methods. However, a simple heuristic combination alone is not sufficient for obtaining a high-quality depth map. A simple example is shown in Fig. 2(e). Although most of the desired depth cues are fused, the noises (i.e., uncertain depth cues) are also included in the depth estimation.

However, even for complex scenes, the human brain can easily evaluate the depth information with less uncertainty. Psychologists [18], [19] have found that the human brain manages this variability on the basis of the reliability of the depth cues. The reliability of a depth cue is locally checked, and the depth cues are fused appropriately by the human brain. In this paper, we propose a human-brain-inspired framework for the fusion of the estimated depth cues by analyzing the reliability of the depth cues locally (i.e., pixel by pixel). Similar to the human brain, the proposed method can generate reliable depth information even for complex scenes.

The rest of this paper is organized as follows: Section II addresses how the human brain analyzes depth to generate depth perception and the challenges in mimicking the brain. Section III describes the proposed system, which mimics the manner in which the brain analyzes the depth information. The simulation results and comparisons with other methods are then presented in Section IV. Finally, we present our conclusions in Section V.

II. Preliminary Knowledge and Challenges—Depth Perception in Human Brain

Fig. 3 shows how the depth perception in the human brain is generated. Psychological researches show that the brain extracts the depth information from a variety of cues [18], [19]. First, each depth cue is estimated individually. Many depth cues exist, such as occlusion, accommodation, binocular disparity, and convergence. Depth perception can be generated by these depth cues. For example, occlusion by other objects and the accommodation of eyes are the cues that provide the information about the relative distance and the distance to the focusing plane, respectively. The human brain attempts to find where occlusion occurs and how the eyes accommodate and then gets the cues for estimating the depth. Second, since the

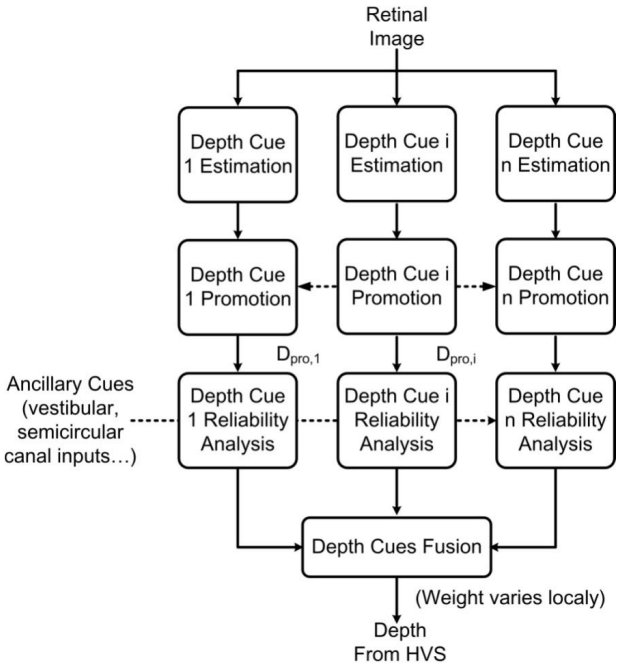

Fig. 3. Depth perception in the human brain [17].

estimated depth from each cue individually is usually not reliable because the context of the environment can vary widely, the considered depth cues are promoted by the other depth cues. The promotion can suppress the uncertainties in the depth cues because the human brain analyzes the context more accurately and then adjusts the model of each depth cue to fit the context better. Next, the reliability of the depth cues is estimated using ancillary cues, for example, vestibular inputs. Notably, the reliability plays a significant role for the final fusion of the depth cues in the following stage. A dynamically weighted fusion is finally applied to the depth cues in order to generate the depth perception in the human brain. A higher reliability implies higher weighting. In summary, the human brain estimates the depth accurately on the basis of promotion and reliability-based weighted fusion of the depth cues. The promotion suppresses uncertainties, and the reliability-based weighted fusion enhances robust depth cues and suppresses non-robust ones.

However, there are still many challenges in mimicking the manner in which the brain reconstructs depth perception. First, the mechanism of promotion is not well defined for general scenes while only the purpose of the promotion is known well, i.e., noise suppression for each depth cue.

Secondly, ancillary cues such as vestibular inputs do not exist for 2D-to-3D video conversion. The reliability, which is important for the fusion of depth cues, is difficult to estimate; hence, the weighting for a fusion of depth cues cannot be defined. In this paper, probability distributions of depth cues are employed to meet the abovementioned challenges. We propose probabilistic noise suppression, reliability analysis, and reliability-based weighted fusion for the depth cues to mimic the manner in which the human brain generates depth information by probability distributions.

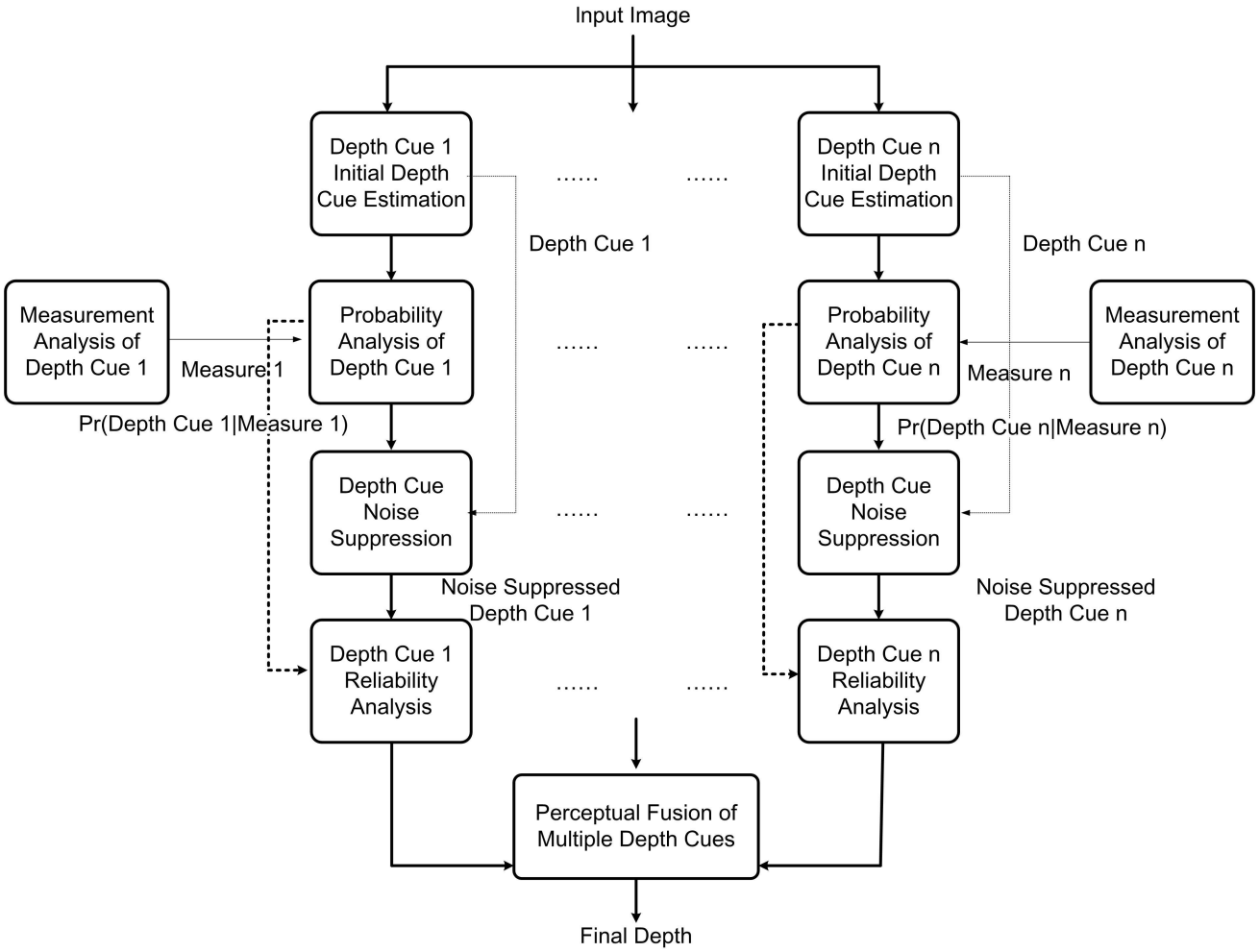

Fig. 4. Schematic representation of the proposed system for depth cue estimation, promotion, and combination..

III. Proposed Perceptual Fusion of Multiple Depth Cues

We attempt to mimic the depth generation process of the human brain in this paper. Depth cue estimation, noise suppression, and fusion are applied in sequence, as shown in Fig. 4. First, the greatest possible value of each depth cue is estimated individually in Stage A. Second, in Stage B, we study the measurement of each depth cue, such as the magnitude of the motion of depth from motion. In addition, the conditional probability for each depth cue given its measurement is modeled. In Stage C, instead of the promotion in the human brain, the noises in each depth cue are suppressed on the basis of the conditional probability for a depth cue given its measurement. In Stage D, the reliability of the depth cues is estimated on the basis of the above-mentioned conditional probabilities and the consistency of the estimated values of the depth cues for similar pixels, instead of vestibular inputs. Finally, in Stage E, the final depth is generated from a dynamically weighted fusion of multiple depth cues in the same manner as in the human brain.

TABLE I
Ranking of the Monocular Depth Cues on the Basis of the Relative Importance [20]

Monocular Depth Cues	Personal Space	Action Space
Occlusion	**1**	**1**
Relative size	**3**	**3.5**
Relative density	5	5
Height in visual field	N.A.	**2**
Aerial perspective	6	6
Motion	**2**	**3.5**
Accommodation	**4**	7

A. Initial Estimation for Each Depth Cue

Motion, accommodation, and height in the visual field are selected as the referenced depth cues in this work. The reason for this is discussed in the following paragraphs. In the human brain, 3D perception is generated from a variety of monocular and binocular depth cues, including occlusion, relative size, relative density, height in visual field, aerial perspective, motion, accommodation, binocular disparity, and convergence. To simplify the brain-mimicking process for depth generation, we attempt to find important depth cues. Cutting et al. [20] provided a deep analysis for the relative importance of the depth cues. Notably, the importance depends on the distance from the observer. Table I (entry [20] onwards) indicates the ranking of the monocular depth cues on the basis of their

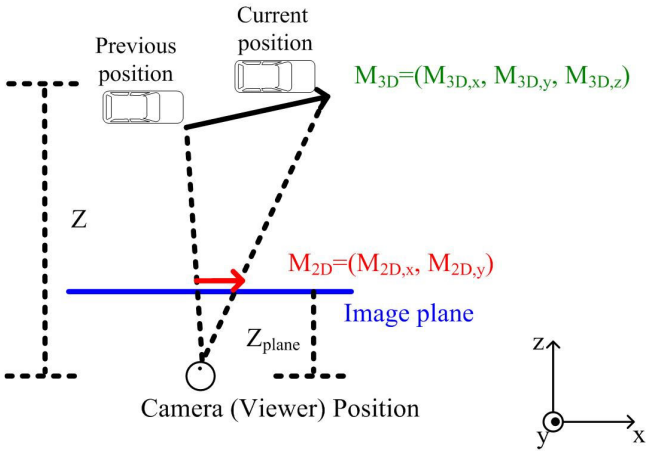

Fig. 5. Illustration of the relationships among actual motion in 3D space (i.e., a three-dimensional vector map, denoted as M_{3D} = ($M_{3D,x}$, $M_{3D,y}$, $M_{3D,z}$)), projected 2D motion vectors (i.e., a two-dimensional vector map, denoted as M_{2D} = ($M_{2D,x}$, $M_{2D,y}$)), and the distances from the given object to the camera (denoted as Z). (Z_{plane} denotes the distance from the image plane to the camera.)

relative importance in the human brain. Smaller ranking values correspond to higher relative importance. The ranking is mainly based on the just noticeable difference (abbreviated as JND) of each depth cue in the personal space and action space, respectively. The JND of each depth cue is defined as the smallest detectable difference of depth cue between a starting and secondary depth level in their experiments. Larger JND implies less relative importance since a depth cue with larger JND means that it is harder to be detected in the human vision system. Two spaces are defined in Table I. Personal space is defined as the zone immediately surrounding the observer's head, generally within arm's reach and slightly beyond, and action space is defined as a space of an individual's public action that lies immediately beyond the personal space. We only apply the important monocular depth cues in the proposed framework. Hence, aerial perspective and relative density are excluded because of their lower importance in both the spaces. In addition, occlusion and relative size are not considered since they depend on the concept of object perception, which is still an unsolved problem for generic images or videos.

In the proposed framework, the initial depth cue estimations of the referenced depth cues are first employed. We define the values of the depth cues as 8-bit numbers in this paper. For each depth cue, a large value implies a small distance to the camera or the eyes of the viewer. We estimate the greatest possible values of the depth cues on the basis of the conventional algorithms in [5], [8], and [21] for motion, accommodation, and height in the visual field, respectively. Notably, the depth cues are random variables instead of fixed values in this paper.

In the case of depth cue from motion, the Euclidean norms of the motion vectors, which are widely used for video compression, are applied to determinate the depth cue in [5]. For estimating the depth cue, we attempt to explore the relationships among actual motion in the 3D space (i.e., a three-dimensional vector map, denoted as M_{3D} = ($M_{3D,x}$, $M_{3D,y}$, $M_{3D,z}$), projected 2D motion vectors (i.e., a two-dimensional vector map, which is widely used in video compression, denoted as M_{2D} = ($M_{2D,x}$, $M_{2D,y}$), and the distances from the given object to the camera (denoted as Z) as shown in Fig. 5. M_{2D} can be computed from M_{3D} and Z when $M_{3D,z}$ is relatively small, i.e.,

$$M_{2D,x} = M_{3D,x} / (Z / Z_{plane}), \tag{1}$$

$$M_{2D,y} = M_{3D,y} / (Z / Z_{plane}), \tag{2}$$

where Z_{plane} means the distance from the image plane to the camera. Consequently, the projected motion vector becomes small when the distance Z becomes relatively large. With this observation, the greatest possible value of the depth cue from motion D_M can be estimated by the Euclidean norm of M_{2D} as in [5], i.e.,

$$\arg\max_{D_M} \Pr(D_M) = \frac{1}{Z} = \alpha_M \sqrt{M_{2D,x}{}^2 + M_{2D,y}{}^2}, \tag{3}$$

where α_M is a scaling factor from the magnitude of the motion to the estimated depth cue.

In the case of depth cue from accommodation, the amount of defocus blur is generally assumed to be proportional to the distance from the plane of focus. Namely, the boundaries or the texture regions of an object become blurred when the object is moved away from the plane of focus. To estimate the blurriness, a Sobel edge filter is applied as in [8]. For pixels belonging to the boundaries or the texture regions, the edge strength detected by the Sobel edge filter becomes small when it is blurred because of defocus. However, for the other pixels, the detected edge strength is always small. For these pixels, the information of the depth cue cannot be extracted from the strength of the Sobel edge. A simple but useful method is to propagate the information of the depth cue from pixels belonging to the boundaries or the textured regions to the other pixels. With this concept, the strength of the Sobel edge is smoothed by a Gaussian filter and then applied for estimating the magnitude of blur and the greatest possible value of the depth from accommodation D_A in [8], i.e.,

$$\arg\max_{D_A} \Pr(D_A) = \alpha_A \times \text{Gaussian Kernal} \otimes \left\| \left(\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \otimes I(x,y), \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \otimes I(x,y) \right) \right\|, \tag{4}$$

where α_A is a scaling factor from the magnitude of the smoothed edge strength to the estimated depth cue, $I(x,y)$ is the luminance value of pixel (x,y) in the input image, and $\otimes$ means the convolution operator.

In the case of depth cue from height in the visual field, [19] provides the depth measurement using image segmentation by computing a minimum spanning tree (MST) on the basis of pixel color and spatial connectivity. The height of the center of gravity of each segment is then applied to assign the depth. To

reduce the exhaustive computation from computing an MST and to create the Bayesian model for the next step, we classify the pixels into three simple classes, namely the sky ceiling (denoted as *SC*), the ground floor (denoted as *GF*), and the others (denoted as *TO*) with the representative colors $\hat{C}_s$ and $\hat{C}_g$ for the classes of the sky ceiling and the ground floor, respectively, instead of the graph-based segmentation discussed in [21]. The representative colors $\hat{C}_s$ and $\hat{C}_g$ are defined in terms of three-dimensional vectors in the RGB color space as follows:

$$\hat{C}_s = (R_s, G_s, B_s) = \frac{1}{|TOP|} \sum_{(x,y)\in TOP} \hat{C}(x,y) , \tag{5}$$

$$\hat{C}_g = (R_g, G_g, B_g) = \frac{1}{|BOT|} \sum_{(x,y)\in BOT} \hat{C}(x,y) , \tag{6}$$

where R_s or R_g, G_s or G_g, and B_s or B_g are the red, green, and blue channels for the representative color vector $\hat{C}_s$ or $\hat{C}_g$. Vector $\hat{C}(x,y) = (R(x,y), G(x,y), B(x,y))$ is composed of the RGB color channels of pixel (x,y). *TOP* and *BOT* are the sets composed of the top 20% and bottom 20% pixels of the given image, respectively.

A pixel will be assigned to the classes of the sky ceiling and the ground floor if and only if the color of the pixel is sufficiently similar to the corresponding representative colors. The classification is discussed in detail at the stage of probability analysis. After the classification, the greatest possible value of the depth from height in the visual field D_H for each pixel (x,y) is assigned on the basis of the result of the classification as follows:

$$\arg\max_{D_H(x,y)} \Pr(D_H(x,y)) = \begin{cases} 0 \text{ if } (x,y)\in SC \\ c + \left(\dfrac{(255-c)\times y}{\max\{q \mid (p,q)\in I\}}\right) \\ \quad \text{if } (x,y)\in GF \\ D_H((x,y+1)) \text{ others} \end{cases} , \tag{7}$$

where c is selected to be 64 in our implementation for approximating the depth generated in [21], I means the input image, and the origin is placed at the top-left corner of the image with the x-axis pointing to the right and the y-axis pointing down.

B. Measurement and Conditional Probability of Depth Cue Measurements

Next, the measurement of each depth cue is defined for further probabilistic analysis. For depth cue from motion, the measurement M_M is defined as the magnitude of the motion vector $(M_{2D,x}, M_{2D,y})$ for each pixel, i.e.,

$$M_M = \sqrt{{M_{2D,x}}^2 + {M_{2D,y}}^2} . \tag{8}$$

For depth cue from accommodation, the measurement M_A is defined as the magnitude of the smoothed value of the Sobel edge for each pixel, i.e.,

$$M_A = \text{Gaussian Kernal} \otimes \left\| \left(\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \otimes I(x,y), \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \otimes I(x,y) \right) \right\| . \tag{9}$$

where $\otimes$ means the convolution operator.

For depth cue from height in the visual field, the measurement M_H is defined as the color difference between a given pixel with color $\hat{C}(x,y)$ and the representative colors $\hat{C}_S$ and $\hat{C}_G$ defined in (5) and (6), i.e.,

$$M_H = (\| \hat{C}(x,y) - \hat{C}_s \|, \| \hat{C}(x,y) - \hat{C}_g \|). \tag{10}$$

Next, the conditional probability of each depth cue given its measurement is modeled using the Bayesian Theorem, with two assumptions. The first is that the prior probability of each depth cue is uniform, i.e., $\Pr(D_M)$, $\Pr(D_A)$, and $\Pr(D_H)$ are constant for every possible cue value. This implies that there is no preferred value of the depth cue before the measurement. The other is that we employ a Gaussian prior probability density for the amount of measurement with a given value for each depth cue (i.e., $\Pr(M_M|D_M)$, $\Pr(M_A|D_A)$, or $\Pr(M_H|D_H)$).

The relationship between depth cues and the corresponding measurements is then explored to determine the parameters in the Gaussian prior. Two types of relationships exist; they are defined as zero measurement preference (abbreviated as ZMP) and non-measurement preference (abbreviated as NMP) in this paper. We discuss them in the following paragraphs. Note that on the basis of the conditional probability, the optimal value of each depth cue (i.e., the value with the greatest possibility) can be derived. This value is supposed to be consistent with the initial estimated depth cue based on Bayesian Probability, named as a depth cue consistent condition (abbreviated as D3C). The conditional probability density functions, $\Pr(M_M|D_M)$, $\Pr(M_A|D_A)$, and $\Pr(M_H|D_H)$ are then modeled using the above relationship and condition.

In the case of the accommodation and the motion, we observe that the preferred values of the measurements (i.e., smoothed edge strength or motion vector) are zero, i.e., ZMP. In the case of the accommodation, the reason that the ZMP exists is that the area of non-boundary and the non-textured regions is large in general cases. Similarly, in the case of the motion, the reason that the ZMP exists is that the area of the static objects or regions is also large in general. This implies that the probability of the measurement value for a given depth cue has a peak at zero, irrespective of the depth cue value. With D3C, the conditional probabilities and the initially estimated depth cues must satisfy the following equations:

$$\arg\max_{D_M(x,y)} \Pr(D_M(x,y)) = \arg\max_{D_M'(x,y)} \{\Pr(D_M'(x,y) \mid M_M(x,y))\}, \tag{11}$$

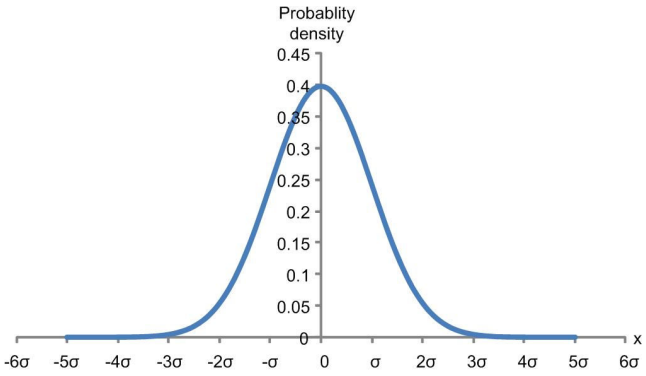

Fig. 6. Zero-mean normal distribution with arbitrary variance σ^2, $N_{0,\sigma^2}(x)$

$$\arg\max_{D_A(x,y)} \Pr(D_A(x,y)) = \arg\max_{D_A'(x,y)}\{\Pr(D_A'(x,y)\,|\,M_A(x,y))\}, \tag{12}$$

where D_M' and D_A' denote the value of the depth cue from motion and accommodation, respectively. Based on Bayesian Theorem, the probability terms in (11) and (12) can be reduced as follows, i.e.,

$$\Pr(D_M'\,|\,M_M) = \frac{\Pr(M_M\,|\,D_M')\Pr(D_M')}{\Pr(M_M)}, \tag{13}$$

$$\Pr(D_A'\,|\,M_A) = \frac{\Pr(M_A\,|\,D_A')\Pr(D_A')}{\Pr(M_A)}, \tag{14}$$

where $\Pr(M_M|D_M')$ and $\Pr(M_A|D_A')$ are zero-mean normal distributed with variances σ_M^2 and σ_A^2, and $\Pr(D_M')$, $\Pr(D_A')$ are uniformly distributed with the above-mentioned assumptions. Note that σ_M and σ_A are functions of D_M' and D_A', respectively. For further discussions, the relationships between the variances and the values of the depth cues are defined as the functions f_M and f_A, i.e.,

$$\sigma_M = f_M(D_M'), \tag{15}$$

$$\sigma_A = f_A(D_A'). \tag{16}$$

Further, the values of $\Pr(M_M)$ and $\Pr(M_A)$ are not important from the perspective of the maximum likelihood for any given measurements M_M and M_A. On the basis of (11)–(16) and the above discussions, the most possible values of D_M and D_A can be reformulated as follows:

$$\begin{aligned}\arg\max_{D_M(x,y)} \Pr(D_M(x,y)) &= \arg\max_{D_M'(x,y)}(\Pr(M_M(x,y)\,|\,D_M'(x,y))) \\ &= \arg\max_{D_M'(x,y)}\left(\frac{1}{\sqrt{2\pi}\sigma_M}\exp\left(-\frac{M_M(x,y)^2}{2\sigma_M{}^2}\right)\right) \\ &= \arg\max_{D_M'(x,y)}\left(\frac{1}{\sqrt{2\pi}f_M(D_M')}\exp\left(-\frac{M_M(x,y)^2}{2f_M(D_M')^2}\right)\right),\end{aligned} \tag{17}$$

$$\begin{aligned}\arg\max_{D_A(x,y)} \Pr(D_A(x,y)) &= \arg\max_{D_A'(x,y)}(\Pr(M_A(x,y)\,|\,D_A'(x,y))) \\ &= \arg\max_{D_A'(x,y)}\left(\frac{1}{\sqrt{2\pi}\sigma_A}\exp\left(-\frac{M_A(x,y)^2}{2\sigma_A{}^2}\right)\right) \\ &= \arg\max_{D_A'(x,y)}\left(\frac{1}{\sqrt{2\pi}f_A(D_A')}\exp\left(-\frac{M_A(x,y)^2}{2f_A(D_A')^2}\right)\right).\end{aligned} \tag{18}$$

Thus, we attempt to link the greatest possible values of the depth cues (i.e., D_M and D_A) and the corresponding measurements M_M and M_A by the variances σ_M^2 and σ_A^2. For further discussion, we first prove the following lemma:

Lemma 1: For any positive real number x and σ, $N_{0,\sigma^2}(x) \le N_{0,x^2}(x)$ and $N_{0,\sigma^2}(x) = N_{0,x^2}(x)$ iff $\sigma = x$, where $N_{0,\sigma^2}(x)$ denotes the probability density functions of a zero-mean normal distribution with arbitrary variance σ^2, as shown in Fig. 6.

Proof: We define

$$G_x^N(\sigma) = N_{0,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{19}$$

Then, we have

$$\frac{\partial G_x^N(\sigma)}{\partial\sigma} = \frac{\partial N_{0,\sigma^2}(x)}{\partial\sigma} = \frac{x^2-\sigma^2}{\sqrt{2\pi}\sigma^4}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{20}$$

The maximum value of $G_x^N(\sigma)$ occurs if and only if $\sigma = x$ by using (20). This completes the proof.

On the basis of lemma 1 and (17)–(18), once we have the measurements M_M and M_A for pixel (x,y), the greatest possible value of the depth cues can be reformulated as the following equations:

$$\begin{aligned}&\arg\max_{D_M(x,y)} \Pr(D_M(x,y)\,|\,M_M(x,y)) \\ &= \arg\max_{D_M'(x,y)}\left(\frac{1}{\sqrt{2\pi}f_M(D_M')}\exp\left(-\frac{M_M(x,y)^2}{2f_M(D_M')^2}\right)\right),\end{aligned} \tag{21}$$

$$\begin{aligned}&\arg\max_{D_A(x,y)} \Pr(D_A(x,y)\,|\,M_A(x,y)) \\ &= \arg\max_{D_A'(x,y)}\left(\frac{1}{\sqrt{2\pi}f_A(D_A')}\exp\left(-\frac{M_A(x,y)^2}{2f_A(D_A')^2}\right)\right),\end{aligned} \tag{22}$$

where $\sigma_M = f_M(D_M') = M_M$ if D_M' = the greatest possible D_M corresponding to M_M, and $\sigma_A = f_A(D_A') = M_A$ if D_A' = the greatest possible D_A corresponding to M_A.

Moreover, by (3) and (4), we infer that $\sigma_M = M_M = D_M'/\alpha_M$, and $\sigma_A = M_A = D_A'/\alpha_A$. The conditional probability density functions $\Pr(M_M|D_M')$ and $\Pr(M_A|D_A')$ for given D_M' and D_A' are then derived as follows:

$$\Pr(M_M \mid D_M') = N_{0,(\frac{D_M'}{\alpha_M})^2}(M_M), \quad (23)$$

$$\Pr(M_A \mid D_A') = N_{0,(\frac{D_A'}{\alpha_A})^2}(M_A). \quad (24)$$

In contrast, in the case of the height in the visual field, the depth cue is directly generated on the basis of the classifications as in (7). Notably, the result of the classification is determined on the basis of the color difference between a pixel and the representative colors $\hat{C}_s$ and $\hat{C}_g$, i.e., measurement M_H. To build the relationship between this depth cue $D_{cue,H}$ and the measurement $M_H = (M_{H,s}, M_{H,g}) = (\|\hat{C}(x,y) - \hat{C}_s\|, \|\hat{C}(x,y) - \hat{C}_g\|)$, we first derive the conditional probability density function of M_H given the result of the classification, denoted as $\Pr(M_H \mid Class)$, where $\Pr(M_{H,s} \mid Class = SC)$ and $\Pr(M_{H,g} \mid Class = GF)$ are normally distributed. In addition, $\Pr(M_{H,s} \mid Class = TO)$ and $\Pr(M_{H,g} \mid Class = TO)$ are uniformly distributed for each pixel. It is noticeable that *SC* mean the class of sky ceiling, *GF* means the class of ground floor, and *TO* means the class of the others. Since the representative colors $\hat{C}_s$ and $\hat{C}_g$ are independent in general cases, a reasonable assumption that $\Pr(M_{H,s} \mid Class = SC)$ and $\Pr(M_{H,g} \mid Class = GF)$ are also independent. Notably, from (7) and D3C, $\Pr(M_{H,s} \mid Class = SC)$, $\Pr(M_{H,g} \mid Class = GF)$, $\Pr(M_{H,s} \mid Class = TO)$, and $\Pr(M_{H,g} \mid Class = TO)$ can also be described as $\Pr(M_{H,s} \mid Depth(x,y) = 0)$, $\Pr(M_{H,g} \mid Depth(x,y) = c+y/(\max\{q \mid (p,q)\in I\} \times (255-c))$, $\Pr(M_{H,s} \mid Depth(x,y) = Depth(x,y+1))$, and $\Pr(M_{H,g} \mid Depth(x,y) = Depth(x,y+1))$, respectively. The mean of the two Gaussian priors $\Pr(M_{H,s} \mid Class = SC)$ and $\Pr(M_{H,g} \mid Class = GF)$ is zero, and the constant variance σ_H for all the pixels reflects the cross-correlation between the measurement and the value of the depth cue (i.e., the classification result). A smaller correlation implies a larger variance. In our experiments, σ_H is empirically selected to be 8 when the values of all the color channels are 8-bit numbers. We call this relationship an NMP relationship because the relationship is not based on the preference of the measurement as compared to ZMP. Notably, the classification is also done here by Bayesian's Theorem, i.e.,

$$SC = \{(x,y) \mid \frac{\Pr(M_{H,s} \mid D_H(x,y) = 0)}{\Pr(M_{H,g} \mid D_H(x,y) = c + \frac{(255-c)\times y}{\max\{q \mid (p,q)\in I\}})} \geq 1 \wedge \frac{\Pr(M_{H,s} \mid D_H(x,y) = 0)}{\Pr(M_{H,s} \mid D_H(x,y) = D_H(x,y+1))} \geq 1\}, \quad (25)$$

$$GF = \{(x,y) \mid \frac{\Pr(M_{H,s} \mid D_H(x,y) = 0)}{\Pr(M_{H,g} \mid D_H(x,y) = c + (\frac{(255-c)\times y}{\max\{q \mid (p,q)\in I\}})} < 1 \wedge \frac{\Pr(M_{H,g} \mid D_H(x,y) = c + (\frac{(255-c)\times y}{\max\{q \mid (p,q)\in I\}})}{\Pr(M_{H,g} \mid D_H(x,y) = D_H(x,y+1))} \geq 1\}, \quad (26)$$

$$TO = \{(x,y)\} - SC - GF, \quad (27)$$

The greatest possible D_H with the given measurement M_H is then estimated on the basis of the result of the classification, as shown in (7).

C. Noise Suppression for Depth Cues

We first define the similarity measure *SM* of two given pixels (x, y) and (x', y') on the basis of the color and spatial proximity [22], i.e.,

$$SM((x,y),(x',y')) = \exp(-\frac{|(x'-x)|^2}{2\sigma_d^2} - \frac{|(y'-y)|^2}{2\sigma_d^2} - \frac{|I(x',y') - I(x,y)|^2}{2\sigma_i^2}), \quad (28)$$

where σ_d *and* σ_i are empirically selected to be 10 and 30 for the spatial coordinates, x and y, and the pixel value $I(x,y)$, respectively. A probability-weighted averaging is then applied to suppress the uncertainties of the estimated value of the depth cues D_{est}, instead of promoting the cues by the depth cue interaction as the brain does. The noise-suppressed depth D_{sup} is denoted as follows:

$$D_{sup}(x,y) = \frac{\sum_{(x',y')\in N(x,y)} \frac{\Pr(D_{est}(x',y') \mid M_c(x',y'))}{\Pr(D_{est}(x',y') \mid M_c(x',y'))} \times SM((x,y),(x',y')) \times D_{est}(x',y')}{\sum_{(x',y')\in N(x,y)} \frac{\Pr(D_{est}(x',y') \mid M_c(x',y'))}{\Pr(D_{est}(x',y') \mid M_c(x',y'))} \times SM((x,y),(x',y'))}, \quad (29)$$

where $N(x,y)$ and $M_c(x,y)$ denote the spatial neighborhood of the pixel (x,y) and the corresponding measurement for pixel (x,y), respectively.

D. Reliability Analysis for Fusion of Depth Cues

A reliability analysis is applied for the fusion of the depth cues. The reliability values are necessary for fusing the depth cues in order to mimic the manner in which the brain generates depth perception. Instead of analyzing the reliability from ancillary cues such as vestibular inputs, we attempt to obtain the reliability by referencing the analyzed probability distributions. Moreover, the distributions of the depth values for similar pixels are considered. We define the reliability with variance and denote it as reliability variance (abbreviated as RV) as follows, and the value of the depth cue with a relatively small RV means that this value is more reliable. To estimate the reliability, both the conditional probability discussed in Sec. III-B and the consistency of the value of the depth cue for similar pixels are considered along with the concept of variance. For the conditional probability, we define probability variance (abbreviated as PV) by applying the variances of the probability density distributions in Sec. III-B, i.e.,

$$\mathrm{PV} = \mathrm{var}(\Pr(D_{est}(x',y') \mid M(x',y')) \quad (30)$$

Next, we attempt to measure the consistency of the value of the

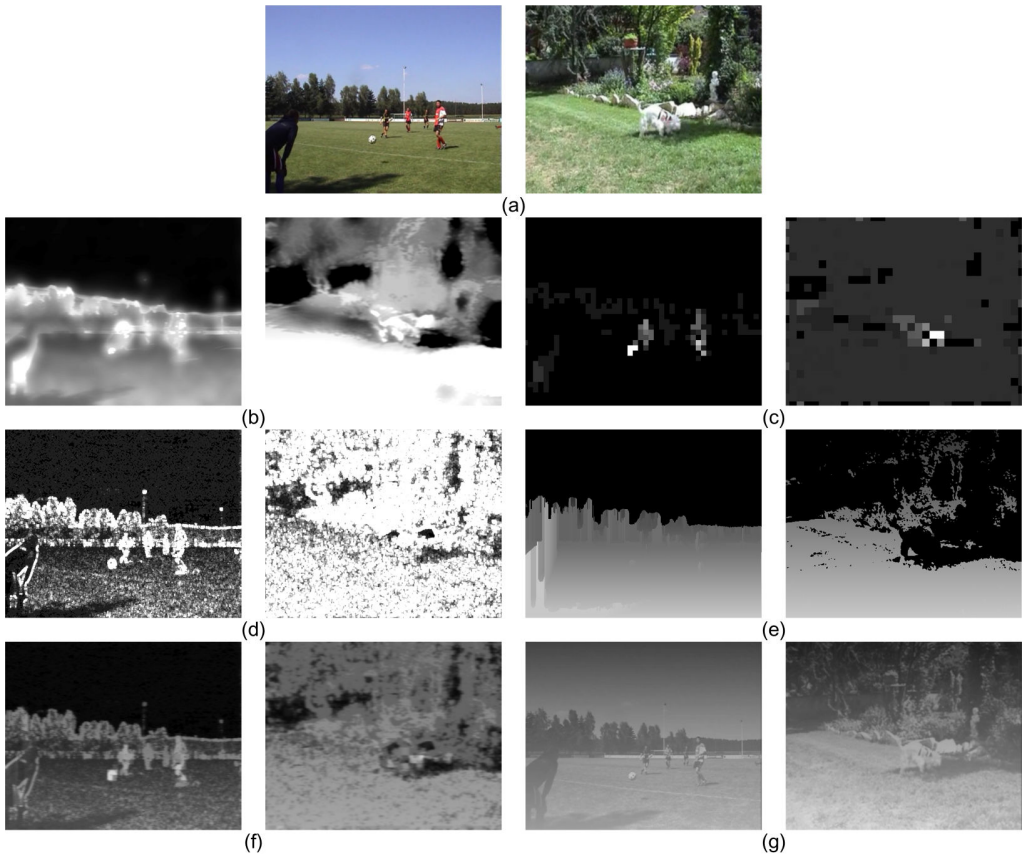

Fig. 7. Experimental results: (a) test sequences Soccer and Jojo; (b) depth from the proposed method; (c) depth from motion [5]; (d) depth from edge [8]; (e) depth from height in the visual field[21] (f) linear combination based on [15], of (c), (d), and (e); and (g) depth from color [10].

depth cue for similar pixels. This consistency is still defined with the concept of variance, named similarity variance (abbreviated as SV). To estimated SV, we first find similar pixels, where both the color similarity and the spatial proximity are exploited. A dissimilarity measurement function between the two pixels (p_x,p_y) and (q_x,q_y), denoted as $K((p_x,p_y), (q_x,q_y))$, is defined, i.e.,

$$K((p_x,p_y),(q_x,q_y)) = \sqrt{(p_x-q_x)^2+(p_y-q_y)^2+\lambda_{lum}{}^2(I(p_x,p_y),\ I(q_x,q_y))^2} \quad (31)$$

where $I(p_x,p_y)$ and $I(q_x,q_y)$ represent and the luminance for the pixels (p_x,p_y) and (q_x,q_y), respectively, and λ_{lum} is the weighting coefficient for the luminance similarity, which defines the luminance-importance-to-space-importance ratio for the pixel similarity. Since σ_d and σ_i are selected to be 10 and 30 in the similarity measure SM, the ratio λ_{lum} is set as 1/3 for the sake of consistency. Then, the pixel similarity $PS((p_x,p_y), (q_x,q_y))$, which defines how pixel (q_x,q_y) is similar to a given pixel (p_x,p_y), can be defined and normalized to [0,1] with $K((p_x,p_y), (q_x,q_y))$. A larger value of $K((p_x,p_y), (q_x,q_y))$ implies a smaller value of $PS((p_x,p_y), (q_x,q_y))$. Next, the SV from the value of the depth cue values for similar pixels can be computed by a weighted averaging for each pixel, i.e.,

$$\mathrm{SV} = \frac{\sum_{(q_x,q_y)\in N((p_x,p_y))} PS((p_x,p_y),(q_x,q_y)) \times [D_{est}((p_x,p_y)) - D_{est}((q_x,q_y))]^2}{\sum_{(q_x,q_y)\in N((p_x,p_y))} PS((p_x,p_y),(q_x,q_y))}, \quad (32)$$

where $N(x,y)$ and $D_{est}(x,y)$ mean the spatial neighborhood of the pixel (x,y) and the estimated value of depth cue for pixel (x,y), respectively. Finally, the RV can be calculated by summing the PV and the SV variance, i.e.,

$$\mathrm{RV} = \mathrm{PV} + \mathrm{SV} \quad (33)$$

E. Reliability-Based Fusion of Multiple Depth Cues

We attempt to mimic the manner in which the brain fuses depth cues with reliability at this stage. Since psychologists have found that this process can be simplified as a locally weighted fusion of the promoted depth cues, we solve the optimal weightings for each pixel in the Bayesian sense. Since we have assumed that each depth cue has a Gaussian-prior probabilistic distribution above, here, the reliability variance (i.e., RV) calculated in Sec. III-D is used for updating the variance of the distribution. Then, finding the greatest possible weighting becomes a Bayesian interference problem. The weighting $w(x,y)$ can be approximated to be proportional to the reciprocal of the variance $\mathrm{RV}(x,y)$, i.e.,

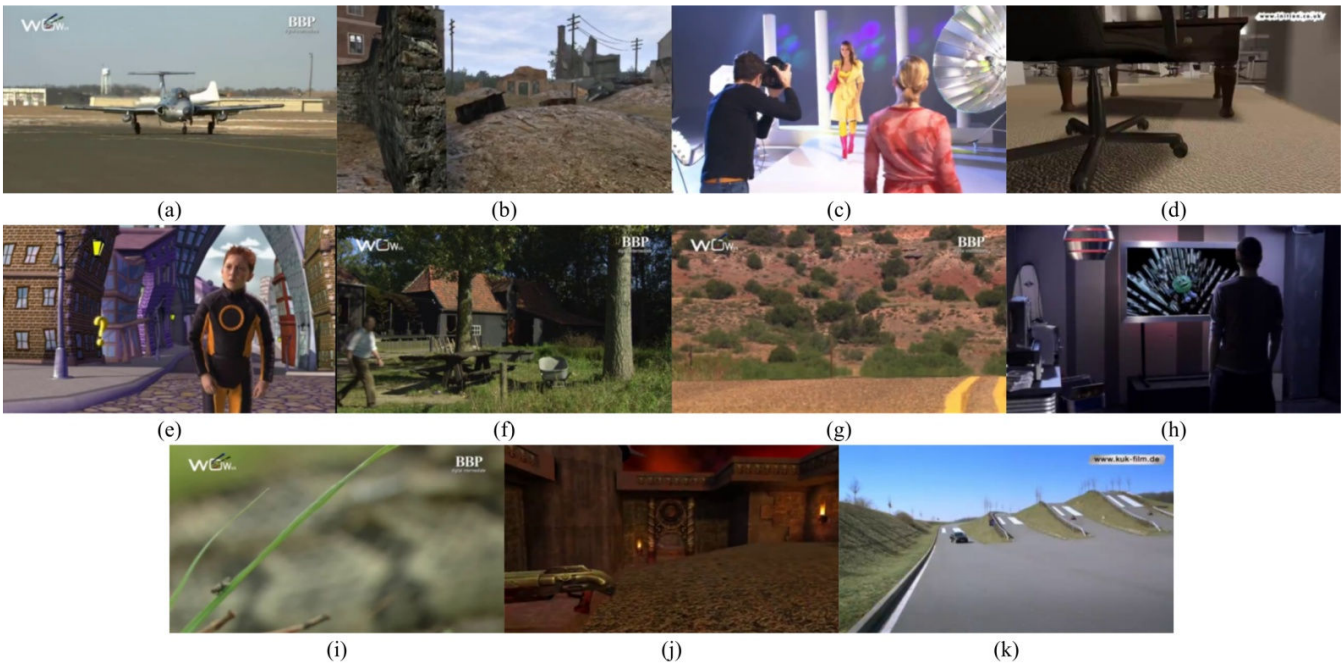

Fig. 8. Test sequences illustrating [24] (a) airshow, (b) cod, (c) fashion, (d) showreel, (e) spykids, (f) watermelon, (g) corvette, (h) kuk, (i) nature, (j) quake, and (k) kukbo.

$$w(x,y)=\frac{1}{RV(x,y)}. \tag{34}$$

Finally, the noise-suppressed depth cues are fused with the optimal weightings in (34) to generate the final depth D_F, i.e.,

$$D_F(x,y)=\frac{\sum_{i=M,A,H} w_i(x,y)D_{sup,i}(x,y)}{\sum_{i=M,A,H} w_i(x,y)}. \tag{35}$$

where w_M, w_A, and w_H are the weightings for motion, accommodation, and height in the visual field, respectively, and $D_{sup,M}$, $D_{sup,A}$, and $D_{sup,H}$ are the noise-suppressed values of depth cue from motion, accommodation, and height in the visual field, respectively.

IV. Experimental Results and Comparisons

In this section, the qualities of the depth maps generated by the proposed method are measured by both subjective and objective assessments as compared to the conventional deterministic methods. First of all, some sample frames of the tested sequences and the corresponding depth maps generated using various types of depth cues (depth from motion [5], depth from edge [8], depth from height in the visual field [21], depth from linear combination [15] of motion, accommodation, height in the visual field, and depth from color [10] and this work) are shown in Fig. 7. For the sequences Soccer and Jojo, only parts of the players on the soccer field and the dog on the grass field appear to protrude from motion [5], as shown in Fig. 7(c). There are a large number of noises in the sky, trees, and ground in the case of depth from edge [8], as shown in Fig. 7(d). In the case of depth from height in the visual field [21], the players, the dog, the wall, and the stones appear to vanish into the background, as shown in Fig. 7(e). Some of the players and trees also appear to vanish into the background in the case of depth from color [10], as shown in Fig. 7(g). As discussed in Sec. I, the lack of some depth cues does lead to a deficient depth estimation. A simple example of the heuristic combination [15] is shown in Fig. 7(f). Although most of the desired depth cues are integrated, the noises (i.e., uncertain depth cues) are also included in the depth estimation. The proposed method can overcome the problems and provide better depth perception, as shown in Fig. 7(b). From the experimental results, it is concluded that a combination of multiple depth cues (without reliability) (i.e., [15]) retains more information and provides a more satisfactory depth perception than methods that use a single depth cue (i.e., [5], [8], and [21]). However, some incorrect information is also retained. In contrast, [10] retains a sharper depth on object boundaries and provides more satisfactory results. The proposed method, which fuses depth cues with reliability, retains robust depth cues and provides satisfactory object boundaries on the depth map. Consequently, it performs better in subjective views. In Fig. 7, the depth map generated by the proposed method outperforms in both boundary preservation and noise suppression, although there

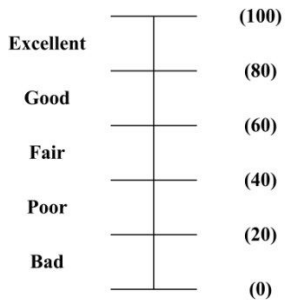

Fig. 9. Five-segment rating scale used for assessing the depth quality.

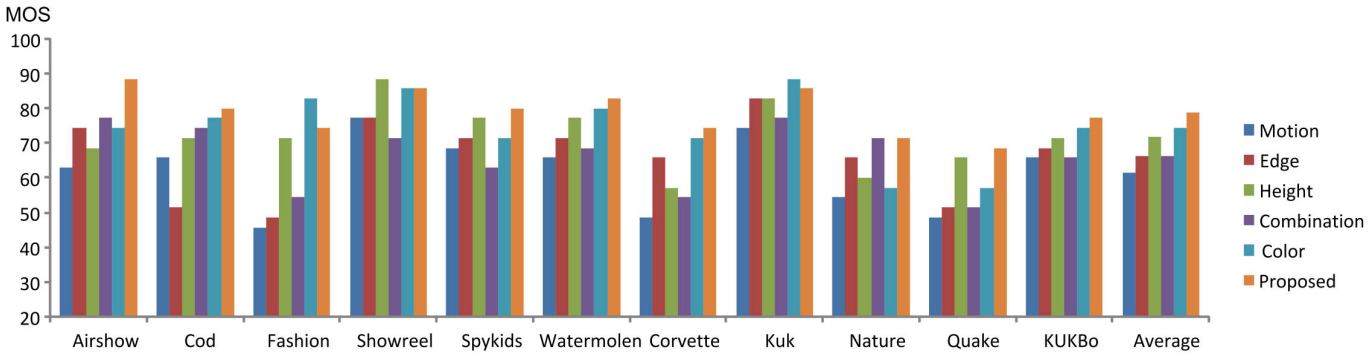

Fig. 10. Subjective evaluations: mean opinion score.

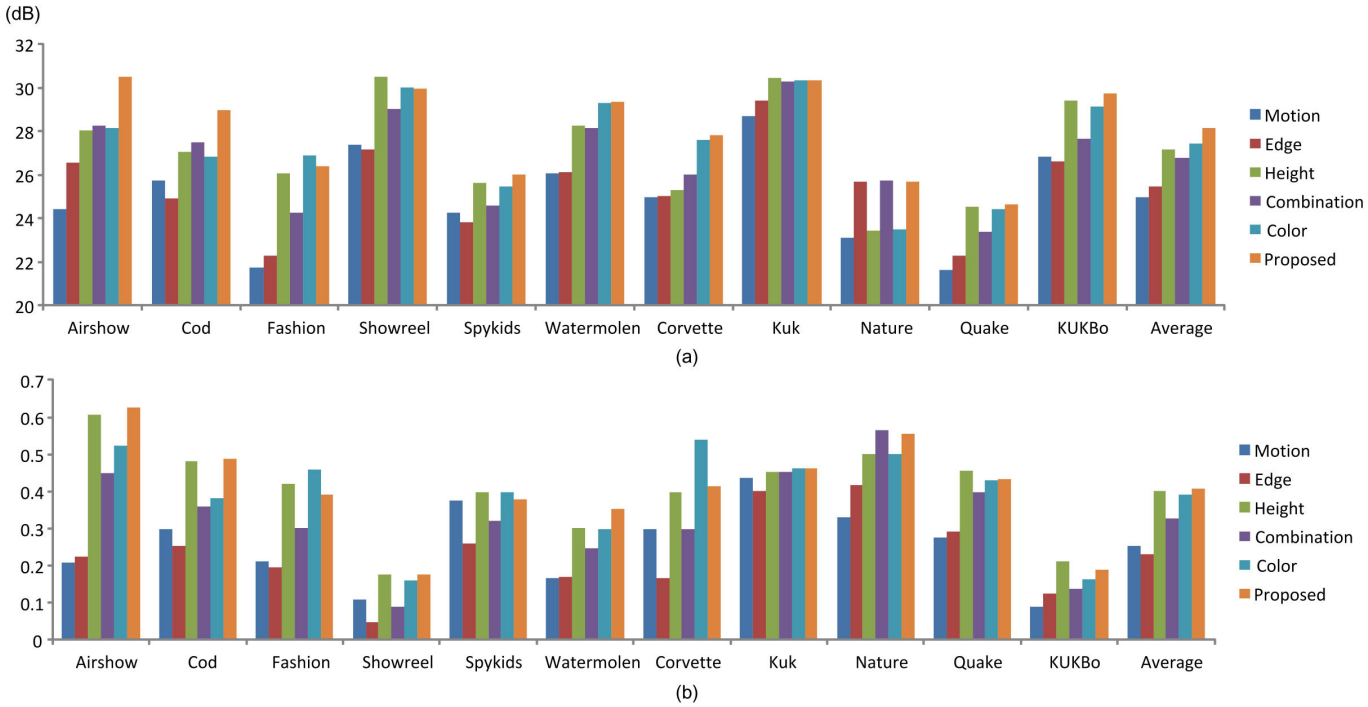

Fig. 11. Objective measurements: (a) modified PSNR and (b) depth distortion model.

are still some regions that are over-smoothed in depth (this is less objectionable), especially the low-textured regions. However, even if the proposed method successfully integrates the above depth cues, there are still some regions with a deficient depth estimation, such as the street lights in Soccer. The reason for this may be that some of the important depth cues (see Table I), such as occlusion, are not included in the proposed system. The probabilistic model for these cues will be the next critical issue for improving the proposed system. For a further subjective comparison, more reconstructed 3D videos are posted on our website [23].

For the further subjective assessments, a subjective evaluation was also performed by fourteen people with normal or correct-to-normal visual acuity and stereo acuity. The participants watched the converted stereoscopic video from various types of depth maps in a random order and were asked to rate each video according to two factors, i.e., depth perception and visual comfort. Sample frames of the tested sequences [24] are shown in Fig. 8. The overall quality for depth perception was assessed using a five-segment scale, as shown in Fig. 9, and the overall mean opinion scores (MOSs) obtained in the experiments for the evaluation sequences are shown in Fig. 10. This shows that the proposed method outperforms the other methods in most sequences because of the appropriate fusion of multiple depth cues.

For an objective assessment, the modified peak signal-to-noise ratio (PSNR) [25] and disparity distortion model (DDM) [26] are checked with an ideal depth map from [24]. Depth maps with higher values in terms of modified PSNR and DDM are better. Higher modified PSNR means that the boundaries of the estimated depth are consistent with the boundaries of the objects in the test image, and higher DDM implies that the estimated depth is more consistent with the ideal depth map in terms of the depth layers. The results are shown in Tables II and III and Fig. 11. The proposed method also provides better results than the conventional deterministic methods in most cases. The experimental results show that the proposed method outperforms up to 0.70-3.14 dB and 0.0059-0.1517 in terms of the modified PSNR and disparity distortion model (DDM) as compared to the existing

TABLE II
EXPERIMENTAL RESULTS IN TERMS OF MODIFIED PSNR

Approaches/ Sequences	Motion	Edge	Height	Combi-nation	Color	**Pro-posed**
Airshow	24.41	26.53	28.06	28.26	28.15	**30.47**
Cod	25.76	24.89	27.07	27.47	26.83	**28.99**
Fashion	21.75	22.28	26.07	24.28	**26.87**	26.37
Showreel	27.37	27.18	**30.50**	29.00	30.03	29.94
Spykids	24.24	23.80	25.60	24.60	25.48	**26.00**
Watermelon	26.06	26.10	28.26	28.13	29.31	**29.33**
Corvette	24.94	25.00	25.30	26.00	27.60	**27.83**
Kuk	28.70	29.40	30.46	30.26	30.33	**30.35**
Nature	23.10	25.70	23.41	25.71	23.48	**25.68**
Quake	21.65	22.27	24.55	23.38	24.43	**24.63**
Kukbo	26.83	26.63	29.38	27.68	29.15	**29.76**
Average	24.98	25.44	27.15	26.80	27.42	**28.12**

(Ideal depth map is ∞; unit: dB)

TABLE III
EXPERIMENTAL RESULTS IN TERMS OF DEPTH DISTORTION MEASURE

Approaches/ Sequences	Motion	Edge	Height	Combi-nation	Color	**Pro-posed**
Airshow	0.2086	0.2243	0.6070	0.4479	0.5244	**0.6252**
Cod	0.2988	0.2544	0.4820	0.3587	0.3809	**0.4884**
Fashion	0.2128	0.1951	0.4207	0.3011	**0.4576**	0.3922
Showreel	0.1086	0.0474	0.1763	0.0886	0.1585	**0.1769**
Spykids	0.3737	0.2606	**0.3984**	0.3202	0.3961	0.3768
Watermelon	0.1651	0.1696	0.3004	0.2475	0.2990	**0.3525**
Corvette	0.2972	0.1655	0.3987	0.2996	**0.5403**	0.4152
Kuk	0.4361	0.4024	0.4526	0.4512	0.4609	**0.4628**
Nature	0.3317	0.4168	0.4996	**0.5639**	0.4999	0.5548
Quake	0.2754	0.2917	**0.4550**	0.3963	0.4312	0.4338
Kukbo	0.0879	0.1240	**0.2111**	0.1361	0.1622	0.1882
Average	0.2542	0.2320	0.4002	0.3283	0.3919	**0.4061**

(Ideal depth map is 1.0000)

algorithms.

The performance of the two important steps in the proposed framework, noise suppression of each depth cue and reliability-based fusion of multiple depth cues, are discussed in detail in the following. To clarify the importance of noise suppression of depth cues, we convert each test monoscopic sequence into 3D by each depth cue with/without the proposed noise suppression respectively. The modified PSNR and DDM are applied for the objective quality assessment.

Tables IV and V show the experimental results. The conversion with the proposed suppression provides better results in most cases. The experimental results show that the proposed suppression outperforms 0.22-1.62 dB in terms of the modified PSNR and has compatible results in terms of DDM. That means the suppression refines the boundaries of the estimated depth well with preserving the reconstructed depth layers.

We also try to clarify the importance of the proposed reliability-based fusion of depth cues. Each test monoscopic sequence is converted into 3D by linear combination and the reliability-based fusion of the noise-suppressed depth cues, respectively. Both the modified PSNR and DDM are applied for the objective quality assessment. Tables VI and VII show the experimental results. The reliability-based fusion provides better results in most cases. The experimental results show that the proposed fusion outperforms 0.58 dB in terms of the modified PSNR and also has compatible results in terms of DDM. That also implies the reliability-based fusion can make the boundaries of the estimated depth more consistent with the

TABLE IV
EXPERIMENTAL RESULTS IN TERMS OF MODIFIED PSNR – WITH / WITHOUT NOISE SUPPRESSION OF DEPTH CUES

Approaches/ Sequences	Motion	Motion + Suppression	Difference	Edge	Edge + Suppression	Difference	Height	Height + Suppression	Difference
Airshow	24.41	**25.93**	1.52	26.53	**28.73**	2.20	**28.06**	27.94	-0.12
Cod	25.76	**28.95**	3.19	24.89	**27.41**	2.52	27.07	**28.17**	1.10
Fashion	21.75	**23.34**	1.59	22.28	**24.38**	2.10	26.07	**26.12**	0.05
Showreel	27.37	**28.89**	1.52	27.18	**28.75**	1.57	30.50	**30.67**	0.17
Spykids	24.24	**25.31**	1.07	23.80	**25.55**	1.75	25.60	**26.05**	0.45
Watermelon	26.06	**26.29**	0.23	26.10	**28.13**	2.03	28.26	**28.76**	0.50
Corvette	24.94	**25.33**	0.39	25.00	**27.14**	2.14	25.30	**25.76**	0.46
Kuk	28.70	**29.32**	0.62	29.40	**30.01**	0.61	**30.46**	30.33	-0.13
Nature	23.10	**23.25**	0.15	**25.70**	25.47	-0.23	**23.41**	23.39	-0.02
Quake	21.65	**22.88**	1.23	22.27	**23.73**	1.46	24.55	**24.64**	0.09
Kukbo	26.83	**28.23**	1.40	26.63	**28.31**	1.68	**29.38**	29.26	-0.12
Average	24.98	**26.16**	1.18	25.44	**27.06**	1.62	27.15	**27.37**	0.22

(Ideal depth map is ∞; unit: dB)

TABLE V
EXPERIMENTAL RESULTS IN TERMS OF DEPTH DISTORTION MEASURE – WITH / WITHOUT NOISE SUPPRESSION OF DEPTH CUES

Approaches/ Sequences	Motion	Motion + Suppression	Difference	Edge	Edge + Suppression	Difference	Height	Height + Suppression	Difference
Airshow	0.2086	**0.2109**	0.0023	0.2243	**0.2354**	0.0111	0.6070	**0.6191**	0.0121
Cod	0.2988	**0.3770**	0.0782	0.2544	**0.3982**	0.1438	0.4820	**0.4924**	0.0104
Fashion	0.2128	**0.2143**	0.0015	0.1951	**0.2252**	0.0301	0.4207	**0.4219**	0.0012
Showreel	**0.1086**	0.0926	-0.0160	0.0474	**0.0659**	0.0185	0.1763	**0.1801**	0.0038
Spykids	0.3737	**0.3911**	0.0174	0.2606	**0.2962**	0.0356	**0.3984**	0.3960	-0.0024
Watermelon	**0.1651**	0.1443	-0.0208	0.1696	**0.2321**	0.0625	0.3004	**0.3155**	0.0151
Corvette	**0.2972**	0.2571	-0.0401	**0.1655**	0.1497	-0.0158	0.3987	**0.4255**	0.0268
Kuk	**0.4361**	0.4253	-0.0108	0.4024	**0.4426**	0.0402	**0.4526**	0.4446	-0.0080
Nature	0.3317	**0.3362**	0.0045	0.4168	**0.4243**	0.0075	**0.4996**	0.4895	-0.0101
Quake	**0.2754**	0.2722	-0.0032	0.2917	**0.3488**	0.0571	**0.4550**	0.4431	-0.0119
Kukbo	**0.0879**	0.0769	-0.0110	0.1240	**0.1269**	0.0029	**0.2111**	0.2088	-0.0023
Average	0.2542	**0.2544**	0.0002	0.2320	**0.2678**	0.0358	0.4002	**0.4033**	0.0031

TABLEVI
EXPERIMENTAL RESULTS IN TERMS OF MODIFIED PSNR - LINEAR COMBINATION VERSUS RELIABILITY-BASED FUSION

Approaches/ Sequences	Linear Combination	Reliability-Based Fusion	**Difference**
Airshow	28.56	**30.47**	1.91
Cod	**29.88**	28.99	-0.89
Fashion	24.97	**26.37**	1.40
Showreel	29.57	**29.94**	0.37
Spykids	**26.13**	26.00	-0.13
Watermelon	28.90	**29.33**	0.43
Corvette	27.15	**27.83**	0.68
Kuk	29.95	**30.35**	0.40
Nature	25.45	**25.68**	0.23
Quake	23.85	**24.63**	0.78
Kukbo	28.52	**29.76**	1.24
Average	27.54	**28.12**	0.58

(unit: dB)

TABLEVII
EXPERIMENTAL RESULTS IN TERMS OF DEPTH DISTORTION MEASURE - LINEAR COMBINATION VERSUS RELIABILITY-BASED FUSION

Approaches/ Sequences	Linear Combination	Reliability-Based Fusion	**Difference**
Airshow	0.4634	**0.6252**	0.1618
Cod	**0.5248**	0.4884	-0.0364
Fashion	**0.4140**	0.3922	-0.0218
Showreel	0.0982	**0.1769**	0.0787
Spykids	**0.4960**	0.3768	-0.1192
Watermelon	0.3450	**0.3525**	0.0075
Corvette	0.3107	**0.4152**	0.1045
Kuk	**0.4929**	0.4628	-0.0301
Nature	**0.5550**	0.5548	-0.0002
Quake	**0.5447**	0.4338	-0.1109
Kukbo	0.1450	**0.1882**	0.0432
Average	0.3991	0.4061	0.0070

boundaries of the objects.

In summary, our proposed framework provides significant improvement in both subjective and objective experimental results. A considerably better depth perception is achieved.

V. CONCLUSIONS

In this paper, a brain-inspired 2D-to-3D conversion is presented, which generates depth information by exploiting the reliability of multiple depth cues. The algorithm is significantly better than the conventional deterministic methods such as depth from edge or motion, and outperforms the conventional combination of previous methods. It also outperforms the color-based method, especially for regions where the warm–cool color assumption for depth fails. We also show that the proposed algorithm can simulate the best depth perception generator, the human brain, by depth cue estimation, noise suppression, and fusion. The proposed algorithm does not increase noise in the depth information or produce distortion in the synthesized views for complex scenes as some other conventional multiple depth cue combination methods do, although some over-smoothed depth effects (less objectionable) are observed in the low-textured regions. Although there are still several semi-automatic depth generators found in the literature that are likely to perform better than the proposed method, the cost of human-in-the-loop is too high to be used for converting all the conventional 2D videos into 3D. With consideration of both the efficiency and the 3D quality, the proposed 2D-to-3D conversion is more suitable for creating a significant number of 3D videos from 2D.

However, there are still some challenges to overcome. One of the most important challenges is the completeness of depth cues. Depth cues such as occlusion and relative size also play significant roles in the depth perception mechanism, but they are still not well modeled in the proposed method. We believe that the quality of the depth could be improved by considering the two depth cues. Another challenge is the computation. Since the proposed methods reconstruct depth on the basis of a reliability-form probabilistic analysis as shown above, a considerable amount of exponential computation is required. The amount of exponential computation is in the order of the resolution of input video multiplied with the size of the neighborhood for each frame when calculating the similarity measure in (28). For an 1080p HD video, the amount of exponential computation is approximately 531 million (1920×1080×16×16 when the size of the neighborhood is 16×16). Thus, the computation complexity is not sufficiently low for a real-time implementation, and, therefore, a computation reduction is required for the implementation of this method in video-processing systems in the future.

ACKNOWLEDGMENT

The authors thank Professor Su-Ling Yeh, Dr. Shin-Yi Liao, and Yong-Hao Yang for providing human perceptual knowledge to help develop the algorithm. They are members of Perception & Attention Laboratory, Department of Psychology, National Taiwan University, Taipei, Taiwan, R.O.C.

REFERENCES

[1] Royal Philips Electronics, "Whitepaper WOWvx BlueBox" [Online]. Available: http://www.business-sites.philips.com/global/en/gmm/images/3d/3dcontentceationproducts/downloads/BlueBox_white_paper.pdf.

[2] C. Wu, G. Er, X. Xie, T. Li, X. Cao, and Q. Dai, "A novel method for semi-automatic 2D to 3D video conversion," 3DTV Conference: *The True Vision - Capture, Transmission and Display of 3D Video*, 2008, pp. 65–68, May 28–30, 2008.

[3] H. M. Wang, C. H. Huang, and J. F. Yang, "Depth maps interpolation from existing pairs of keyframes and depth maps for 3D video generation," *in Proc. IEEE Int. Symp. Circuits Syst.*, pp. 3248–3251, May 2010.

[4] H. M. Wang, C. H. Huang, and J. F. Yang, "Block-based depth maps interpolation for efficient multiview content generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1847–1858, Dec. 2011.

[5] I. A. Ideses, L. P. Yaroslavsky, B. Fishbain, and R. Vistuch, "3D from compressed 2D video," *in Proc. SPIE: Stereoscopic Displays and Applications XIV*, vol. 6490, 64901C, 2007.

[6] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 406–417, 2006.

[7] J. Park and C. Kim, "Extracting focused object from low depth-of-field image sequences," *in Proc. SPIE Visual Communications and Image Processing*, vol. 6077, pp. 607710-1–607710-8, Jan. 2006.

[8] W. J. Tam, A. S. Yee, J. Ferreira, S. Tariq, and F. Speranza, "Stereoscopic image rendering based on depth maps created from blur and edge information," *in Proc. SPIE: Stereoscopic Displays and Applications XII*, vol. 5664, pp. 104–115, 2005.

[9] W. J. Tam, C. Vázquez, and F. Speranza, "Three-dimensional TV: A novel method for generating surrogate depth maps using color information," *in Proc. SPIE: Stereoscopic Displays and Applications XX*, vol. 7237, 2009.

[10] C. C. Cheng, C. T. Li, and L. G. Chen, "An ultra-low-cost 2D-to-3D conversion system," *International Symposium, Seminar, and Exhibition of Society for Information Display*, pp. 766–769, May 23–28, 2010.
[11] J. Zhang, Y. Yang, and Q. Dai, "A novel 2D-to-3D scheme by visual attention and occlusion analysis," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2011, pp. 1–4, May 16–18, 2011.
[12] Y. Feng, J. Ren, and J. Jiang, "Object-based 2D-to-3D video conversion for effective stereoscopic content generation in 3D-TV applications," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 500–509, June 2011.
[13] S. Battiato, S. Curti, M. La Cascia, E. Scordato, M. Tortora, "Depth Map Generation by Image Classification" *in Proc. SPIE: Electronic Imaging 2004 – Three-Dimensional Image Capture and Applications VI*, vol. 5302-13, January 2004;
[14] S. Battiato, A. Capra, S. Curti, M. La Cascia, "3D Stereoscopic Pairs by Depth-map Image Generation," *IEEE 3DPVT'04, 2nd International Symposium on 3D Data Processing Visualization & Transmission*, pp. 124-131, September 2004.
[15] C. C. Cheng, C. T. Li, P. S. Huang, T. K. Lin, Y. M. Tsai, and L. G. Chen, "A block-based 2D-to-3D conversion system with bilateral filter," *IEEE International Conference on Consumer Electronics, 2009. ICCE '09.* Digest of Technical Papers, pp. 1–2, Jan. 10–14, 2009.
[16] Y. L. Chang, J. Y. Chang, Y. M. Tsai, C. L. Lee, and L. G. Chen, "Priority depth fusion for the 2D-to-3D conversion system," *SPIE 20th Annual Symposium on Electronics Imaging*, vol. 6805, 680513, 2008.
[17] T. Iinuma, H. Murata, S. Yamashita, and K. Oyamada, "Natural stereo depth creation methodology for a real-time 2D-to-3D image conversion," *International Symposium, Seminar, and Exhibition of Society for Information Display*, vol. 43, pp. 1212–1215, May 14–19, 2000.
[18] L. T. Maloney and M. S. Landy, "A statistical framework for robust fusion of depth information," *in Proc. SPIE: Visual Communications and Image Processing IV*, vol. 1199, pp. 1154–1163, 1989.
[19] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: in defense of weak fusion." *Vision Res.*, vol. 35, pp. 389–412, 1995.
[20] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth," *in W. Epstein and S. Rogers (Eds.), Handbook of Perception and Cognition. Vol. 5: Perception of Space and Motion*, pp. 69–117, San Diego.
[21] C. C. Cheng, C. T. Li, and L. G. Chen, "A 2D-to-3D conversion system using edge information," *IEEE International Conference on Consumer Electronics. ICCE 2010,* Digest of Technical Papers, pp. 377–378, Jan. 9–13, 2010.
[22] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," Computer Vision, 1998. Sixth International Conference, pp. 839–846, Jan 4–7, 1998,
[23] DSP/IC LAB, National Taiwan University, "Demo sequences for brain-inspired framework for fusion of multiple depth cues" [Online]. Available at: http://video.ee.ntu.edu.tw/~ztdjbdy/3D_demos/.
[24] Royal Philips Electronics, "Philips 3D Solutions" [Online]. Available at: http://www.business-sites.philips.com/3dsolutions/.
[25] C. T. Li, Y. C. Lai, C. Wu, C. C. Cheng, and L. G. Chen, "A quality measurement based on object formation for 3D contents," *International Symposium, Seminar, and Exhibition of Society for Information Display* , pp. 1265–1268, May 16–20, 2011.
[26] S. L. P. Yasakethu, D. V. S. X. De Silva, W. A. C. Fernando, and A. Kondoz, "Predicting sensation of depth in 3D video," *Electronics Letters*, vol. 46, no. 12, pp. 837–839, June 10, 2010.

Chung-Te Li was born in Taipei, Taiwan, ROC in 1984. He received the B.S. degree in Electronics Engineering from National Taiwan University, Taiwan, Taiwan, R.O.C., in 2006. He is currently a Ph.D. student of Graduate Institute of Electronics Engineering from National Taiwan University. His major research interests include digital signal processing, computer vision, 3D image/video processing algorithm and architecture.

Yen-Chieh Lai was born in Taipei, Taiwan, ROC in 1988. He received the B.S. and M.S. degree in Electronics Engineering from National Taiwan University, Taiwan, Taiwan, R.O.C., in 2009, and 2011, respectively. His research interests include stereo vision and 3D signal processing.

Chien Wu was born in Taipei Taiwan in 1987. He received the B.S. degree in electrical and electronics engineering from National Tai-wan University, Taipei, Taiwan in 2009, where he is working toward the M.S. degree at Graduate Institute of Electronics Engineering. His major research interests include computer vision and image/video processing.

Sung-Fang Tsai was born in Hsinchu, Taiwan in 1983. He received the B.S. and M.S. degree in electrical and electronics engineering from National Taiwan University, Taipei, Taiwan in 2005 and 2007, where he is working toward the Ph. D. degree at the Graduate Institute of Electronics Engineering. His major research interests include high efficiency video coding, and design of 3D/FTV video system.

Tung-Chien Chen was born in Taipei, Taiwan, in 1979. He received the B.S., M.S., and Ph.D. degrees in electrical engineering, National Taiwan University, Taipei, Taiwan in 2002, 2004, and 2009. He was a visiting scholar in University of California, Santa Cruz, from 2006 to 2008. He is doing postdoctoral research in electrical engineering, National Taiwan University from 2009 till present. Before 2006, he worked on the algorithm and architecture design for multimedia applications such as video codecs and intelligent video sensors. He is currently working for the algorithm and architecture design for smart body sensors and closed-loop neuroprosthetic devices.

Shao-Yi Chien (S'99--M'04) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 1999 and 2003, respectively. During 2003 to 2004, he was a research staff in Quanta Research Institute, Tao Yuan County, Taiwan. In 2004, he joined the Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, as an Assistant Professor. Since 2008, he has been an Associate Professor. His research interests include video segmentation algorithm, intelligent video coding technology, perceptual coding technology, image processing for digital still cameras and display devices, computer graphics, and the associated VLSI and processor architectures. He has published more than 120 papers in these areas.

Dr. Chien serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology and Springer Circuits, Systems and Signal Processing (CSSP). He also served as a Guest Editor for Springer Journal of Signal Processing Systems in 2008. He also serves on the technical program committees of several conferences, such as ISCAS, A-SSCC, and VLSI-DAT.

Liang-Gee Chen (S'84–M'86–SM'94–F'01) received the B.S., M.S., and Ph.D. degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1979, 1981, and 1986, respectively. In 1988, he joined the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan. From 1993 to 1994, he was a Visiting Consultant in the DSP Research Department, AT&T Bell Labs, Murray Hill, NJ. In 1997, he was a Visiting Scholar of the Department of Electrical Engineering, University of Washington, Seattle. Currently, he is Professor with National Taiwan University. His current research interests are DSP architecture design, video processor design, and video coding systems.

Dr. Chen has served as an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology since 1996, as Associate Editor of the IEEE Transactions on VLSI Systems since 1999, and as Associate Editor of IEEE Transactions Circuits and Systems II since 2000. He has been the Associate Editor of the Journal of Circuits, Systems, and Signal Processing since 1999, and a Guest Editor for the Journal of Video Signal Processing Systems. He is also the Associate Editor of the Proceedings of the IEEE. He was the General Chairman of the Seventh VLSI Design/CAD Symposium in 1995 and of the 1999 IEEE Workshop on Signal Processing Systems: Design and Implementation. He is the Past-Chair of Taipei Chapter of IEEE Circuits and Systems (CAS) Society and is a member of the IEEE CAS Technical Committee of VLSI Systems and Applications, the Technical Committee of Visual Signal Processing and Communications, and the IEEE Signal Processing Technical Committee of Design and Implementation of SP Systems. He is the Chair-Elect of the IEEE CAS Technical Committee on Multimedia Systems and Applications. From 2001 to 2002, he served as a Distinguished Lecturer of the IEEE CAS Society.