DEPTH MAP ESTIMATION IN DIBR STEREOSCOPIC 3D VIDEOS USING A COMBINATION OF MONOCULAR CUES

Mohammed Aabed, Dogancan Temel, Mashhour Solh and Ghassan AlRegib

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA, 30332-0250 USA
{maabed, dtemel3, msolh, alregib}@gatech.edu

***Abstract*—We propose a method to reconstruct the depth map from multiple estimated depth maps relying on monocular cues. Based on extracted depth cues from luminance, chrominance, motion and texture, we obtain an optimal depth estimation by analytically deriving the best combinations. We first analyze a ground truth depth map to extract a set of depth cues. Then, using these depth cues, we process the colored reference video to reconstruct the depth map. We tested this approach on different video sequences with different monocular properties. The results show that the extracted depth maps generate a 3D video with quality close to the video rendered using the ground truth depth map. We report subjective and objective results using 3VQM.**

***Index Terms*—Monocular cues, sensor fusion, depth map, 3D imaging, perceptual quality**

I. Description

A. Streaming 3D Video with Implicit Depth

A stereoscopic 3D video can be generated using a single video sequence and its corresponding depth map sequence using a technique know as depth image based rendering (DIBR). In this work, we show that transmitting the depth map is unnecessary in DIBR-based stereoscopic 3D display. Instead, we will extract depth cues from the depth map at the sender and send them as a side information with the reference colored video. At the receiver, the depth map is reconstructed using the depth cues and information extracted from monocular cues. In addition to the bandwidth saving, taking the depth map out of the transmitted data eliminates an additional source of error. This process corresponds to blocks (A) and (B) in the system depicted in Figure 1 and we focus in this paper on these two blocks. Particularly, the discussion herein is concerned with the 3D content video coding in block (A), and generating virtual views in block (B).

B. Depth from Monocular Cues in the HVS

The Human Visual system (HVS) exploits a set of visual depth cues to perceive 3D scenes. These depth cues can be classified into two classes: binocular and monocular cues. Binocular cues are the disparities that exist between the two views seen by both eyes of a particular scene. HVS extracts the depth information by comparing two views of a particular scene. It is believed that the human mind creates the illusion of 3D by exploiting horizontal disparity between the scenes. In today's technologies the 3D visual simulation is triggered by projecting two views with a slight horizontal disparity on the left and right eyes. The HVS can also extract depth using a single eye. The depth information that can be extracted from a single view is known as monocular cue. Monocular depth cues are numerous and the following is a list of the important cues within the context of our discussion [1].

- *Motion Cue:* The HVS can distinguish depth from relative motion of objects since near objects move faster across the retina than far objects.
- *Texture Gradient Cue:* Texture is also an important depth cue. As the surface gets farther away from the observer the texture gets finer and appears smoother. As a result, depth can be extracted from an image by evaluating the texture attributes.
- *Color and Intensity Cues:* Depth can also be estimated from luminance and color variations in the scene. By a phenomenon known as atmospheric scattering, scenes in the foreground tend to have higher contrast as compared to scenes in the background. In addition, brighter or higher luminance values are often closer to the foreground. Color cues can also be learned heuristically by prior knowledge such as color of the sky, mountain, land and others.

The goal of the depth estimation from monocular cues is to convert depth cues contained in video sequences into actual depth values of a captured scene. The extraction of depth from monocular depth cues for 2D-to-3D conversion is a complex challenge, one that has attracted a lot of attention in the last decade [2]. Most of the work on 2D-to-3D conversion from monocular cues has focused on extracting depth from a single depth cue by analyzing the 2D scene. On the contrary, the *HVS* uses a combination of information from individual monocular cues and prior cognitive knowledge to get a single estimate of the depth map of the scene. Therefore, the challenge is to find an optimal way to combine the information that we obtain from individual monocular cues. In order to do

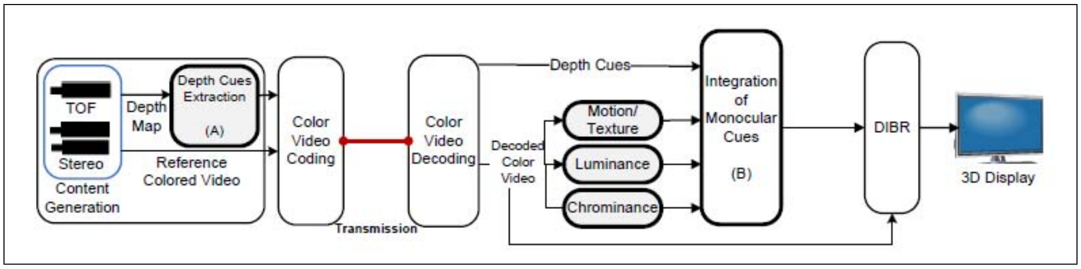


Fig. 1. A block diagram demonstrating the proposed framework for depth maps extraction. The proposed algorithm has two components labeled as (A) and (B), respectively.

that, we should return to the basics and understand how this combination mechanism works in the *HVS*. In this work, we propose an approach that remotely mirrors the *HVS* depth estimation through a combination of several depth cues and prior knowledge (cues) extracted from the depth map. This approach will be described in the next section.

II. Proposed Approach

We propose a depth extraction approach based on intensity, motion and texture. First, we begin by processing a given depth map to extract a set of depth *cues*. These cues are transmitted as a side information with the reference view. Then, they are used alongside motion and texture to reconstruct the depth map at the receiver side. We independently estimate texture-based and motion-based depth maps. These depth maps are then combined to give a single depth map. It can be considered as an example of sensor fusion problem for two correlated cues. Luminance and chrominance channel contents are also used with the depth cues to reconstruct the depth map. This is achieved by matching the statistical feature of the depth map and the intensity component. At the final stage, all of the individual monocular cues are combined to obtain a single depth estimate that can be used for rendering the 3D view.

A. Depth-Map Cues Extraction

The depth-map cues extraction process involves calculating the number of depth planes (N) and the corresponding depth value for each plane in the depth map D_n. The exact number of planes in an image varies, and for natural scene images, the number of planes is finite. Based on the information collected from stereo images and videos from different sources, we have found that the average number of planes in an image is less than or equal to four $(N <= 4)$. In principle, our initial approach was to quantize the depth levels to come up with depth cues that can be used for depth map reconstruction. The first step is to construct the histogram of the depth intensity values. Then assuming $N = 4$, we quantize the histogram by extracting local maximas and minimas in the histogram to evaluate D_n. The histogram is then scanned from 0 to 255 and each D_n value is chosen as the median of the range constituting the two minimas and a single maxima.

B. Luminance/Chrominance-Based Depth Extraction

At the receiver, the parameters are extracted as it is explained in Section II-A and these parameters are used with the luminance channel to reconstruct the depth map. The process is as follows:

- For each plane N and its corresponding depth values D_n, a corresponding range of depth values is extracted from the luminance component of the colored video. Thus, a set of planes $(\mathbf{L}_n)$ are extracted from the luminance channel. This can be done by following similar steps to the one used in depth cues extraction.
- A luminance to depth mapping is then performed by replacing the high luminance planes with the near depth, the low luminance pixels with far depth and middle luminance planes with the corresponding intermediate depth values. The output of this mapping is the estimated depth-map, $\overline{Z_Y}$.
- The estimation of the depth map is done by replacing luminance values between 0 and L_0 with D_0, values between L_0 and L_1 with D_1, values between L_1 and L_2 with D_2 and values between L_2 and L_3 with D_3.

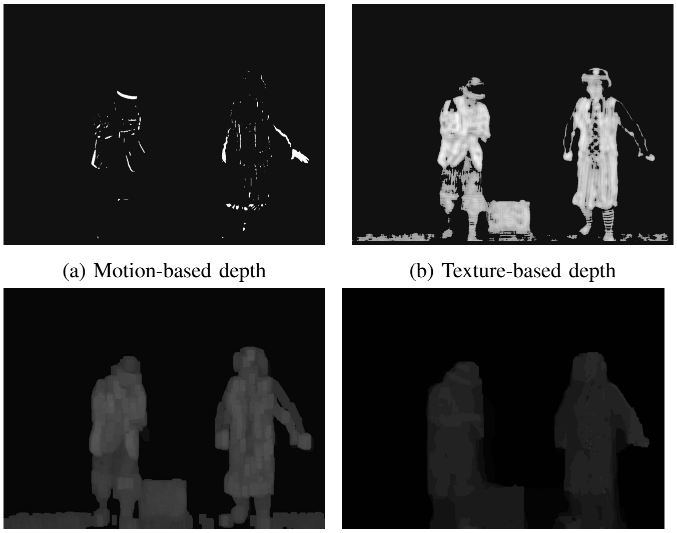
Chrominance-based depth extraction is performed in the same way as *luminance* and we obtain two depth estimates for C_R and C_B channels.

C. Texture Based Depth Extraction

Texture structure in an image can provide information about the depth of a scene. The authors in [3] used texture information for the depth-matching experiments, and concluded that it provides information about the depth of a scene by itself. In this work, we first analyze the colored reference frame for texture content. The image is then segmented into structured, smooth and highly textured regions. A texture randomness index is then assigned to non overlapping macroblocks, the object closer to the camera would have a higher texture value compared to the same object at the background. Therefore, we can estimate the relative depth. Next, we can map the texture value to the absolute depth by using the extracted depth cues.

D. Motion Based Depth Extraction

Extracting depth from motion starts by evaluating the scene for optical flow and then near objects are the ones moving faster than the farther objects. Block-based motion estimation methods such as exhaustive search and three step search do not provide an accurate estimation of the depth because of the extensive amount of the noise. For the purpose of this work, the motion information can be calculated by taking the difference of the pixel values between the luminance channel



(a) Motion-based depth

(b) Texture-based depth

(c) Combined depth

(d) Ground truth depth

Fig. 2. Estimated depth maps from texture and motion for the `Pantomime` sequence.

of the reference colored frame and the i^{th} subsequent frame. Extracted depth cues are used to map linearly the relative motion index to absolute depth estimation.

E. Combining Motion and Texture Based Depths

The basic idea behind this method is to combine different depth estimates pixel by pixel in a search window. If both motion and texture have zero values, we scan the estimates separately in a search window. For every pixel, we will look at the rest of the pixels in the search window and assign a new value to the central pixel accordingly. Then, scaled versions of the estimates will be summed up for the updated value of the central pixel. If one of the depth estimates is non-zero at one pixel and zero at the other pixel, we directly assign the non-zero value to the final estimate. Otherwise, we weight the estimates and sum them up. From our experiments, the combination of motion and texture outperforms the separate mapping. Figure 2 shows these estimated depth maps. Motion-based depth estimation detects changes in the scene corresponding to the motion of foreground objects. For example, the hands and the arms are well represented in the motion-based depth map (Figure 2 (a)). However, texture-based depth estimation fails to account for mild-textured areas of foreground objects, such as the regions around the hand in Figure 2(b). When we combine texture and motion, individual estimates complement each other and we get a more convincing depth map as shown in Figure 2(c). Interestingly, the ground truth depth map does not differentiate between the wall or the background and the floor as shown in Figure 2(d). It is important to notice that the ground truth is not the perfect depth map and thus it might not be error-free as in this case. For this particular case, our combined depth estimate differentiates between the background and the floor as illustrated in Figure 2(c).

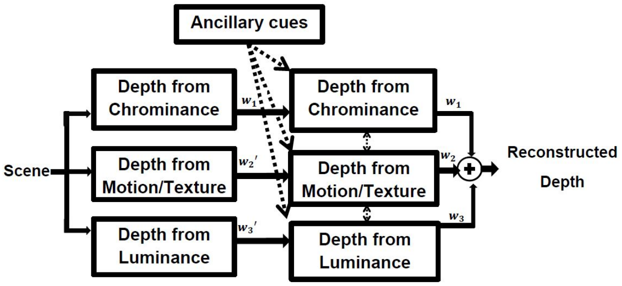


Fig. 3. Modified Weak fusion model for integration of monocular cues.

F. Combining Monocular Cues

The simplest approach to model the combination procedure is the *Weak Fusion (WF)* model which assumes that depth cues are isolated from each other and the final estimate is obtained by simply averaging the individual estimates. On the other hand, some researchers modeled the combination as a *Strong Fusion (SF)* problem which is based on interactive and holistic processing [4]. According to our observations, *WF* model oversimplifies the problem and *SF* model makes the situation more problematic to be a practical approach. We need to combine *WF* and *SF* in such a way that we will benefit from the the advantages of both of the methods. *Modified Weak Fusion (MWF)* model is the solution to our problem which enables processing monocular cues in an isolated way and uses the information coming from other cues as ancillary cues. These ancillary cues are used to fill the missing parameters to convert the relative depth estimates to the absolute ones. In our proposed work, we implemented *MWF* model as in Figure 3. Initial depth maps are estimated from chrominance, luminance, motion, and texture. Extracted cues are used to promote the monocular cues so that we obtain the absolute depths. As a result of the promotion process, we will get the estimates represented by w_1, w_2 and w_3. Now, we need to calculate the reliability metrics to combine the depth estimates. We focused on combining these cues in an optimal way with the reliability metrics that are based on *SO*, *TO* and *TI*. *SO* is the standard deviation of depth map errors, *TO* is the standard deviation of two depth map errors in time domain and *TI* is the standard deviation of two depth values in two time instances. These metrics are pooled together to assess the quality of the rendered view so that we can converge to the optimal combination by reassigning the weights and the pooled metric is called as *3VQM* [5].

III. Preliminary results

In all our experiments, we wrap the virtual view by using the same reference frame and depth map estimates. At first we render the view using DIBR [6] and then we apply HHF [7] to eliminate the holes in the rendered view. As a reference, *3VQM* and *PSNR* values for the rendered views from monocular cues and ground-truth are provided in Table I. We combine luminance and motion-texture to obtain the estimate and then we use chrominance instead of motion-texture to observe the

Sequence	3VQM					PSNR (dB)				
	M-T	Y	C_R	C_B	Ground truth	M-T	Y	C_R	C_B	Ground truth
`Balloons`	4.24	2.16	0.65	2.43	4.52	24.23	27.33	25.16	26.28	24.00
`Cafe`	4.69	3.03	3.52	1.97	4.54	28.75	24.88	25.56	28.04	28.64
`Champ. Tower`	4.07	0.41	2.70	3.81	4.67	27.34	21.64	21.97	27.77	23.95
`Kendo`	4.21	2.90	1.64	3.88	4.41	27.46	28.99	24.58	31.49	25.76
`Love Birds`	4.84	3.98	3.04	0.90	4.75	28.32	23.09	25.68	24.15	34.94
`Pantomime`	4.44	0.17	3.01	3.27	4.73	30.67	22.17	21.32	23.80	22.87

TABLE I
3VQM AND PSNR VALUES FOR SIX DIFFERENT VIDEOS.

Sequence	Metric	M-T	Y	C_R	C_B	Ground truth
Balloons	TI	0.99	0.96	0.88	0.96	1.00
Res. 1024 × 768	TO	0.99	0.97	0.93	0.98	1.00
Frames: 500	SO	0.99	0.96	0.91	0.97	1.00
Cafe	TI	0.95	0.98	0.98	0.95	0.99
Res. 1920 × 1080	TO	0.99	0.98	0.98	0.96	1.00
Frames: 300	SO	0.95	0.98	0.99	0.96	1.00
Champ. Tower	TI	0.99	0.87	0.97	0.99	1.00
Res. 1280 × 960	TO	1.00	0.90	0.97	0.99	1.00
Frames: 300	SO	0.99	0.90	0.97	0.99	1.00
Kendo	TI	0.99	0.97	0.93	0.98	0.99
Res. 1024 × 768	TO	0.99	0.98	0.96	0.99	0.99
Frames: 400	SO	0.99	0.98	0.96	0.99	1.00
Love Birds	TI	0.96	0.99	0.98	0.92	1.00
Res. 1024 × 768	TO	0.99	0.99	0.98	0.93	1.00
Frames: 300	SO	0.95	0.99	0.98	0.92	1.00
Pan to Mime	TI	0.99	0.82	0.97	0.98	1.00
Res. 1280 × 960	TO	1.00	0.87	0.98	0.98	1.00
Frames: 300	SO	0.99	0.88	0.98	0.98	1.00

TABLE II
TEMPORAL INCONSISTENCY (TI), TEMPORAL OUTLIERS (TO), AND SPATIAL OUTLIERS (SO) FOR THE DIBR SYNTHESIZED FOR SIX DIFFERENT VIDEO SEQUENCE FROM 3D MOBILE PROJECT.

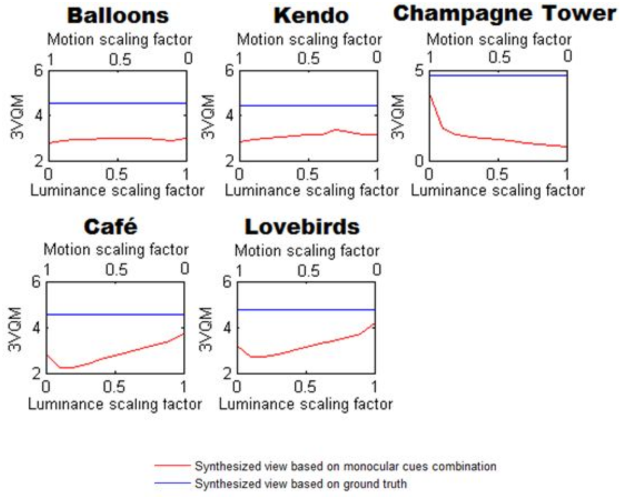


Fig. 4. Luminance-Motion/Texture Combination

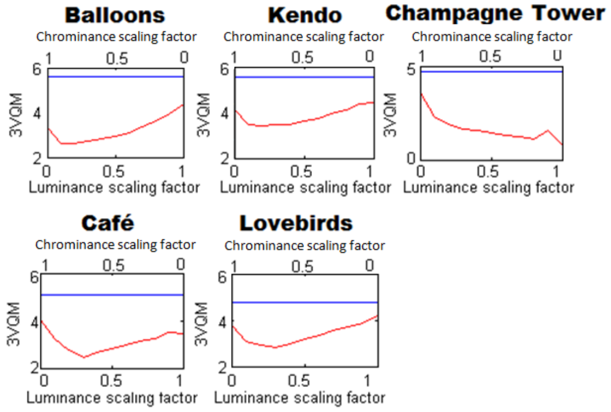


Fig. 5. Luminance-Chrominance Combination

difference between these two cases. We vary the weights of the estimates so that we can observe the changes in the objective and subjective quality assessment results. *3VQM* results for luminance and motion/texture combination are given in Figure 4 and the results for luminance and chrominance combination are shown in Figure 5. Weight of the motion-texture combination goes from *1.0* to *0.0* while luminance weight is simply *1.0* minus the other weight. *"GT"* corresponds to the rendered view from ground truth and it is shown with the *blue (thick) line* on the graph. *Red curves* illustrate the *3VQM* values corresponding to different combination scenarios under varying weights. For *Champagne tower*, motion-texture based estimate is more accurate than the luminance estimate as it can be understood from the individual *3VQM* values. Therefore, we observe a significant decrease in the quality as we go from left to the right. For *Cafe* and *Lovebirds*, luminance values are more reliable and we see an increase as we go from left to right. The slight decrease can be due to the limited number of images in the sample set. For *Balloons* and *Kendo*, luminance and motion-texture estimates are comparable so we can only observe low level fluctuations. In case of luminance and chrominance combination, we see a similar decrease for *Champagne tower*. Other sequences tend to have a decrease and then an increase in the objective quality assessment results and final values are close to initial ones.

Table II illustrates three metrics to measure the visual discomfort of the rendered 3D video by combining three quality criteria: Spatial Outliers (SO), Temporal outlier (TO), and Temporal Inconsistency (TI). For each of the three indices, the individual pixel values were calculated as proposed in [5]. SO measures the visible distortions due to spacial artifacts. TO quantifies the errors due to depth map noise and hole filling. TI is a measure of visual discomfort in the form of fast changing disparities caused by errors in stereo matching, hole filling algorithms and depth compression.

The results show that quality of the rendered videos resulting from the combination of texture and motion tends to be more temporally and spatially consistent than the cues from luminance and chrominance. However, these variations are dependent on the nature of the scenes in these videos. The subjective evaluation of the videos as well as the objective results presented here have shown that the optimal combination will vary from one sequence to another. These variations can be predicted by evaluating the scene itself for motion, texture, luminance and chrominance.

The *Cafe* sequence, for example, is rich in colors, texture, and luminance but motion is restricted to one object. In this case, an optimal combination of cues would assign more weight for cues from color, texture and luminance less for motion cues. The same conclusions about the *Cafe* sequence can be derived by looking at the figures and the tables.

The *Champagne* sequence scene can be described as an indoor, studio light, high detailed, and complex object motion. The depth cues for luminance in this scene is inaccurate because studio lights are not natural, as a result atmospheric scattering of light is manipulated. An optimal combination for the *Champagne* sequence would stress motion, texture and chrominance cues because the scene has complex object motion, and detailed textures. The same can also be concluded by analyzing the *Pantomime* video sequence which is very similar in nature to the *Champagne* sequence. It is an indoor, studio light, medium complex object motion and medium details. Table I shows that both the *Champagne* and *Pantomime* sequences have similar performance for individual cues.

Looking at the results in the Table I, Figure 4, and Figure 5 we find that the *Lovebirds* sequence has a more accurate depth estimation from luminance as compared to the rest of the sequences. The *Lovebirds* scene can be described as outdoor, natural light, simple object motion, and highly detailed. The increased accuracy of depth from luminance is a result of atmospheric scattering of natural light. An optimal combination in the case of *Lovebirds* would emphasize cues form motion, texture, and luminance.

IV. Conclusion

Overall, the rendered views from the reconstructed depth maps are comparable to the rendered view from the ground truth depth map. However, in order to guarantee a permanent high-quality standard, we have to perform a robust linear combination of these individual cues. The most important conclusion that we can get from the previous observations is the varying reliability of the monocular cues. Therefore, we need to take into account the interactions between the monocular cues to perform the linear combination adaptively.

References

[1] 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, june 2011.

[2] Q. Wei, “Converting 2d to 3d: A survey,” *Delft University of Technology,The Netherlands, Project Report*, Dec 2005.

[3] J. et al., “Optimal Integration of texture and motion cues to depth ,” *Vision Research*, vol. 39, pp. 4062–4075, 1999.

[4] 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 Research*, vol. 35, pp. 389–412, 1995.

[5] M. Solh, G. AlRegib, and J. M. Bauza, “3VQM: A vision-based quality measure for DIBR-based 3D videos,” in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, July 2011, pp. 1 –6.

[6] C. Fehn, “Depth-image-based Rendering (DIBR), Compression, And Transmission For A New Approach On 3DTV,” *Proc. of SPIE*, vol. 5291, pp. 93–104, 2004.

[7] M. Solh and G. AlRegib, “Hierarchical Hole-Filling (HHF): Depth Image Based Rendering without Depth Map Filtering for 3D-TV,” in *IEEE MMSP’10*, Saint-Malo, France, October 4-6 2010.