VISUAL PERTINENT 2D-TO-3D VIDEO CONVERSION BY MULTI-CUE FUSION

Zhebin Zhang[*,+] *Yizhou Wang*[+] *Tingting Jiang*[+] *Wen Gao*[+]*, Fellow, IEEE*

[*]Key Lab. of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[*]Graduate School, Chinese Academy of Sciences, Beijing, 100039, China
[+]National Engineering Lab. for Video Technology,
[+]Key Lab. of Machine Perception (MoE), Sch'l of EECS, Peking University, Beijing, 100871, China
zbzhang@jdl.ac.cn, {yizhou.wang, ttjiang, wgao}@pku.edu.cn

ABSTRACT

We describe an approach to2D-to-3D video conversion for the stereoscopic display. Targeting the problem of synthesizing the frames of a virtual 'right view' from the original monocular 2D video, we generate the stereoscopic video in steps as following. (1) A 2.5D depth map is first estimated in a multi-cue fusion manner by leveraging motion cues and photometric cues in video frames with a depth prior of spatial and temporal smoothness. (2)The depth map is converted to a disparity map with considering both the displaying device size and human's stereoscopic visual perception constraints. (3)We fix the original 2D frames as the 'left view' ones, and warp them to "virtually viewed" right ones according to the predicted disparity value. The main contribution of this method is to combine motion and photometric cues together to estimate depth map. In the experiments, we apply our method to converting several movie clips of well-known films into stereoscopic 3D video and get good results[1].

Index Terms— 2D-to-3D, Stereoscopic, Depth, Disparity, Virtual View Synthesis

1. INTRODUCTION

2D-to-3D video conversion is necessary for 3D video industry for two main reasons. (1) Indespite of many stereoscopic 3D movies produced in the recent few years, the amount isnot enough, especially for 3DTV industry. For instance, the Sky TV's 3D Channel could not start its officially broadcasting until it is able to supply more than 6 hours 3D programs per day. As many broadcasting companies plan to start 3DTV Channel in the near future, the desire for 2D-to-3D conversion technology is urgent. (2) 3D reproduction of the conventional well-know 2D movies or TV programs is appealing both to the potential audience and the producers, and also this will gives these films renascence after years of the premiere.

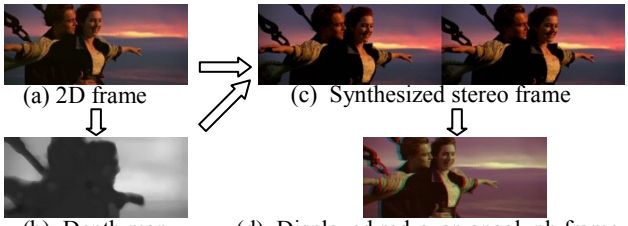

Figure1: Converting *Titanic* into stereoscopic 3D video

Lots of researchers study on the 2D-to-3D video conversion problem and some companies also have released several conversion softwareand the chips integrated in 3DTV. These methods can be classified into two categories. One class is fully automatic conversion [2][3]. Some commercial software, such as DDD's TriDef 3D player and Samsung's 3DTV, can generate stereoscopic views from monocular videos in realtime. However, these products are not able to robustly estimate pixel disparities. The other category of methods involves user interaction during stereoscopic conversion [4]. At key frames, they manually initialize the depth values of the objects in the scene, and propagate the depth information to other frames. The IMAX also develops a manually labeling system [22] and achieves better visual impression than those fully automatic ones.

Converting a 2D video into stereoscopic one usually takes two steps, disparity map estimation for each frame and virtual view synthesis. The major challenge of 2D-to-3D video conversion lies in the disparity estimation of monocular videos, which is closely related to the scene depth estimation [1]. Since the disparity can't be directly extracted from the monocular image or video, the problem is cast into first estimating scene's depth map and then converting it to disparity according to their proportional relation. Precisely estimation will produce the stereo image pair without visual artifacts. Researchers take advantage of various cues and context to estimate depth maps from single-viewed images/video, such as photometric measurement [14][15][17][18],structure (geometry)[6][16], motion [2][3], and appearance [4][7][8],etc.

[1]All the videos or movie trailers presented in this paper are collected from internet and only used for research purposes.

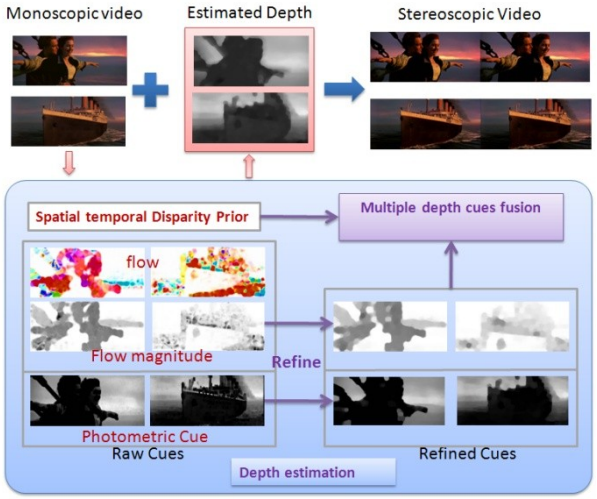

Figure 3: The flowchart of theproposed conversion system.

Photometric measurement. Objects in the image may not be all in focus [17]. [18] useswavelet analysis and edge defocus estimation to obtain a relative depth. Besides, scene atmospheric light also can provide a depth cue. Atmospheric radiance images are usually degraded in the atmosphere, especially for the outdoor scenes.The irradiance received by the camera from objects in the sceneis attenuated along the line of sight. Such phenomena can be used a cue for scene understanding. For instance, He *et al* [14] use a dark channel prior to remove the haze and also produce a depth map of the scene.

Structure. Structures in the scene can be leveraged to estimate the depth. For example, parallel linesalong the principal coordinate direction are always the main cues for single view 3D reconstruction, by estimating the vanishing points/line from the lines [5][6][16]. However, these structures always appear in limited scenes.

Motion. Motion cue is always used in 2D-to-3D video conversion. Under the condition of constrained camera motion and assuming that scenes are static, there are two ways to estimate disparity maps, (i) using Structure from Motion (SfM) [3][5], and (ii) leveraging motion parallax [2]. However, in real scenario such as movies and TV programs, the constrained camera motion condition and static scene assumption are often violated, which leads to the failure of applying the two methods in disparity estimation.

Appearance. Hoiem [7] casts the 3D structured outdoor scene reconstruction from a single image as a multi-label classification problem by using categories of appearance feature of superpixel. Saxena et al. [8] use MRF to model the unstructured scene and directly learn the relation between the planar 3D structuresand the super-pixel's texture and color features. Both of their works need training data to learn the model, which makes it difficult to apply their methods to large scale movie/TV programs with significant scene appearance variance.

Although there is great progress in depth estimation from a single view image or monocular video, recovering

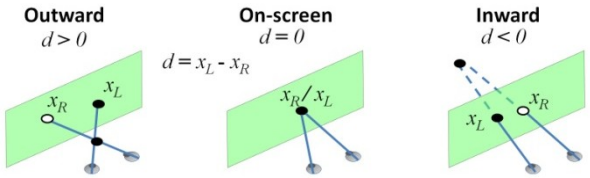

Figure 2: Illustration of eyes' visual perception when human watch the stereoscopic 3D video. Impertinent disposal of the disparity of objects in a scene causes visual artifacts.

the depth map of the unconstrained scene is still a challenging problem for 2D-to-3D video conversion. Scenes in video always change, especially for the movie and TV programs. Using any single category of depth cue is not suitable and adaptive for the 2D-to-3D video conversion. Multiple cues should be combined together to estimate the depth.In this paper, considering the difficultyof extending the learning based methods to handle different scene categories in the conventional 2D movies, we propose a conversion method based on directly predicting depth map by fusing motion and photometric cues with a spatial temporal depth prior. The flowchart of the proposed method is shown in Figure 2.

2. VISUAL PERTINENT 3D VIDEO

Before describing the details of our method, we should know that converting 2D videos into stereoscopic 3D ones requires that

1. The generated stereo view image pair should be similar to the original frame both in appearance and structure without any apparent artifacts,
2. The generated 3D video should look comfortable without causing any visual fatigue,
3. The generated video also should look with sort of stereoscopic visual effects (e.g., the objects are pop out from the screen), so that the conversion makes sense.

Hence, taking the biological stereoscopic visual perception into account is necessary for a 2d-to-3d conversion system. As shown in Figure 3, the displayed disparity perceived by eyes produces stereoscopic impression. The sign of the disparity decides the corresponding pixel would be perceived as inward (positive), outward (negative) or on (zero) the screen. Its absolute value controls the perceived distance from the pixel to the screen. The displayed disparity value is required to be within the range of pertinent visual perception, which should not exceed the maximum value that human eyes can comfortably perceive,meanwhile to be able produce the in/out screen effects.If the absolute value is too large, the sight lines of two eyes will not converge and the artifacts of ghost images are produced.Absolute valuesof displayed disparities arerelated to both the image disparity (pixels) and the screen size (inches). Generally, an absolute disparity should not exceed 3% of the image size for the screens larger than 77 inches, and 5% for the screen smaller than 77 inches.

3. 2D-TO-3D VIDEO CONVERSION

As shown in Figure 2, the 2D-to-3D video conversion method is performed in steps as follows:

(i) A 2.5D depth map$dp(I)$ is first estimated in a multi-cue fusion manner by leveraging motion and photometric cues in video frames with a spatial temporal depth smoothness prior.

(ii) Depth map $dp(I)$ is converted to a disparity map$d(I)$ with considering both the displaying device size and human's stereoscopic visual perception constraints. For every pixel x in a image I, its disparity value is calculated as

$$d(x) = s \cdot w_i \left(\frac{dp(x) - dp_{min}(I)}{dp_{max}(I) - dp_{min}(I)} - f \right), \quad (1)$$

wherew_i is the image width,$dp_{max}(I)$ is the maximum of depth values. As we described in Section 2, the absolute disparity value should be restricted according to the screen size. Here, s is the factor that controls the maximum absolute disparity value, and $f(0 \le f < 1)$ is a parameter that shifts the disparity to negative value.

Directly using any local cues to predict the depth is always unstable and make the synthesized frames floating. Watching such frames will cause visual fatigue, and viewers will feel dizzy. To avoid such a problem, a temporal smoothness prior is used to smooth the estimated disparity value, which is calculated as

$$d_t(x) = \delta d_{t-1}(x) + (1 - \delta) d_t(x). \quad (2)$$

In Equation (2), δ is the similarity of neighboring frames and can be evaluated by the intersection of color histograms of the two frames.

(iii)We fix the original 2D frames as the 'left view' ones, and warp them to the 'virtually viewed' right ones according to the predicted disparity value,

$$I_r(x + d(x)) = I_l(x). \quad (3)$$

3.1. Depth Estimation by Multi-cue Fusion

Two cues are used to estimate the depth map$dp(I)$, dark-channel prior and motion magnitude. The first one indicates the distance from objects to the camera due to variance of atmospheric light. The second one is always used to predict the pseudo-disparity value which is inversely proportional to the depth value. We combine the two cues together to estimate depth map of the scene in a video frame.

As described in [14], the irradiance in a scene is attenuated along the sight. Here we also use it as one of cues to estimate the depth. Different from [14], which needs about 20 seconds to solve the global optimal depth map for a 600×400 image, we do not pursue a global optimal depth map. The depth map is solved by

$$dp_a(x) = -\alpha \, ln\big(t(x)\big), \quad (4)$$

whereα is a the scattering coefficient of the atmosphere and$t(x)$ is the medium transmission, which is estimated by directly using dark-channel prior

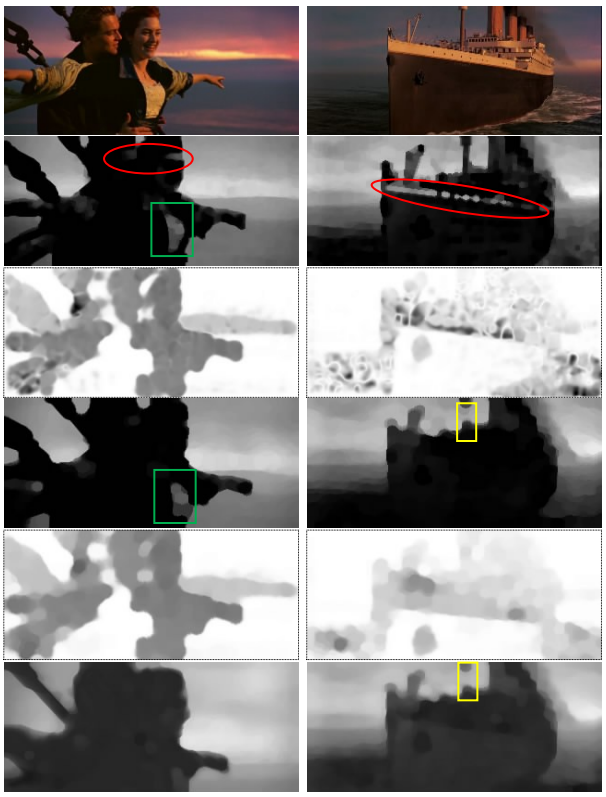

Figure 4: Estimated depth maps and their refinement. (a) Original frames. (b) Depth maps by using dark-channel. (c) Pseudo-depth maps estimated by motion magnitude. (d) and (e) Refined maps by using morphological opening and closing operators. (f) Final depth maps after fusing (d) and (e).

$$t(x) = 1 - \omega \, min_c \left(min_{y \in \Omega(x)} \left(\frac{I^c(y)}{A^c} \right) \right). \quad (5)$$

$\omega$is a constant parameter (0<ω<1),y is a pixel in a local patch centered on x, $I^c(y)$is the intensity value in the color channel c (RGB), and A^c is the atmospheric light (please refer to [14] for the details of estimatingA^c). Fig. 4(b) shows examples of the predicted d_a.

Motion is also a cue for depth estimation [2]. Although it can't provide exact and reliable depth value, motion can be used to estimate a pseudo-depth[23].Depth value is calculated proportionally to optical flowmagnitude, $dp_m(x) = \beta m(x)$, where β is a normalized factor.

Finally, the multi-cue fusion depth is

$$dp(x) = 0.5 \times (dp_a(x) + dp_m(x)). \quad (6)$$

The dark-channel prior is invalidated for estimating depth when regions with gray color appear near the camera or local light variances appear (the red eclipse regions in Fig.4 (a) and (b)).[24] shows that morphological filters can efficiently remove the tiny noisy region when segmenting images. Here we can also use it to refine the depth map. The depth values in neighboring positions should be locally smooth. Given such a spatial prior, we perform morphological opening and closing operations on the depth map.

Figure 5 (best viewed in color): The synthesized stereoscopic frames (shown in red-cyan anaglyph) of 3 film clips. Top row: *Titanic*. Middle row: *Harry Potter*. Bottom row: *Inception*.

4. EXPERIMENT RESUALTS

In the experiment, we use the proposed method to convert clips of several famous movies into stereoscopic 3D ones. Figure 5 shows the red-cyan anaglyphs of the synthesized stereoscopic frames.

Figure 4 shows the depth prediction results for *Titanic* frames (a) by directly using dark-channel prior (b) and motion cue (c), the results after morphological refinement ((d) and (e)) and by using the proposed multi-cue fusion method (e).We find that morphological operations are able to remove the tiny noisy depth values but can't handle the invalidation problem caused by the large gray regions on the near-camera objects (green rectangles in Fig.4 (b) and (d)), whereas the multi-cue fusion manner achieve better results (f).From Fig. 4(f), we can see that the invalid depth values in green rectangles of Fig. 4(b) and (d) can be replaced by more reasonable ones.

5. CONCLUSION

The proposed multi-cue fusion method combines motion and dark-channel prior to estimate the depth map. Experiments show that this method gets better results than directly using any one of them. Several generated 3D videos show the feasibility of our method. The morphological operation improves the depth map by removing tiny noisy regions, but the thin objects are also blurred or removed (yellow rectangle regions in Fig. 4 (d) and (f)), which need to be improved in our future work.

6. ACKNOWLEDGEMENT

We'd like to thank for the support from research grants NSFC-60872077 and Doctoral Fund of Ministry of Education of China KEJ200900029.

7. REFERENCES

[1] 3D4YOU: 3D4You Project, WP2: Requirements on post-production and formats conversion, 2008.

[2] Kim, D., Min, D., Sohn, K.: A stereoscopic video generation method using stereoscopic display characterization and motion analysis, IEEE T. BROADCASTING , 2008.

[3] Rotem, E., Wolowelsky, K., Pelz, D.: Automatic video to stereoscopic video conversion, SPIE2005.

[4] Guttmann, M., Wolf, L., Cohen-Or, D.: Semi-automatic stereo extraction from video footage, ICCV2009.

[5] Hartley, R.I., Zisserman, A.: Multiple view geometry in computer vision (2000) Cambridge University Press.

[6] Criminisi, A., Reid, I., Zisserman, A.: Single view metrology (2000) IJCV.

[7] Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop on scene interpretation (2008) CVPR.

[8] Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3-D scene structure from a single still image (2008) PAMI.

[9] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense twoframe stereo correspondence algorithm (2002) IJCV.

[10] Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation (2004) IJCV.

[11] Burges, C.: A tutorial on support vector machines for pattern recognition (1998) Knowledge Discovery and Data Mining.

[12] Chang, C.C., Chih-Jen: A library for support vector machines (2000) http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[13] Scharstein, D., Szeliski, R.: Middlery stereo benchmark (2005) http://vision.middlebury.edu/stereo/eval/.

[14] K. He, J. Sun, X. Tang, Single Image Haze Removal Using Dark Channel Prior, CVPR2009.

[15] G. Guo, N. Zhang, L. Huo, Wen Gao, 2D to 3D Convertion Based on Edge Defocus and Segmentation, ICASSP 2008.

[16] G. Guo, L. Liu, Z. Zhang, Y. Wang, W.Gao, An Interactive Method For Curve Extraction, ICIP2010.

[17] A. P. Pentland, "A New Sense for Depth of Field," IEEE Trans. PAMI, vol. 9, pp. 523-531, 1987.

[18] S. A. Valencia, R. M. Rodríguez-Dagnino, Synthesizing Stereo 3D Views from Focus Cues in Monoscopic 2D images, Proc. SPIE, vol. 5006, pp. 377-388, 2003.

[19] Z. Zhang, Y. Wang, T. Jiang, W.Gao, Stereoscopic Learning for Disparity Estimation, accepted by ISCAS2011.

[20] Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels, CVPR2010.

[21] Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image, ICCV2005.

[22] IMAX, Harry potter and the half-blood prince: An Imax 3D experience.http://www.imax.com/movie/HarryPotterAndTheHalfBloodPrince3D/synopsis.

[23] Xuming He, Alan Yuille, Occlusion Boundary Detection using Pseudo-Depth, ECCV2010.

[24] A. Criminisi, T. Sharp, C. Rrother, Geodesic Image and Video Editing. Transaction on Graphics, 2010.