3D-TV CONTENT GENERATION: 2D-TO-3D CONVERSION

Wa James Tam, Liang Zhang

Communications Research Centre Canada
3701 Carling Avenue
Ottawa, Ontario, Canada, K2H 8S2

ABSTRACT

The next major advancement in television is expected to be stereoscopic three-dimensional television (3D-TV). A successful roll-out of 3D-TV will require a backward-compatible transmission and distribution system, inexpensive 3D displays that are equal or superior to high-definition television (HDTV), and an adequate supply of high-quality 3D program material. With respect to the last factor, we reckon that the conversion of 2D material to stereoscopic 3D could play an important role. In this paper we provide (a) an overview of the fundamental principle underlying 2D-to-3D conversion techniques, (b) a cursory look at a number of approaches for depth extraction using a single image, and (c) a highlight of the potential use of surrogate depth maps in depth image based rendering for 2D-to-3D conversion. This latter approach exploits the ability of the human visual system to combine reduced disparity information that are located mainly at edges and object boundaries with pictorial depth cues to produce an enhanced sensation of depth over 2D images.

1. INTRODUCTION

Three-dimensional television (3D-TV) is expected by many to be the next step in the advancement of television. Stereoscopic images that are displayed on 3D-TV are expected to increase visual impact and heighten the sense of presence for viewers [1]. It is also expected that 3D-TV displays will provide multiple stereoscopic views, offering motion parallax as well as stereoscopic information.

The successful adoption of 3D-TV by the general public will depend not only on technological advances in stereoscopic 3D[1] displays but also on the availability of a wide variety of program contents in 3D. We rationalize that the conversion of two-dimensional (2D) images to 3D images is one way to alleviate the predicted lack of program material in the early stages of 3D-TV rollout. Furthermore, good 2D-to-3D conversion techniques can be profitable for content providers who are always looking for new sources of revenue for their vast library of program material.

2. FUNDAMENTAL PRINCIPLE

The fundamental principle underlying 2D-to-3D conversion techniques rests on the fact that stereoscopic viewing involves binocular processing by the human visual system of two slightly dissimilar images. The slight differences between the left-eye and right-eye images (horizontal disparities) are transformed into distance information such that objects are perceived at different depths and outside of the 2D display plane.

Thus, the various methods of converting 2D to stereoscopic 3D images involves the fundamental, underlying principle of horizontal shifting of pixels to create a new image so that there are horizontal disparities between the original image and a new version of it. The extent of horizontal shift depends on the distance of the feature of an object to the stereoscopic camera that the pixel represents. It also depends on the inter-lens separation (i.e., camera-baseline) because it will determine the new image viewpoint.

3. CUT-AND-PASTE TECHNIQUE

The most straightforward way to create a stereoscopic 3D image is to use the original image as a left-eye view and to generate a new image as the right-eye view by horizontally shifting local regions of the original image, using a cut-and-paste process. Using this method, stereoscopic depth can be created and any artifact in the new image would tend to be masked by the higher picture quality of the original image presented to the left eye [2].

2D-to-3D conversion with this technique is easier now because of the availability of digital processing and computer software. Object segmentation techniques can be used to isolate regions that can then be shifted horizontally to a new position. The extent of horizontal shift reflects the

[1] The term "3D" in this manuscript denotes "stereoscopic." It does not refer to the depth derived from pictorial cues to depth nor does it refer to 3D modeling in computer graphics.

level of horizontal disparity and, therefore, the relative depth of the local region/object in the scene. This technique is very effective when the objects in the depicted scene are well segregated. However, more laborious techniques are needed to deal with images that have multiple small objects, large areas with low textures, and gentle gradations of depth.

4. DEPTH MAPS AND DEPTH IMAGE BASED RENDERING

Since the extent that a pixel is shifted depends on the depth information for that pixel, it follows that an auxiliary image that provides the depth information for every pixel of an original, full-color, 2D image can be used to generate new views. These auxiliary images are referred to as "depth maps"[2] and the depth information is coded as luminance intensity level, usually lighter values for closer distances and darker values for farther distances.

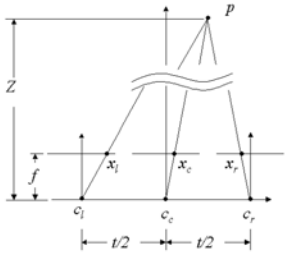

Figure 1. Top-view of camera configuration for generation of virtual stereoscopic images. P is a point of interest in the scene.

New views can be generated from an original image and its depth map using a process called depth image based rendering (DIBR). As shown in Figure 1, left-eye and right-eye images at virtual camera positions c_l and c_r can be generated for a specific camera-baseline indicated by *t, if* knowledge of the focal length, *f*, and the depth, Z, from the depth map is provided [3]. The geometrical relationship shown in Figure 1 can be expressed mathematically as in Equation 1 and the extent of pixel shifting can be computed:

$$x_l = x_c + \frac{t}{2}\frac{f}{Z} \quad , \quad x_r = x_c - \frac{t}{2}\frac{f}{Z} \qquad (1)$$

Since new views can be generated, DIBR is particularly useful for multiview stereoscopic systems that typically require between eight and sixteen views of a visual scene. There has also been great interest in the use of DIBR with respect to the development of a practical 3D-TV system because the method allows for efficient transmission and storage [4][5]. In general, a set of colour images and their associated depth maps are more efficient to compress than two (or more) streams of colour images for 3D-TV [6].

5. DEPTH MAP GENERATION

The use of DIBR in converting 2D images to 3D, however, requires the availability of depth maps. Depth maps can be generated using direct methods, such as using a ZCam that directly measures the distance of objects in a scene by employing the time required to bounce a beam of infra-red light back to a sensor located on the camera [7]. Another direct method is to project a structured light pattern onto the scene so that the depths of the various objects could be recovered by analyzing the distortions of the light pattern created by the 3D shape of objects in the scene [8]. However, aside from requiring specialized hardware, the direct methods have other drawbacks such as restrictive scene lighting and the need to have objects within a restrained distance.

In the literature, there are also many papers describing techniques for generating depth maps from disparity estimation, e.g., [9]. These techniques require both the left-eye and the right-eye images (stereoscopic pair) so that corresponding points/pixels in the two images could be determined and, thus, the disparity could then be calculated. However, the basis for 2D-to-3D conversion is that at the onset only a single 2D image or a stream of images for one eye is available. Thus, the issue of 2D-to-3D conversion can be viewed as an issue of depth map generation.

6. METHODS FOR DEPTH MAP GENERATION

There are as many ways to generate depth maps as there are pictorial cues to depth. For artists there are several pictorial cues, alone or in combination, that can be used to generate an effective impression of depth in two-dimensional pictures and images. A list that is not intended to be exhaustive includes texture gradient, atmospheric haze, shading, blur, and geometric perspective. The principal problem in the generation of depth maps is how to extract and convert these expressions of depth in nature into luminance intensity values in the form of a depth map.

6.1 Depth from Blur

A prominent area of research for determining the depth contained in a 2D image is that based on blur analysis. In extracting depth data from blur (commonly referred to as "depth from focus"), the depth information in a visual scene is obtained by modeling the effect that a camera's focal parameters have on the image, e.g., [10]. The underlying principle is that, for a given camera focal length, there is a direct relationship between the depth of an object, i.e., its distance from the camera, and the amount of blur of that object in the image. One common method in determining depth from blur is the use of inverse filtering to determine

[2] In this paper we do not distinguish between depth maps containing relative depth information ("disparity maps") and those containing absolute depth information ("range maps"). Theoretically, absolute depth information can be derived from the relative depth information if sufficient camera and capture information are provided.

the defocus operator so as to arrive at an estimate of depth [11]. However, there are other methods that can be used to estimate blur without relying on camera lens parameter. For example, blur can be estimated based on a multi-resolution wavelet analysis of local regions, whereby a high-frequency (sharp) region will have a large number of non-zero wavelet coefficients and a low-frequency (blurred) region would have a much lower number of counts [12].

Blur, and therefore depth, can also be estimated based on luminance intensity gradients in local regions containing edges and textures because increasingly sharp edges would have increasingly sharp gradients. We have used this latter method relatively successfully to generate depth information for depth image based rendering of stereoscopic views [13]. One of the central issues with this approach is the problem of filling in regions where there is no edge or texture information.

It should be noted that it is not an easy task to extract depth from blur because the blur found in images can also arise from other factors, such as lens aberration, atmospheric interference, fuzzy objects, and motion. In addition, equivalent extent of blur can arise even though in one case the object is farther away and in another case is closer than the position that produces the sharpest image. However, there are methods to overcome some of these problems and to arrive at more accurate and precise depth by examining controlled changes in blur using two or more images [14].

6.2 Depth from geometric perspective

An interesting approach to generating depth maps is through exploitation of gradient and linear perspective cues. This approach is intended for qualitative depth extraction, which is expected to be adequate for 3D-TV. With this approach, a single input image is first classified (as indoor, outdoor, or outdoor with geometry appearance) based on color segmentation [15]. Guided by the category of the image, the vanishing points and lines are then determined by identifying straight lines in the image. The region with the most number of intersections is considered to be the vanishing "point", and the major straight lines passing close to the vanishing point are considered to be vanishing lines that provide linear perspective of depth. In general, converging lines that are actually parallel indicate a surface that recedes in depth. Thus, depending on the slopes of the vanishing lines, different depth gradient planes can be generated with the vanishing point being at the farthest distance. This geometric depth information can then be fused with depth information of "objects" generated by the initial image segmentation and classification process to end up with a convincing "natural" depth map [15][16]. The limitation of this approach is the small number of image categories that can currently be used for 2D-to-3D conversion. However, this could be expanded.

6.3 Depth from edge information

In contrast to the above method of using depth gradients to fill a depth map, we have proposed the use of sparse depth maps for DIBR [13][17]. These so-called "surrogate" depth maps contain depth information that is concentrated mainly at edges and object boundaries in the 2D images. Although surrogate depth maps have large regions with missing and/or incorrect depth values, the perceived depth of the rendered stereoscopic images was judged to be adequate. It was speculated that the visual system combines the depth information available at the boundary regions together with pictorial depth cues to fill in the missing areas [13][17].

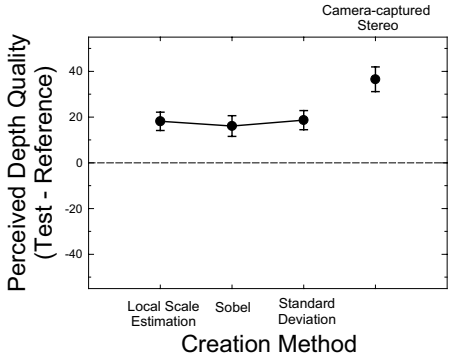

Figure 2. Ratings of depth quality of stereoscopic images for three different methods used in obtaining surrogate depth maps, expressed as a difference between ratings of a rendered test and a corresponding 2D reference image. The symbols represent the mean and the error bars are standard errors based on the ratings from 18 viewers, averaged over seven test images. The symbol on the far right refers to data for two stereoscopic images that were captured with real cameras.

It is worth noting that the surrogate depth maps were initially generated as a result of estimating the level of blur in local regions using a luminance gradient (or local scale) method [18], to obtain estimates of depth as discussed in Section 6.1. Later, we found that using a measure of variance of the luminance distribution within a block of pixels, and repeated for the whole image, was sufficient to produce effective surrogate depth maps. Interestingly, even using a standard Sobel edge detector to generate outlines of objects was adequate as a surrogate depth map [13][17]. See Figure 2.

With this approach of using object outlines and edges, the surrogate depth maps are relatively simple and easy to generate. All of this is computed, without knowledge of camera-capture parameters, to produce effective stereoscopic images. Admittedly, the perceived depth is qualitative and is suitable for applications where depth *accuracy* is not critical. Nevertheless, more research is needed with this method to extend it to image sequences.

7. 2D-TO-3D IMAGE SEQUENCES

Once an effective method has been developed for 2D-to-3D conversion of a single image, it can then be extended to image sequences. The most important issue in this step is to

ensure that the depth maps are reliable over frames such that there are no spurious fluctuations in depth. Clearly, one method is to smooth the depth maps over frames, such as through the use of a median filter. More complex methods involve "depth tracking" utilizing optical flow or block matching methods [19]. However, several issues were found with these complex methods, for example, difficulty in tracking because of fast motion or fuzzy object boundaries.

Another issue associated with the extension of 2D-to-3D conversion processes to image sequences is whether there is a need to generate one depth map for each colour image frame. This is a topic for future research given the fact that precision in depth maps may not be critical. In particular, the 2D-to-3D conversion process for surrogate depth maps, discussed in section 6.3, involves smoothing of the surrogate depth maps with an asymmetrical Gaussian filter before the rendering process [3][20]. Thus, the depth maps do not have sharp edges and it might be adequate to use the same depth map for the next *n* frames (where *n* needs to be empirically determined) and not create a visible artifact.

Despite these potential issues and problems for image sequences, there have been apparent successes with proprietary 2D-to-3D conversion techniques by some companies that provide services to the entertainment business. For example, Dynamic Digital Depth (DDD) Inc. has developed a semi-automatic process that involves manually inputting the depth of local areas within key frames (depth maps) of an image sequence. Computer algorithms then compute the depth information for larger regions within the key frames and for the whole length of the sequence based on the key frames [6].

As a final note, for video sequences there are other methods that do not always rely on depth maps for 2D-to-3D conversion. For example, there are methods that utilize motion parallax information [21]. However, these techniques have their own problems, such as issues of non-linear motion and the separation of camera motion from object motion.

8. CONCLUSIONS

Generation of depth maps is an important pre-requirement for depth image based rendering which is a useful technique for 2D-to-3D conversion of images and video. Techniques involving an analysis of blur, vanishing points and lines, or edges at object boundaries are useful for generating effective depth maps. The problem of extending 2D-to-3D conversion using single images to video is trickier. Techniques are required to avoid user interaction as far as possible and in reducing computational complexity. Furthermore, there is the additional issue of how to stabilize the depth of objects over frames. Despite these difficulties there are a few success stories and the future looks promising for the use of these techniques for 3D-TV.

9. REFERENCES

[1] S. Yano, & I. Yuyama, "Stereoscopic HDTV: Experimental system and psychological effects," *Journal of the SMPTE*, Vol. 100, pp. 14-18, 1991.

[2] L. B. Stelmach, W. J. Tam, D. Meegan, & A. Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 10(2), pp. 188-193, 2000.

[3] L. Zhang & W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Transactions on Broadcasting*, Vol. 51, pp. 191-199, 2005.

[4] P. Harman, "Home based 3D entertainment—An overview," *IEEE Conference on Image Processing*, Vol. 1, pp. 1-4, 2000.

[5] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Stereoscopic Displays and Virtual Reality Systems XI*, Vol. 5291, pp. 93-104, 2004.

[6] J. Flack, P. Harman, & S. Fox, "Low bandwidth stereoscopic image encoding and transmission," *Stereoscopic Displays and Virtual Reality Systems X*, Vol. 5006, pp. 206-214, 2003.

[7] G. Iddan & G. Yahav, "3D imaging in the studio," *Videometrics and Optical Methods for 3D Shape Measurement*, Vol. 4298, pp. 48-55, 2001.

[8] D. Scharstein & R. Szeliski, "High-accuracy stereo depth Maps using structured light," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp. 195–202, 2003.

[9] D. Scharstein & R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, Vol. 47(1–3), pp.7–42, 2002.

[10] S. H. Lai, C. W. Fu, & S. Chang, "A generalized depth estimation algorithm with a single image," *PAMI,* Vol.14(4), pp. 405-411, 1992.

[11] J. Ens, & P. Lawrence, "An investigation of methods for determining depth from focus," *PAMI,* Vol.15(2), pp. 97-108, 1993.

[12] S. A. Valencia & R. M. R. Dagnino, "Synthesizing stereo 3D views from focus cues in monoscopic 2D images," *Stereoscopic Displays and Virtual Reality Systems X*, Vol. 5006, pp. 377-388, 2003.

[13] W. J. Tam, A. Soung Yee, J. Ferreira, S. Tariq, and F. Speranza, "Stereoscopic image rendering based on depth maps created from blur and edge information," *Stereoscopic Displays and Applications XII*, Vol. 5664, pp.104-115, 2005.

[14] S. Chaudhuri & A. Rajagopalan, "Depth from defocus: a real aperture imaging approach," Springer Verlag, 1999.

[15] S. Battiato, S. Curti, E. Scordato, M. Tortora, and M. La Cascia, "Depth map generation by image classification", *Three-Dimensional Image Capture and Applications VI*, Vol. 5302, pp. 95-104, 2004.

[16] S. Battiato, A. Capra, S. Curti, and M. La Cascia, "3D stereoscopic image pairs by depth-map generation", *Second International Symposium on 3D Data Processing, Visualization and Transmission,* pp. 124-131, 2004.

[17] W. J. Tam, F. Speranza, L. Zhang, R. Renaud, J. Chan, & C. Vazquez, "Depth image based rendering for multiview stereoscopic displays: Role of information at object boundaries", *Three-Dimensional TV, Video, and Display IV*, Vol. 6016, pp. 75-85, 2005.

[18] J. H. Elder & S. W. Zucker, "Local scale control for edge detection and blur estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, pp.699-716, 1998.

[19] M. Lambert, "Synthese von 3D videos," Graduate thesis, http://www.marlam.de/thesis.pdf, December 15, 2005.

[20] W.J. Tam, G. Alain, L. Zhang, T. Martin, & R. Renaud, "Smoothing depth maps for improved stereoscopic image quality," *Three-Dimensional TV, Video and Display III*, Vol. 5599, pp.162-172, 2004.

[21] L. Zhang, B. Lawrence, D. Wang, & A. Vincent, "Comparison study on feature matching and block matching for automatic 2D to 3D video conversion," *IEE European Conference on Visual Media Production*, pp. 122-129, 2005.