HYBRID VIEW- SYNTHESIZING APPROACH FOR MULTIVIEW APPLICATIONS

Iliya Koreshev[1], *Mahsa T. Pourazad*[1,2], *Panos Nasiopoulos*[1]

[1]University of British Columbia
[2]TELUS Communications Inc.

ABSTRACT

The need for synthesizing virtual views is more pronounced with the advent of multiview displays, as it is not practical to capture all of the possible views when filming multiview content for different multiview display technologies. Unfortunately, during the process of view synthesis, areas, which were occluded by foreground objects in the real views, become visible in the virtual view as holes. Filling these holes with natural looking color and texture information is challenging. We present a new hybrid approach for synthesizing multiview videos, which utilizes an effective hole filling method that preserves the perceptual quality of 3D content. Subjective evaluations confirm that our approach outperforms the current state-of-the-art interpolation-base synthesizing method.

Index Terms — 3D TV, multiview TV, view synthesizing, hole filling.

1. INTRODUCTION

Three-dimensional (3D) video provides users with a more engaging and realistic impression of scenes than traditional two-dimensional (2D) video. Users can perceive depth in 3D videos the same way as they would perceive depth if they were looking at a live scene. As this technology evolves and grows so do the expectations of consumers. Watching 3D content without wearing cumbersome glasses is one of the key features that 3D technology consumers demand. In this regard, researchers and display manufacturers are working towards developing multiview displays that provide viewers with a wider viewing angle and do not require them to wear glasses for watching 3D content; however, to support this technology several views of the scene are required to be captured simultaneously. Multiview content production is expensive and highly demanding in terms of camera configuration and post processing. In summary, it is not practical to capture and transmit all the required views for multiview display applications. To address these problems, the 3D Video (3DV) ad-hoc group (part of ISO/IEC Moving Pictures Experts Group (MPEG)), recommended the use of two or three views plus a depth map (3DV data format) to synthesize high-quality views for multiview applications [1]. While this type of data format reduces the limitations regarding camera inputs and transmission bandwidth, it introduces a new challenge, which is synthesizing high quality views. The main issue with the synthesizing process is related to estimating the information of the occluded areas. During the synthesizing process, areas of the background that were occluded by foreground objects in the available views become visible in the synthesized views. These areas (holes) must be filled with realistic data to avoid noticeable artifacts. A well-practiced solution is to apply interpolation to estimate the missing texture. This approach has been utilized in the existing state-of-the-art view synthesis reference software (VSRS), which has been selected by the MPEG-3DV group to synthesize test sequences for future 3D video compression standardization activities [2]. VSRS uses the depth and texture information of available views to generate intermediate virtual views. The foreground and background objects are segmented using the depth data and then are horizontally shifted based on their depth range to create virtual views. This shifting is what produces areas with missing texture called holes. VSRS uses the nearest neighbor interpolation approach to fill these holes (by selecting neighboring pixels and assigning their average value to the hole pixels).

The downfall of interpolation-based hole-filling methods is that the interpolated texture does not resemble the true texture of the occluded areas, but instead looks as if a clone tool was applied to those areas, in a sense that small parts of the neighboring texture are simply replicated over and over. This approach usually produces a similar looking color to the true background, but fails to reproduce texture that exists in those areas, thus reducing the quality of the synthesized views and hampering the overall 3D effect. To avoid creation of holes in the synthesized view, a group of researchers from the Disney Research lab in Zurich have proposed to use a warping technique to generate synthesized views from the available views [3]. This method first requires a sparse saliency map to be created. This saliency map helps with separating foreground and background objects. During the second stage, the saliency map information is used to stretch or compress some parts of the picture. The result is that foreground objects are shifted either to the left or right, depending on which view is being synthesized. Shifting the objects instead of moving them, changes the size/shape of the objects which now are either stretched (enlarged) or compressed (shrunk). The end result is that this method does not produce holes. However, due to warping (stretching or shrinking), some deformation may be evident in the generated virtual views. This is more prominent around foreground objects that have large disparity (need to be shifted more) and which are also close to background objects with well-defined vertical edges. In such cases, since variant amounts of warping is applied to the foreground objects, the vertical edges are deformed and become wavy rather than straight.

It is recognized that the quality of 3D content is extremely important, as low quality 3D videos can produce eyestrain, headaches, and generally unpleasant viewing experience for the viewers [4]. Thus to enable multiview technology, there is a strong need for an effective view synthesizing approach that does not compromise the quality of the generated view with inadequate/poor hole-filling.

In this paper, we propose a new view synthesizing approach with very efficient hole-filling performance, which is the hybrid form of the interpolation approach in [2] and the warping technique (Disney approach) in [3]. In our method a view is synthesized by shifting the objects based on their depth map similar to [2]. The generated holes are filled by stretching/warping the existing background texture, a concept that is

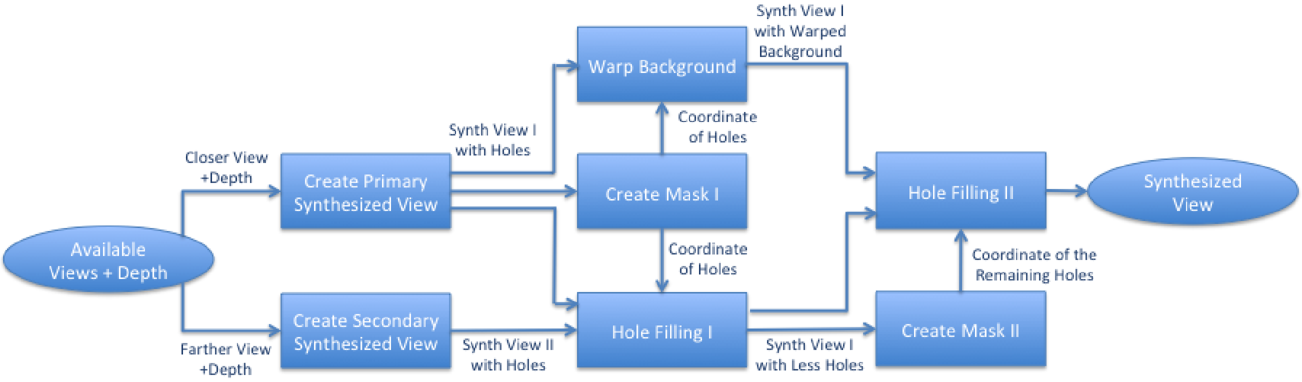

Figure 1. The flowchart of our proposed hybrid view synthesis method

derived from the work described in [3]. Note that unlike [3] which applies warping to both background and foreground, we only warp the background so that the shape and the size of the foreground objects are intact by our hole filling process. By warping the existing background texture, the holes are filled with texture similar to that of the surrounding background region. Since the texture of filled areas is similar to that of surrounding areas, the filled areas look more natural and the overall quality of the synthesized views is improved. To evaluate the performance of our algorithm, we conduct subjective tests and compare our results with synthesized views generated by the state-of-the-art view synthesis reference software (VSRS) [5].

The remainder of this paper is divided into four sections. Section 2 elaborates on the proposed method. In Section 3 the performance of our method is evaluated and compared with that of the existing state of the art technology. The results of our experiments and future work are discussed in Section 4, and conclusions are drawn in Section 5.

2. HYBRID VIEW SYNTHESIS

The need for synthesizing virtual views is more pronounced with the advent of multiview displays. It is not practical to capture all of the views when filming the multiview content. In this study we intend to improve on the existing view synthesizing techniques by using a new approach for hole filling to generate virtual views that resemble real views more closely. In our study some general ideas are taken from an existing interpolation technique and a warping method to create a hybrid view synthesis approach that allows generation of high quality synthesized views. Figure 1 shows the flowchart of the proposed method. More details are provided in the following subsections.

2.1 Creating primary synthesized view

In the view synthesizing problem, we have multiple views captured with multiple cameras (usually with a parallel setup) and we try to synthesize additional views from the available ones as if there were more cameras in the multiview camera setup. The closer the real camera views are to the virtual camera view the more accurate is the synthesized view. To this end, we create a primary synthesized view based on the closest camera view and its depth map. To do this in a way similar to VSRS, the appropriate shifting amount for different objects in the scene is calculated using the depth and texture information as follows [6]:

$$p_{pix} \approx -x_B \frac{N_{pix}}{D}\left(\frac{m}{255}\left(k_{near} + k_{far}\right) - k_{far}\right) \qquad (1)$$

where p_{pix} is the shift parameter at depth level m, D is the viewer distance from the display, k_{near} and k_{far} are the distance of the closest and farthest object to the camera, and N_{pix} is the user defined parameter controlling the maximum parallax based on the screen width. The maximum parallax determines the depth of the closest object in the scene when watched on the screen. This shifting process creates holes (pixels with missing color and texture information) in way a similar to a regular interpolation-based synthesizing approach. The coordinate of the pixels corresponding to these holes are registered by creating a mask with the value of zero for the hole pixels and the value of one for the rest of the pixels. This mask is called “Mask I” (see Figure 1).

2.2 Matching based hole filling

In order to fill up the holes, the first step is to use the information of the farther available view. To this end, a secondary synthesized view is generated solely based on the farther view by following the same procedure as creating the primary synthesized view. Once the secondary synthesized view is generated, the holes in the primary synthesized view (registered in Mask I) are filled by corresponding available areas in the secondary synthesized view, with the condition that the depth of these areas matches the depth of the neighboring objects to the holes in the primary synthesized view. Note that this condition is not always met, since the secondary view is generated based on the farther view, which covers different areas in the scene. The coordinates of the remaining holes in the primary synthesized view are registered by creating “Mask II” (similar to Mask I).

2.3 Warping the background

To fill up the holes, our hybrid approach also applies warping to the background area of the primary synthesized view. To do this, Mask I is used to generate the list of warping points. The warping start-points (the points where the hole-areas start with a small overlap towards the background) and the warping end-points (the points where the hole-areas end with a small overlap towards the foreground) are identified using Mask I. To avoid vertical parallax, the warping process for filling the holes should be done in the horizontal direction, so the vertical coordinate of the warping start-point and end-point are equal. We also restrict the warping process to not use the information of the corner of the synthesized image. We do that because there is not enough texture data at the corners that can guarantee effective warping. Then, Piecewise cubic Hermite interpolation [7] is applied to the primary synthesized view which takes a matrix containing all the points in the image as well as a matrix with the corresponding warp points and produces a matrix containing all the interpolated points between the warp points as follows:

$$\mathrm{A} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}, B = \begin{bmatrix} x'_1 & y'_1 \\ \vdots & \vdots \\ x'_n & y'_n \end{bmatrix} \quad \text{number of nonzero elements in B} = \mathrm{m} \qquad (2)$$

$$C = \begin{bmatrix} x''_1 & y''_1 \\ \vdots & \vdots \\ x''_n & y''_n \end{bmatrix} \quad \mathrm{m} < \mathrm{n}$$

where *n* is the total number of pixels in the primary synthesized view (including hole pixels), and *m* is the total number of warp points. Matrix A contains (x_i, y_i) which are the coordinate of pixel *i* in the primary synthesized view, matrix B is a sparse matrix containing (x_j', y_j') which are either the corresponding coordinate of the warping points (identified based on Mask I) or if no warping points exist then just zero, and matrix C contains (x_i'', y_i'') which are the interpolated locations of every *x* and *y* point in the warped image. Basically with the help of Piecewise cubic Hermite interpolation all the points in the primary synthesized view are mapped to *(x'', y'')*. Once the new coordinates in the warped image are obtained, the pixel values of primary synthesized view are mapped accordingly to create the warped image.

2.4 Creating hybrid synthesized view

To create the final synthesized view as shown in Figure 1, the hole areas in the secondary virtual view (which are marked in Mask II) are filled with the data from the warped image. Once this process is complete, we obtain a virtual view where all the holes are filled either with data from the secondary virtual view or from the warped image. As it can be observed from Figure 2, unlike VSRS, our hybrid approach generates a realistic looking texture for hole areas without hampering the quality of more visually important foreground objects.

3. SUBJECTIVE EVALUATION

The performance of our method is evaluated based on subjective tests and is compared to that of the existing VSRS package (version 3.5) [5]. For this evaluation we used three test sequences, namely "Balloons" (1024x768, 30fps, 300 frames), "Kendo" (1024x768, 30fps, 300 frames) and "GT_Fly" (1920x1088, 25fps, 250 frames). These test streams along with their depth information are selected from the database provided by MPEG for the Call for Proposals (CfP) on 3D video coding [8]. All the videos are in YUV 4:2:0 format and progressive.

The viewing conditions were set according to the ITU-R Recommendations BT.500-13 [9]. Twenty volunteer subjects, ranging from the age of 18 to 57 participated in the evaluations. All subjects had none to marginal 3D image and video viewing experience. They all were screened for color and visual acuity (using Ishihara and Snellen charts), and for stereo vision (Randot test – graded circle test 100 seconds of arc). The evaluation was performed using a 46" Full HD Hyundai 3D TV (Model: S465D) with passive glasses. The TV settings were as follows: brightness: 80, contrast: 80, color: 50, R: 70, G: 45, B: 30. The 3D display and the settings are based on MPEG recommendations for subjective evaluation of the proposals submitted in response to the 3DV CfP [8].

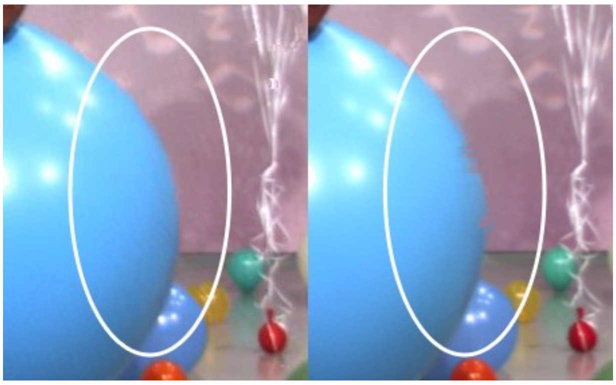
Figure 2. Artifacts in the VSRS generated image on the right and our hybrid approach on the left

At the beginning of each evaluation session, a demo sequence ("Dancer", 1920x1088, 25fps) with different levels of synthesizing artifacts was played for the subjects to become familiar with the artifacts and the testing process. The process of rating the sequences was explained during that time so that the subjects would be ready when the first rated sequence was played. After the demo sequences were done, a 5 second break interval was shown which informed the subjects that the next sequence they see should be rated. The "Dancer" test sequence was then omitted from the actual evaluation procedure to maintain the purity of the results.

After that, the viewers were shown the synthesized stereoscopic test sequences in random order, so that they would watch two different synthesized versions of the same sequence consecutively, without knowing which video was generated by our method or VSRS. Between test videos, a ten-second gray interval was provided to allow the viewers to rate the perceptual quality of the content and relax their eyes. Here, the perceptual quality reflects whether the displayed scene looks pleasant in general. In particular, subjects were asked to rate a combination of "naturalness", "depth impression" and "comfort" as suggested by Hyunh-Thu et al. [10]. For ranking, there were 11 quality levels, 10 indicating the highest quality and 0 the lowest quality. Three test scenarios were examined: 1) right-view is synthesized, 2) left-view is synthesized and 3) both views are synthesized. Switching the synthesized view between the right and the left eye compensated for the effect of eye dominancy (out of the twenty volunteers we had 13 left-eye dominant and 7 right-eye dominant subjects).

4. RESULTS AND DISCUSSION

The first step after collecting the experimental results is to remove the outliers according to the ITU-R Recommendations BT.500-13; there were 2 outliers [9]. The mean opinion scores from the viewers were then calculated with a 95% confidence interval as shown in Figure 3. As it can be observed, the results from the subjective tests show that, with 95% confidence interval, the scenes generated using our hybrid approach scored consistently higher than those generated using VSRS, which confirms the superior performance of our technique. Even in the case where both views are synthesized, the MOS score for our hybrid approach is higher than that of VSRS. Moreover, the subjective tests show that the MOS scores for the case where both views are synthesized using our hybrid approach are similar or higher than those for the case where only one view is synthesized by VSRS.

As Figure 3 shows, in general the MOS scores for the case where both views are synthesized are a bit lower than the case where only one view is synthesized. This is due to binocular rivalry [11]. In the latter case, the information of the dominant picture (original view in our case) suppresses the information of the less-dominant view (synthesized view), thus the perceived quality of the overall picture is higher than the case both views are synthesized (less-dominant).

An interesting general observation from the results is that for the cases where the right view is synthesized the MOS is higher than the cases where left view is synthesized. We believe this can be explained by the fact that we had more left eye dominant subjects for our tests so that when the synthesized view is shown to their left eye, the artifacts are affecting their 3D perception and they rate the overall quality much lower compared to the case that the synthesized view is shown to their right eye (non-dominant eye).

As the average scores per each test sequence show (see Figure 3), the difference between the quality of our synthesized view and the ones generated by VSRS is higher in the case of “GT-Fly”. This is due to the high accuracy of the depth map of this computer-generated stream. Since our method relies on precise movement of objects at different depth levels, a cleaner depth map allows our method to shift only the specific objects and exclusively warp the background area and create a high quality synthesized view.

As our future work we plan to extend our proposed hybrid approach to include extrapolation of views using only a single view and its depth map. We believe that extrapolation of virtual views is just as important as interpolation due to the increasing demand for converting existing 2D videos to 3D format.

Our subjective results confirm that our hybrid method allows for better hole filling and view synthesis than the current state-of-the-art interpolation-based technique. While our approach cannot replicate the true texture that is missing in the hole areas, it will provide similar looking texture without hampering the perceptual quality of the 3D content.

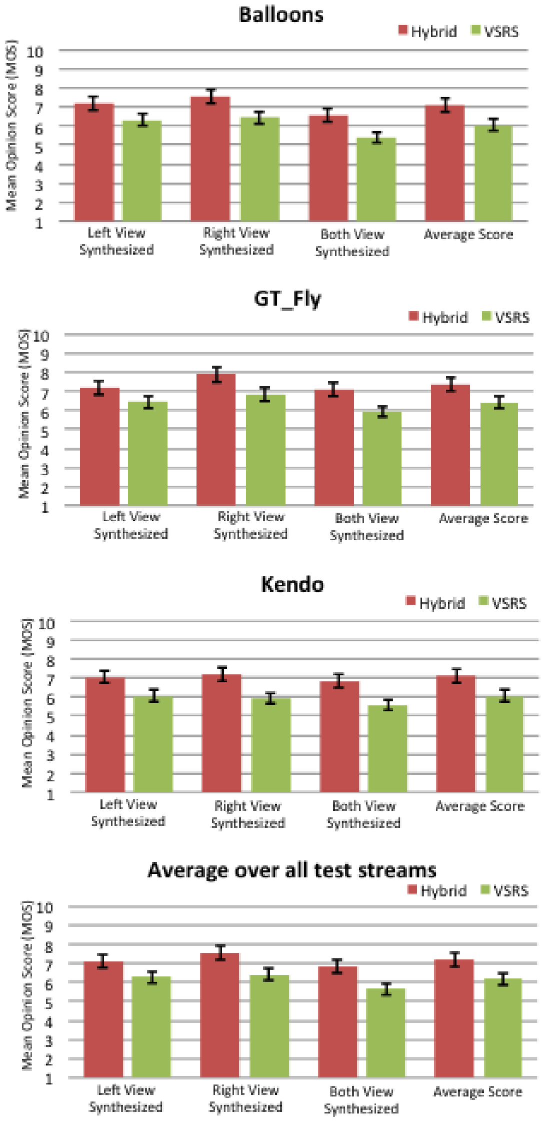

Figure 3. Mean opinion scores for individual scenes and the average mean opinion scores for all scenes combined. The black bar on each graph shows the 95% confidence interval.

5. CONCLUSION

Consumer multiview displays are expected to reach the market in less than 4 years. This necessitates the development of efficient view-synthesizing methods, as capturing multiview scenes is neither practical nor cost-effective. To this end, we proposed a new hybrid view synthesizing approach, which utilizes the merits of two existing techniques while overcoming their downfalls. Our proposed method synthesizes new views in a similar fashion to the interpolation-based view synthesizing techniques. However, to fill the holes it uses an effective warping technique instead of the traditional nearest neighbor interpolation approach. Unlike the Disney proposed approach in [3], which warps both background and foreground objects, our approach only warps the background, thus avoiding deformation of the foreground objects of interest. Since most of holes are present in the areas where foreground objects occlude background objects, warping the background areas keeps intact the more visually important foreground objects. Subjective evaluations confirm the superior performance of our method compared to the current interpolation-based state-of-the-art view synthesizing method.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 MPEG, Document N10357, “Vision on 3D Video,” N10357, 87th MPEG meeting, Geneva, February 2009.

[2] ISO/IEC JTC1/SC29/WG11 MPEG Document N11631, ”Report on Experimental Framework for 3D Video Coding,” Guangzhou, China, October 2010

[3] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. 2010. “Nonlinear disparity mapping for stereoscopic 3D,” ACM SIGGRAPH 2010 papers (SIGGRAPH '10), Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 75 , 10 pages.

[4] S. L. P. Yasakethu, W. A. C. Fernando, B. Kamolrat, and A. Kondoz, “Analyzing perceptual attributes of 3d video,” IEEE Transactions on Consumer Electronics, vol.55, no.2, pp.864-872, May 2009.

[5] ISO/IEC JTC1/SC29/WG11, MPEG, “View Synthesis Software Manual,” Sept. 2009, release 3.5.

[6] ISO/IEC JTC1/SC29/WG11 MPEG, Document N8038, “Committee Draft of ISO/IEC 23002-3 Auxiliary Video Data Representations,” Montreux, Switzerland, April 2006.

[7] F. N. Fritsch and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," SIAM J. Numerical Analysis, Vol. 17, 1980, pp.238-246.

[8] ISO/IEC JTC1/SC29/WG11 MPEG, Document N12036, “Call for proposals on 3D video coding technology,” 96th MPEG meeting, Geneva, March 2011.

[9] ITU-R, “Methodology for the subjective assessment of the quality of television pictures,” ITU-R, Tech. Rep. BT.500-13, 2012.

[10] Q. Hyunh-Thu, P. L. Callet, and M. Barkowsky, "Video quality assessment: from 2D to 3D challenges and future trends," Proc. of 2010 IEEE 17th International Conference on Image Processing, (ICIP), pp.4025-4028, 2010.

[11] H. Asher, “Suppression Theory of Binocular Vision,” Brit. J Ophthalmol, 1953 January, 37(1), pp.37–49.