GENERATION OF LAYERED DEPTH IMAGES FROM MULTI-VIEW VIDEO

Xiaoyu Cheng, Lifeng Sun and Shiqiang Yang

Tsinghua University
Department of Computer Science and Technology
100084, Beijing, China

ABSTRACT

The feature of rendering arbitrary view make layered depth image (LDI) be suitable to present multi-view video data and provide interactivity form such as free view video(FVV). However, visual artifacts occured in the rendered result limit the practicality of LDI. In this paper, we proposed an approach to deal with this problem during generating of LDI, which refines colors of the depth pixels by choosing proper candidate depth pixels, removes matting effects by projecting LDI backward to reference views, and eliminate the gaps or holes dure to disocclusion and undersample by local background interpolation. Expermental results show that our approach is practicable and efficient.

Index Terms— Layered Depth Image (LDI), multi-view video, Three-dimensional image

1. INTRODUCTION

Multi-view video consists of multiple video sequences captured for the same scene synchronously by a camera array at different locations. It is possible to generate scenes of intermediate virtual views from the multi-view video data via view synthesis, which introduces more forms of interactivity into multi-view video.

There are various approaches for view synthesis. The traditional geometry-based rendering (GBR) approach used textured three-dimensional (3-D) geometric models to render virtual view, which need precise 3-D models of objects and scenes. Such models are very difficult to create for photo-realistic scenes, especially for dynamical scenes. A realistic GBR environment requires millions of polygons, complex lighting models, generous texture mapping and great computational cost.

As an attractive alternative to GBR, image-based rendering (IBR) technologies have received much attention for acquisition and visualization of photo-realistic interactive environments, which focus on generating images directly from other images instead of using 3-D geometric models. Among a variety of IBR techniques, the Layered Depth Image (LDI) is one of the most efficient representing and rendering methods for 3-D objects with complex geometries [1]. It is a considerable technology for scenes represention of multi-view video to provide the interactivity efficiency by rendering arbitrary view from its data set. However, some aspects need further refinement in practice, such as video matting and holes/gaps in the rendered results of LDI .

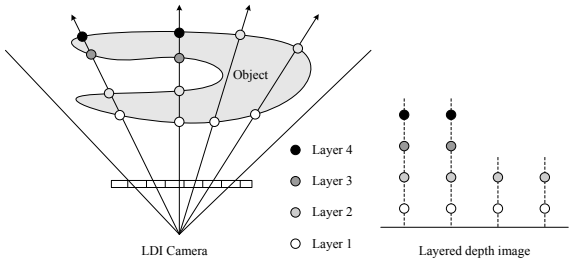

Fig. 1. Layered depth image and LDI camera [1]

Commonly, the LDI data set is generated by stereo vision technique. During stereo vision computation, each pixel is assumed to have a unique disparity. This is generally not the case. Some pixels along the boundary of object will receive contributions from both the foreground and background colors. If pixels with mixed colors are used roughly for rendering, visible artifacts will result. On the other hand, a low-cost sparse camera array is usually used to capture scenes. Partial scene may be absent due to occlusions that perhaps result holes in rendered results when the viewing angle is large enough. Visual artifacts such as partial occlusion and matting at boundary are difficult to overcome [2].

In this paper, we focus on problems of visual artifacts that may result when LDI is used in interactive multi-view video, and propose a solution for generating LDI to make it being practicable in multi-view video applications.

2. BACKGROUND

A layered depth image is a array of layered depth pixels ordered from closest to furthest from a so-called *LDI camera*. Each layered depth pixel contains multiple depth pixels at per pixel location. The farther depth pixels, which are occluded

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant No. 60503063

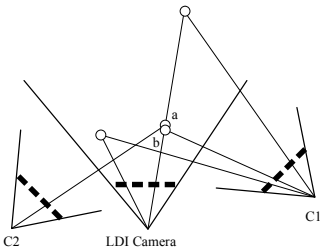

Fig. 2. Generating and rendering of LDI [1]

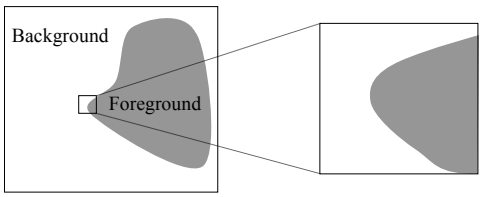

Fig. 3. Matting effect

from the viewpoint at the center of LDI camera, will appear as the viewpoint moves away from the center of LDI camera. This IBR technique is an abstract of depth images, which assigns a z-value in addition to the x and y values normally associated with a pixel.

There are two ways to generate an LDI from natural images: generate LDI from multiple depth images and generate LDI from real images directly [1], illustrated as Fig. 2. If multiple color and depth images from natural images are obtained, the LDI is constructed by warping pixels in other camera locations, such as C1 and C2. Otherwise, The LDI can be generated directly from the input images by an adaptation of Seitz and Dyers voxel coloring algorithm [3], in which the regular voxelization is replaced by a view-centered voxelization similar to the LDI structure. Candidate voxels along outward rays from the LDI camera are projected back into input images. If all input images agree on same color, this voxel is marked as a depth pixel and inserted into the LDI structure.

Rendering view of scene at certain viewpoint can be performed by the fast warping algorithm proposed by McMillan [4]. This approach enables straightforward construction of LDI from images that do not contain depth per pixel [1]. But artifacts maybe occur in the rendered result.

Holes may occur in the rendered images for two reasons [5]. One case is undersampling, this occurs when the sampling rate is larger than that provided by the reference image. Another case occurs when part of the rendered image is revealed but is not visible in the reference image, which is called disocclusion.

Folds (overlaps) occur when two or more pixels that are close together and be rendered in the same spot. The hole-filling problem arising in image-based techniques that use depth maps is commonly solved by a combination of filtering and splatting. The original LDI approach uses splatting algorithm to deal with the visual artifacts of undersampling and folds in 3-D warping. LDI uses nearest neighbor or another picture to fill in holes. Chang et al. proposed the LDI tree for hierarchical rendering and progressive refinement to avoid the artifacts caused by undersampling [6].

Image matting or video matting effects will occur at boundary of objetcs. The problem of video or image matting has been attended for a long time. Some approaches were proposed to deal with the visible artifacts at boundary of objects, such as image priors [7], Bayansian approach [8, 9], and boundary matting [10]. However, these approaches do not aim at LDI structure. Another method is that the LDI was used to present an object in the scene. The whole scene was presented by multiple primitives such as LDIs, depth sprites environment map etc. [1]. It is difficult to apply such a structure to multi-view video. So, we use single LDI as presentation of the scene in multi-view video scenario, and make improvements on the generation and rendering of LDI to reduce the artifacts in synthesized views. Our approach takes some ideas from existing techniques for video matting and splatting.

3. IMPROVEMENT IN THE QUALITY OF LDI

Visual artifacts are essentially due to imperfect LDI data set. Therefore, we first try to solve this problem during the stage of generating LDI. As the limit of LDI technique, we pay attention on the case of diffuse reflection only. The two methods of generating LDI from multiple natural images are equivalence in substance. Depth images can be generated by process of correspondence match and disparity estimation using stereo vision approaches as voxel coloring algorithm used in generating LDI directly from multiple real images. Without loss of generality, we focus on the method using depth images.

If background colors are far different from foreground, matting effects will occur at pixels along the boundary of object in reference image due to contributions from both the foreground and background colors (illustrated as Fig. 3). No matter do the mixed pixels act as foreground or background, visual artifacts will occur in the scene rendered at spots other than original reference view.

Generally, matting effect is due to the mixed pixel near occluded boundary. The corresponding mixed pixels in different reference view are caused by different pairs of foreground and background pixel. As the case shown in Fig. 4, C_1 is blend of F and B_1, and C_2 is blend of F and B_2. Furthermore, some occluded background pixels will be visible in two or more reference views in which the corresponding pixels are far from boundary, such as B_2 is occluded in V_1 but it is visible in V_2 and V_3. In fact, pixels from different depth images will be warped into the same depth pixel position in LDI dur-

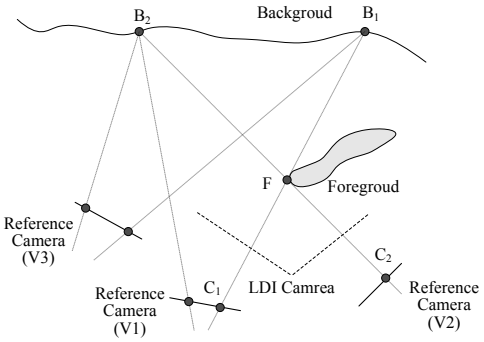

Fig. 4. Projecting LDI to reference views. B2 is visible in V1 and V3. B1 can be viewed from V3.

ing generating stage. If we take into account corresponding pixels from different views, matting effects can be probably reduced.

First, pixels with different colors from referene views are warped to LDI camera. Generally, one camera of the referene views is choosen as LDI camera with the same resolution. When one pixel of referene views are being warped to LDI camera, the corresponding depth pixel will not locate at proper ray from project center. We select the closest one of overlaped depth pixels from different referene view around certain ray of the LDI. Then depth pixels with similar depth values are clustered by a simple algorithm such as *K-means* if these pixels are warped to same depth pixel position. If pixel C_i from reference view V_i is warped to DepthPixel DP in the LDI, we cluster C_i into $\mathrm{R_k}$ $(k = 1, 2, \cdots, m)$. We choose average color of pixels in the larger cluster as the color of the DepthPixel instead of average color of all pixels:

$$\begin{cases} C_{DP} = \dfrac{1}{|R|} \displaystyle\sum_{R} C_i \\ R = \arg\max\limits_{R_i} |R_i| \end{cases} \tag{1}$$

In this case, the proper background color can be found if a larger cluster exists. Using the method of cluster, matting effects can be removed in most case if the reference views are not close together.

After the initial LDI is constructed, we reproduce reference images by reprojecting the LDI to every reference cameras, especially for pixels within boundary regions. Comparing the regenerated reference image to the origin, we can detect mixed pixels using a given threshold. Then, we will remove the matting effects of these pixels and refine the corresponding DepthPixels in LDI.

Assuming the scene is diffuse, the mixed pixel color C can be considered as blend of foreground color F and background color B with alpha value α. It can be modeled by the compositing equation:

$$C = \alpha \cdot F + (1 - \alpha) \cdot B \tag{2}$$

For a given mixed DepthPixel C_{DP}, corresponding pixel C_i in reference view V_i can be considered as blend of foreground F and background B_i.

$$C_i = \alpha_i \cdot F + (1 - \alpha_i) \cdot B_i \tag{3}$$

When the LDI is projected to reference cameras, like that the reference cameras act as LDI cameras, we can get depth images at each reference view. According the regenerated depth maps, we mark the boundary pixels by check discontinuous depth with a threshold. A foreground pixel may be occluded in a certain view. So, only the visible pixels are marked. Clustering method can remove some mixed pixels in background. We can assume that the background pixels in the depth images are perfect and the background texture varies slowly in most cases. B_i can be projected from LDI, or use the average of nearest-neighbor pixels at background depth if it is absent. The unknown F can approximately take the average of nearest-neighbor pixels F_i. The α_i can be calculated approximately according Eq. 3.

$$\alpha_i = (C_i - B_i)/(F_i - B_i) \tag{4}$$

α_i means the proportion of foreground to mixed pixels. It can refine the foreground at sub_pixel level and increase the density of DepthPixels in LDI. To simplify the process, we use α_i to estimate a candidate foreground pixel and make no change in LDI structure. The ideal color F of foreground can be estimated as:

$$F = \sum \alpha_i^2 F_i / \sum \alpha_i^2 \tag{5}$$

A candidate foreground pixel in a reference view is confirmed as foreground with color F if $\alpha_i \geq 0.5$. Otherwise, it will be redefined as background, and takes the color of B_i and average depth of nearest-neighbor background. The DepthPixels correspond to new foreground or background pixels will recalculated accordingly.

Although splatting method can solve the problem of undersampling satisfactorily, the holes still remain in the final images due to partial scene occluded in all reference view and calculation errors. The sizes of gaps or holes caused by calculation errors are very small in common. We can check the gaps or holes by certain threshold in LDI and interpolate small holes with nearest-neighbor pixels. The larger holes are considered as disocclusion. There are no extra information can be used to remove these holes. In general, texture filling is better, and more complex than color filling. So, we use a filling interpolate the hole in the farthest background by average color of boundary pixels around it.

4. EXPERIMENTAL RESULT AND ANALYSIS

We use the data sets *Breakdancers* [2] from MSR in our experiments. *Breakdancers* is captured by 8 cameras within

(a) (b) (c)

Fig. 5. experimental results

about 20cm horizontal spacing. Depth maps computed by stereo matching algorithms and camera parameters are provided. In our experiments, we construct the LDI from depth images and render it at a new viewpoint other than reference views.

Fig. 5(a) shows the result generated with original LDI approach. The gaps caused by computation error and holes caused by disocclusion is obvious. Matting effects can be find around boundary of the hat located at center of the scene. We use cluster approach to refine the LDI data. The result is shown as Fig 5(b). Finally, we fill the holes in Fig. 5(b) by local background color. The refined result is shown as Fig. 5(c).

The data set *Breakdancers* has precise depth maps. Consequently, the experiment result is satisfied. Our approach improve visual effects of rendered image at novel viewpoint. There are some unconspicuous visual artifacts along the occluded boundary, and the filled parts for holes is not so clear. The final result shown as Fig. 5(c) can meet the need of interactive multi-view video applications.

5. CONCLUSIONS AND FUTURE WORKS

The interactivity of multi-view video can be enhanced by using LDI to present multi-view video data. However, visual artifacts occured in the rendered result limit applications of LDI generated from nature images. We proposed an approach to deal with this problem which refines the data of LDI and improves the rendering effects. Experimental results show that our approach is effective. There are still some aspects needing further improvement, for instance, boundary matting artifacts are not completely eliminated if depth images have serious errors in segmentation of the foreground and background. In addition, experiment results indicate that the precise depth maps are also very important issues to LDI applications to deal with. As single LDI is suitable to present scenes within a particular field of view, how to describe converge views of large field with LDI structure will be one of our future works.

6. REFERENCES

[1] J. Shade, S. Gortler, L. Hey, and R. Szeliskiz, "Layered depth images," in *Proceedings of ACM SIGGRAPH'98*, Orlando, Florida, 1998, pp. 231–242.

[2] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "Highquality video view interpolation using a layered representation," in *Proceedings of ACM SIGGRAPH'04*, 2004, pp. 600–608.

[3] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, 1997, pp. 1067–1073.

[4] L. McMillan, "A list-priority rendering algorithm for redisplaying projected surfaces," Tech. Rep. 95-005, University of North Carolina at Chapel Hill, 1995.

[5] H. Schirmacher, W. Heidrich, and H.-P. Seidel, "High-quality interactive lumigraph rendering through warping," in *Proceedings of Graphics Interface*, Montreal, Quebec, Canada, 2000, pp. 87–94.

[6] C. Chang, G. Bishop, and A. Lastra, "Ldi tree: A hierarchical representation for image-based rendering," in *Proceedings of ACM SIGGRAPH'99*, Los Angeles, CA USA, 1999, pp. 291–298.

[7] A. Fitzgibbon, Y. Wexler, and A. Zisserman, "Image-based rendering using image-based priors," in *Proceedings of the International Conference on Computer Vision*, Beijing, China, 2003, pp. 1176–1183.

[8] Y.-Y. Chuang, A. Agarwala, D. H. Curless, B.and Salesin, and R. Szeliski, "Video matting of complex scenes," in *Proceedings of ACM SIGGRAPH'02*, San Antonio, Texas, 2002, pp. 243–248.

[9] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman, "Bayesian estimation of layers from multiple images," in *Proceedings of ECCV 2002*, Copenhagen, Denmark, 2002, vol. 3, pp. 487–501.

[10] S. W. Hasinoff, S. B. Kang, and R. Szeliski, "Boundary matting for view synthesis," in *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW04)*, Washington, DC, USA, 2004, p. 170.