MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

A Trellis-based Approach for Robust View Synthesis

Tian, D.; Vetro, A.; Brand, M.

TR2011-065 September 2011

Abstract

View synthesis is an essential function for a number of 3D video applications including free-viewpoint navigation and view generation for auto-stereoscopic displays. Depth Image Based Rendering (DIBR) techniques are typically applied for this purpose. However, the quality of the rendered views is very sensitive to the quality of the depth image. In this paper, a novel trellis-based view synthesis framework is proposed to overcome the above limitations in depth images and reduce artifacts in the rendered picture. Our results demonstrate that the proposed approach offers visible improvements in rendering quality compared to existing view synthesis techniques.

IEEE International Conference on Image Processing (ICIP)

A TRELLIS-BASED APPROACH FOR ROBUST VIEW SYNTHESIS

Dong Tian, Anthony Vetro, Matthew Brand
Mitsubishi Electric Research Labs
201 Broadway, Cambridge, MA 02139, USA

ABSTRACT

View synthesis is an essential function for a number of 3D video applications including free-viewpoint navigation and view generation for auto-stereoscopic displays. Depth Image Based Rendering (DIBR) techniques are typically applied for this purpose. However, the quality of the rendered views is very sensitive to the quality of the depth image. In this paper, a novel trellis-based view synthesis framework is proposed to overcome the above limitations in depth images and reduce artifacts in the rendered picture. Our results demonstrate that the proposed approach offers visible improvements in rendering quality compared to existing view synthesis techniques.

Index Terms— View synthesis, DIBR, 3D video

1. INTRODUCTION

Three-Dimensional Video (3DV) is the next milestone for the video industry after the massive deployment of High Definition TV (HDTV). A 3D display is capable of presenting a different view for each eye, thereby enabling a viewer to perceive the depth in a scene. In conventional stereo systems, the left and right views are acquired, compressed and either stored or transmitted, before being decoded and ultimately displayed. In more advanced systems, a virtual view from a different viewpoint than the existing input views may be synthesized to enable enhanced 3D features, e.g., free-viewpoint navigation or view generation for autostereoscopic displays.

Depth Image Based Rendering (DIBR) is a technique to synthesize virtual views, which typically requires the depth image of the scene available [1]. It is noted that depth images are likely to exhibit noise, either resulting from errors in computing stereo correspondences or through direct acquisition by range sensors; this noise may produce artifacts in the rendered views. On the other hand, per-pixel depth images cannot always represent depth discontinuities that typically occur at object boundaries, which is another source of artifacts in the rendered views [2].

This paper describes a novel framework that solves the view synthesis problem as a trellis-based optimization. This approach is inspired by trellis-based approaches to stereo disparity estimation, e.g., [3], but as in [4], the trellis costs are constructed to elicit the best synthesized image rather than the best disparity estimate. More specifically, a candidate set of depth values are identified for each sample that needs to be warped based on an estimated depth value for that sample as well as neighboring depth values. The cost for each candidate depth value is quantified based on an estimate of the synthesis quality, then the candidate with the best expected quality is selected.

The rest of this paper is organized as follows. The next section provides background on the view synthesis process. Our proposed trellis-based approach is presented in section 3, and experimental results are provided in section 4. Conclusions are given in section 5.

2. BACKGROUND ON VIEW SYNTHESIS

The process of view synthesis is composed of three steps [5], a warping step, in which the samples to the virtual position are warped from input views based on the geometry of the scene; a blending step, in which the warped pictures from each input viewpoint are combined into a single picture; and a hole filling step, in which any remaining holes in the blended picture are filled. The blending is only invoked when there are multiple input viewpoints from which the synthesized view is generated.

For the warping step, there are two categories of methods: forward warping and backward warping. With forward warping, the sample values in the reference view are mapped to a virtual view via a 3D projection process. However, with backward warping, the sample values in the reference view are not mapped to virtual view directly. Instead, the depth values are mapped to the virtual view first and the warped depth image is then used to find a corresponding sample value in the reference view for each sample location in the virtual view.

Most of the samples in the virtual view are mapped after the warping process. However, some samples do not have a corresponding value, which is caused by the disocclusion from one viewpoint to another. The samples without mapped values are known as holes in the virtual view. When there are multiple reference views, a blending process is used to merge the warping results into a single image. There are various ways to achieve this, e.g., a weighted average may be applied or one of the mapping values is selected

depending on the proximity of the virtual viewpoint location relative to the reference views.

Following the blending process, there are typically still some hole samples that have not been filled, and hence a final hole filling process is then required. In-painting techniques may be used to propagate the surrounding sample values to fill the remaining holes [6]. Other methods may simply propagate the background samples into the hole region if the hole is not very large.

It is well-known that the quality of the rendered views is very sensitive to the quality of the depth image, which is typically estimated through an error prone process. The goal of this work is to provide a more robust view synthesis framework to improve the picture quality of synthesized views in such a way that the synthesized view is free of boundary artifacts and geometrically consistent with the image characteristics that are present in the input viewpoints.

3. TRELLIS-BASED VIEW SYNTHESIS

A trellis-based view synthesis framework is proposed to overcome the limitations of depth images and reduce artifacts in the generated views. With this approach, a candidate set of depth values are identified for each sample that needs to be warped based on the estimated depth value for that sample as well as neighboring depth values. The cost for each candidate depth value is quantified based on an estimate of the synthesis quality, then the candidate with the best expected quality is selected. Our method is applied during the warping step of a DIBR scheme.

An example of a trellis-based structure for view synthesis is shown in Fig. 1. This trellis is constructed for a predetermined number of samples. In practice, one line of image samples are arranged into the trellis and the warping process is performed line-by-line in the image. That is, each column of the trellis represents one image sample. The nodes in each column of the trellis represent the candidate mappings for that sample in the virtual view picture.

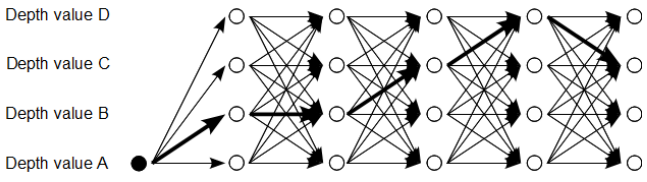


Fig. 1: Trellis based view synthesis

The trellis is used to indicate the set of depth candidates for each sample. The set includes the estimated depth value from the input depth image as well as several other candidates based on neighboring depth values; further details are provided in section 3.1. The number of depth candidates corresponds to the number of rows in the trellis. In the example shown in Fig. 1, each sample has four depth candidates corresponding to the four rows in the trellis. A cost function is used to estimate the synthesis quality and determine the best depth candidate, which is described further in section 3.2.

3.1. Determining the set of depth candidates

In our proposed method, the estimated depth value from the input depth image is always used as one of the candidate depth values. However, this value may be incorrect and lead to artifacts or inconsistencies with respect to the input images. Therefore, additional candidates are considered. A few methods that can be used to determine candidate depth values in this framework are elaborated below.

One method to determine the set of candidate depth values is with a predetermined increase and/or decrease relative to the estimated value from the input depth image. A second method, which we use in our implementation of this framework, is a predicted value based on the depth values from the neighbor samples. For example, the average or median value from neighboring depth values could be used. Our experiments have shown that the depth value from preceding samples of the same line is an effective candidate.

Obviously, an increase in the number of depth candidates will lead to an increase in complexity since each candidate need be evaluated and compared. So, the number of candidates should be selected judiciously.

3.2. View synthesis using dynamic programming

Once a set of depth candidates is determined, each node in the trellis is assigned a metric which estimates the synthesis quality. Then, the view synthesis problem is solved by finding an optimal set of depth values across the trellis. In this work, dynamic programming is used to solve the optimization problem.

In order to estimate the synthesis quality, a cost function is defined. The design of the cost function may depend on whether the warping process is forward warping or backward warping. Without loss of generality, we describe the definition of a cost function assuming backward warping in this paper. This definition is easily applied to forward warping as well.

The cost function in our current work is defined by the Mean Square Error (MSE) between two square blocks. The blocks are upper-left blocks relative to the current sample location. Let (x, y) denote the current sample location, (x', y') denote the warped position using a depth candidate. The first block is located at $(x\text{-}s, y\text{-}s) – (x, y)$ in the synthesized view, where s is the block size, and the second block is located at

(x'-s, y'-s) – (x', y') in the reference view. Cropping is applied if part of the block goes beyond the image area. Energy functions other than MSE of pixel values may also be used in this framework. For instance, the average absolute error is an effective cost function to estimate the synthesis quality. Also, image features can be extracted from the blocks of pixels and a matching process could be used to determine whether the blocks are geometrically consistent. The effectiveness of such methods is beyond the scope of the current work and left open for further study.

Furthermore, since any artifacts in the foreground objects are more easily perceived by human eyes, a method is needed to synthesize the foreground objects in a consistent manner. Therefore, the upper-left blocks are not always the best choice to compute the cost. A sample is classified into three types of areas: flat area, decreasing depth area and increasing depth area, as shown in Fig. 2. For samples at decreasing depth boundaries (right boundaries in Fig. 2) or flat areas, the upper-left block is used; whereas the upper-right block is used for samples at increasing depth boundaries (left boundaries in Fig. 2). In this way, the foreground samples have a higher priority in the quality evaluation and can be more consistently rendered.

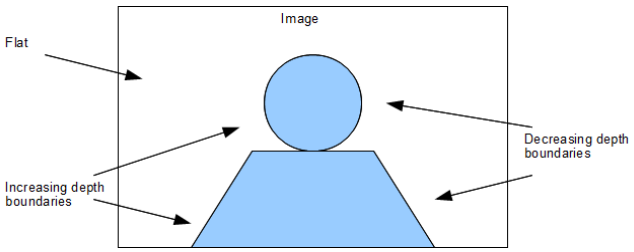


Fig. 2: Different cost functions for three types of areas

In some applications, a confidence map may also be available as an input to the synthesis process in addition to the estimated depth image. The cost function for the depth value from the depth image can then be weighted by a high factor when the depth estimator indicates a high confidence.

3.3. Practical considerations

The complexity of the view synthesis is dependent on the number of depth candidates, the ways in which the depth candidates are determined, as well as the form of the cost function. This section considers two practical designs based on the options described in sections 3.1 and 3.2.

The first design assumes a local optimization and hence has limited complexity. With this design, depth candidate selection is not permitted to depend on the selection of the best depth candidates from previous samples. As a result, the depth candidate assignment and evaluation can be performed in parallel for each sample within one line of the image. A brief description of this design is provided below.

First, three depth candidates, labeled A, B and C, for all samples in the trellis are identified.

- A: estimated depth value from the input image.
- B: depth value of a sample in the same row of the image that has a maximum difference with the value of A. In our simulations, the four previous samples to the left of the current sample are checked.
- C: depth value of a previous sample in the same column of the image that has a maximum difference with the value of A. In our simulations, the four previous samples directly above the current sample are checked.

Given these candidate values, the cost for each depth candidate of each sample is evaluated based on the technique described in section 3.2. Finally, the costs of all the depth candidates for each sample are compared and the corresponding depth value with minimum cost is selected.

In the second design, the best depth values that have been selected for previous samples are utilized in determining the depth candidates for the current sample. In this way, the best depth values from previous samples are used to derive depth candidates B and C, which may lead to a value different from what is signaled in the depth image. The main drawback of this approach is that the warping process cannot be done in parallel. However, improved synthesis quality can be achieved. In the simulation results that are shown in the next section, we compare this design against a reference method.

It is noted that more advanced designs for the trellis-based framework are possible given sufficient resources. In the above two designs, the cost is evaluated and compared per node, which is a local optimization strategy. Alternatively, it is possible to evaluate the cost of each candidate path through the trellis to find the globally optimal path. A node may be assigned with a different depth candidate value when it is crossed by a different path, and hence such a global optimization may involve very high complexity. Further study is needed to find out an efficient way to solve the global optimization problem. Trellis pruning and parallel processing, which have proven effective in stereo disparity estimation, may be considered.

4. SIMULATIONS

We implemented the synthesis methods discussed in this paper within a proprietary system that estimates depth from a stereo pair and utilizes backward warping. For the anchor method, the depth value signaled in the depth images are always used during the warping process. On the other hand, the trellis based synthesis is implemented in the way that the

best depth from previous samples are carried for predicting the next depth, as described in section 3.3. Since there is a lack of commonly used objective metrics for synthesis quality especially when there is no ground truth as reference, sample rendering results are provided to demonstrate the improved subjective quality.

Simulations were conducted over several sequences, including Kendo from Nagoya University, Newspaper and Cafe from GIST, as well as Poznan_Hall1 from Poznan University. For each sequence, a virtual view is synthesized at 1/3 of the baseline distance from the left view in-between the stereo pair.

Figs. 3-6 show sample images of the synthesized results. The left images correspond to the anchor method, where the estimated depth values from depth images are always used; while the right correspond to the proposed trellis-based synthesis method as described in section 4. It is observed that the artifacts can be significantly reduced with the trellis-based method. For example, there are much fewer artifacts along the boundary of the face mask in Kendo (Fig. 3), calendar in Newspaper (Fig. 4), face edge and background text in Cafe (Fig. 5), as well as the wall boundary and logo in Poznan_Hall1 (Fig. 6).

5. CONCLUSIONS

In this paper, a novel view synthesis framework based on trellis optimization is proposed. This framework is able to improve the subjective quality of generated views by coping with errors in the estimated or acquired depth maps and overcomes the limitation of a single depth value representation along depth discontinuities. Simulation results demonstrate visible reduction in artifacts and satisfactory view synthesis results.

Further improvements within this framework could also be done. For instance, global optimization through the trellis is a topic for further study. The extension to video with enforcement of temporal consistency is another area to consider. The cost function may also be designed to better detect geometric inconsistencies with the reference images.

REFERENCES

[1] H.Y. Shum, S.B. Kang, and S.C. Chan, "Survey of Image-Based Representations and Compression Techniques," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 11, November 2003.

[2] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered depth images," *Proc. SIGGRAPH'98*, Orlando, FL, July 1998.

[3] S. Birchfield and C. Tomasi, "Depth Discontinuities by Pixel-to-Pixel Stereo," *International Journal of Computer Vision*, 35(3): 269-293, December 1999.

[4] M.E. Brand, "Image and Video Retargetting by Darting," *Image Analysis and Recognition*, Vol. 5627, July 2009.

[5] D. Tian, P.L. Lai, P. Lopez, and C. Gomila, "View synthesis for 3D video," *Proc. SPIE*, vol. 7443, August 2009.

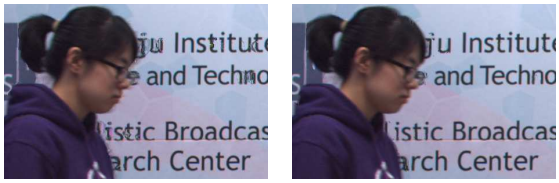
[6] K.-J. Oh, S. Yea, and Y.-S. Ho, "Hole filing method using depth based in-painting for view synthesis in free viewpoint television and 3-D video", *Proc. Picture Coding Symposium*, Chicago, IL, May 2009.

Left: Conventional Right: Trellis-based

Fig. 3: Sample view synthesis results of *Kendo*.

Left: Conventional Right: Trellis-based

Fig. 4: Sample view synthesis results of *Newspaper*



Left: Conventional Right: Trellis-based

Fig. 5: Sample view synthesis results of *Cafe*

Left: Conventional Right: Trellis-based

Fig. 6: Sample view synthesis results of *Poznan_Hall1*