MULTIPLE IMAGE VIEW SYNTHESIS FOR FREE VIEWPOINT VIDEO APPLICATIONS

Eddie Cooke, Noel O'Connor

Centre for Digital Video Processing, Dublin City University, Dublin, Ireland
{ej.cooke, oconnorn}@eeng.dcu.ie

ABSTRACT

Interactive audio-visual (AV) applications such as Free Viewpoint Video (FVV) aim to enable unrestricted spatio-temporal navigation within multiple camera environments. Current virtual viewpoint view synthesis solutions for FVV are either purely image-based implying large information redundancy; or involve reconstructing complex 3D models of the scene. In this paper we present a new multiple image view synthesis algorithm that only requires camera parameters and disparity maps. The Multi-View Synthesis (MVS) approach can be used in any multi-camera environment and is scalable as virtual views can be created given 1 to *N* of the available video inputs, providing a means to gracefully handle scenarios where camera inputs decrease or increase over time. The algorithm identifies and selects only the best quality surface areas from available reference images, thereby reducing perceptual errors in virtual view reconstruction. Experimental results are presented and verified using both objective (PSNR) and subjective comparisons.

1. INTRODUCTION

Increasingly audio-visual (AV) applications are being designed to enable dynamic viewing of a captured event or scene [10]. Such applications allow users to freely navigate within the AV scenes by specifying arbitrary viewpoints and orientations. To this end, the Moving Picture Experts Group (MPEG) of ISO/IEC recently formed the 3DAV work group to investigate the requirement for standardization in this area [11]. Of the AV application scenarios being investigated they identified *Free Viewpoint Video* (FVV) as being the most challenging. FVV enables unrestricted spatio-temporal navigation within a scene captured using a multiple camera setup such as that depicted in Fig. 1.

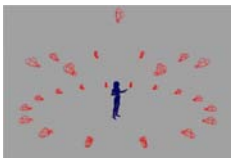


Fig. 1. Free viewpoint video image capturing environment.

All visual information associated with a scene in 3-space can be described by the plenoptic function [1]. This 7D structure may be interpreted as a scene representation method and therefore enables the accurate reconstruction of any arbitrary viewpoint in the 3D world. However, due to the enormous amounts of information involved, it is not possible to capture and store a scene's entire plenoptic function. Chai *et. al.* [2] attempt to resolve this by examining the relationship between the number of samples of the plenoptic function and the amount of geometrical information required to generate a continuous representation of the function. They determined that plenoptic sampling could be used to design an optimized multi-camera setup for image-based rendering (IBR) systems.

Currently available IBR virtual viewpoint view synthesis solutions tend to be either navigationally restrictive and setup dependent [3] or take such a general approach that all the perceptual errors across the available reference views are incorporated into the final virtual view [7]. In this paper we present a new view synthesis algorithm that identifies and selects only the best quality surface areas from available reference images for virtual view reconstruction. The approach can be used in any multi-camera environment and is scalable as virtual views can be created given 1 to *N* of the available video inputs.

This paper is organised as follows: In Section 2 we discuss virtual viewpoint synthesis and current FVV synthesis solutions. Multi-View Synthesis (MVS) a new multiple image view synthesis algorithm is described in Section 3. In Section 4 experimental results are presented and verified by comparison with ground truths. Finally, in Section 5 conclusions and future work are described.

2. VIRTUAL VIEWPOINT SYNTHESIS

IBR systems generate novel views via reference images and corresponding scene geometry information [8]. The virtual view rendered at the desired viewpoint is created using an associated view synthesis process. Virtual view synthesis algorithms are designed for environments where the scene interaction area is well defined and can be captured using an *N* camera setup. In immersive scenarios where the cameras are calibrated and each reference image has a corresponding disparity map there are a number of available view synthesis approaches [6]. These are mainly restricted to combining scene information from a stereo camera setup. Restricting novel view creation to a fixed number of images simplifies the view synthesis process. However, it implies that during scene navigation the best reference image view of the surfaces required for the current virtual viewpoint may be ignored. In contrast, existing multiple image view synthesis approaches combine all available surface information, usually via view orientation weighting, in order to create virtual views [7]. This implies that although more original information is available, errors due to occlusions, depth mismatches etc; are incorporated into the final virtual view. Current FVV view synthesis solutions are based on using either images with no scene geometry information that are obtained via a dense sampling of the scene, implying large information redundancy [12], or reconstructing complex 3D models of the scene from a sparse camera setup [9].

0-7803-9134-9/05/$20.00

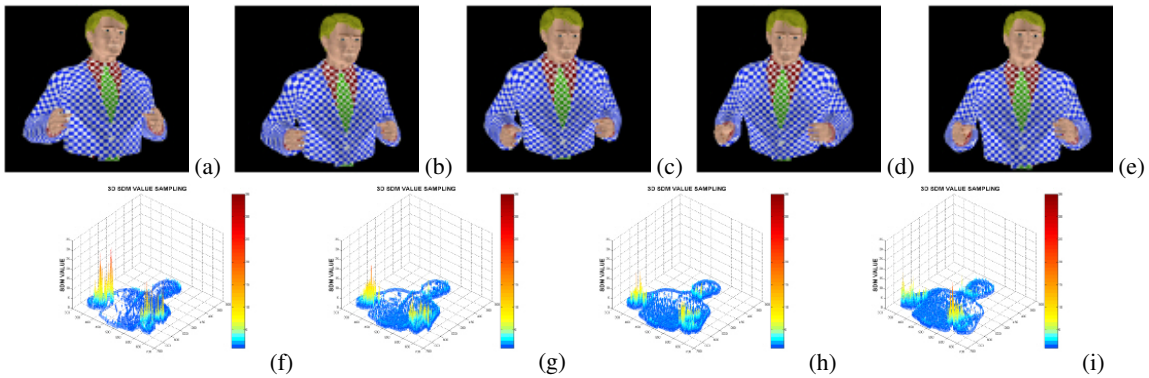


Fig. 2. (a)-(d) Reference images A, B, C and D. (e) Ground truth. (f)-(i) 3D visualisation of SDM after warping (a)-(d) respectively.

3. MULTIPLE IMAGE VIEW SYNTHESIS

A restrictive version of the MVS algorithm that used only image pairs was developed for immersive teleconferencing environments [4]. Here we describe a new algorithm that provides a scalable and flexible approach to *multiple image* view synthesis for FVV [5]. The approach is scalable since it can process an arbitrary number of reference images making it suitable for both sparse and dense camera sampling. At the core of MVS is a flexible definition of what constitutes a 3D scene surface. This flexibility ensures that the best quality virtual view reconstruction is produced from the available views. A common problem with objective evaluation of view synthesis approaches is that there are no original views to compare with the virtual viewpoint results. For this reason Fig. 2(a)-(d) presents four different camera views taken within a test environment that will be used to illustrate the MVS approach. Fig. 2(e) is taken from a camera placed at the position of the final virtual view and will be used as a ground truth to indicate the correctness of the approach. The algorithm consists of four phases which are now described in detail:

3.1. Identification

The surface identification process determines surface areas within reference images after they have been warped to the virtual viewpoint. This process is based on the following observations:

- Corresponding 3D scene points visible across the reference images are warped to the same position in the virtual view.
- These matching virtual view areas/surfaces have varying 2D sampling densities within the warped images.

We now describe how the sampling of the 3D surfaces at the virtual viewpoint is used as a criterion for virtual view surface quality identification. Fig. 3(a) presents a reference image containing two distinct surfaces in 3-space: π_1 (vertical lines) and π_2 (horizontal lines). Sample $\mathbf{q}$ along with its neighbouring samples $\mathbf{q}_1 - \mathbf{q}_4$ and $\mathbf{q}_6 - \mathbf{q}_8$ belong to surface π_1, while sample $\mathbf{q}_5$ lies on π_2. This 3×3 block is the reference image *local surface*. Fig. 3 (b) and (c) illustrate these surface samples after a warp to two distinct virtual viewpoints, where point $\mathbf{q}_i$ is warped to $\mathbf{q}_i'$.

The displacement between a sample $\mathbf{q}' = [x', y']^T$ and any of its warped reference image neighbours $\mathbf{q}_i' = [x_i', y_i']^T$ is determined via Eq. (1):

$$\rho_{\mathbf{q}'}(\mathbf{q}_i') = \sqrt{(x_i' - x')^2 + (y_i' - y')^2} \tag{1}$$

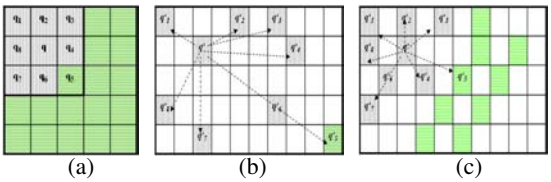


Fig. 3. (a) Two distinct surfaces in 3-space. (b)&(c) The warped surface's sampling density indicates virtual view surface quality.

A measure of the extent of the displacement of $\mathbf{q}'$ with respect to the warped reference local surface, referred to as its *sampling density value*, $\delta(\mathbf{q}')$, is computed using the sampling density function Eq. (2):

$$\delta(\mathbf{q}') = \frac{\sum_{i=1}^{N} \rho_{\mathbf{q}'}(\mathbf{q}_i')}{N} \tag{2}$$

Where N is the number of sample neighbours of $\mathbf{q}$. A sampling density value is computed for every pixel sample in the warped reference image and stored in a corresponding *Sampling Density Map* (SDM). The higher the value at a sample the larger the displacement of its original reference image surface neighbours in the warped view. High values in the SDM imply the surface is either undersampled in the reference image or that it lies on different sides of a depth discontinuity visible from the virtual viewpoint. The SDM enables each warped reference image to define its own representation of the 3D surfaces required at the virtual view position. Returning to the example, Fig. 3(b) depicts the scenario where the local surface $\mathbf{q}_1 - \mathbf{q}_8$ around $\mathbf{q}$ is sparsely sampled in the warped reference image. Hence, the SDM contains a high sampling density value at $\mathbf{q}'$ for this virtual view, reflecting its sparse sampling. In contrast, Fig. 3(c) presents a virtual view where after warping the local surface displacements are relatively constant irrespective of the actual 3D surface represented, π_1 or π_2. Therefore, the SDM for this virtual viewpoint contains a low sampling density value at $\mathbf{q}'$, reflecting its dense sampling. This surface identification approach dynamically groups reference image warped samples into surfaces of similar sampling densities.

Fig. 2(f)-(i) illustrates a 3D visualisation of the SDM of reference images A, B, C and D after they have been warped to the virtual viewpoint. Each sampling density value is presented as a height allowing easy identification of surface areas with extremely sparse sampling, the colour-bar indicates the actual sampling density value at a point.

3.2. Selection

The surface selection process determines the surface areas across the available surface representations to be used for view synthesis. The selection process is based on a sampling density weighting scheme. This weight, α^{δ}, is computed from the sampling density values contained in the SDM of the reference image virtual view surface representations. Examining the SDM of the two reference images A and B the surface sample at position $\delta^{A}_{(i,j)}$ in SDM A and position $\delta^{B}_{(i,j)}$ in SDM B detail the respective reference image's sampling density for the same virtual view position. Hence, a sampling density weighting function for the surface sample Q in the SDM of image A, $\hat{\alpha}^{\delta^A}_{Q}$, can be computed via Eq. (3):

$$\hat{\alpha}^{\delta^A}_{Q} = 1 - \frac{\delta^{A}_{Q}}{\sum_{i=1}^{N} \delta^{i}_{Q}} \tag{3}$$

where N is the number of reference images. Normalising the surface weight across the N SDMs to sum to unity via:

$$\alpha^{\delta^A}_{Q} = \frac{\hat{\alpha}^{\delta^A}_{Q}}{\sum_{i=1}^{N} \hat{\alpha}^{\delta^i}_{Q}} \tag{4}$$

The approach to surface selection is to loop through the N available SDM selecting a non-processed surface area with the highest α^{δ} weight on each pass Eq. (5):

$$\Lambda_{Q} = \Phi(\alpha^{\delta^i}_{Q}) \tag{5}$$

Where Λ_Q represents the final virtual view surface sample at Q, the $\alpha^{\delta^i}_{Q}$ weight is a measure of the sampling density at surface sample Q based on reference view i, where i ranges from $[1, N]$, and Φ is a function which selects the surface sample associated with the maximum weight. This ensures a virtual view related surface selection as opposed to a strict depth or texture surface identification. The approach dynamically groups local sample neighbourhoods into surfaces of similar sampling densities at the virtual viewpoint.

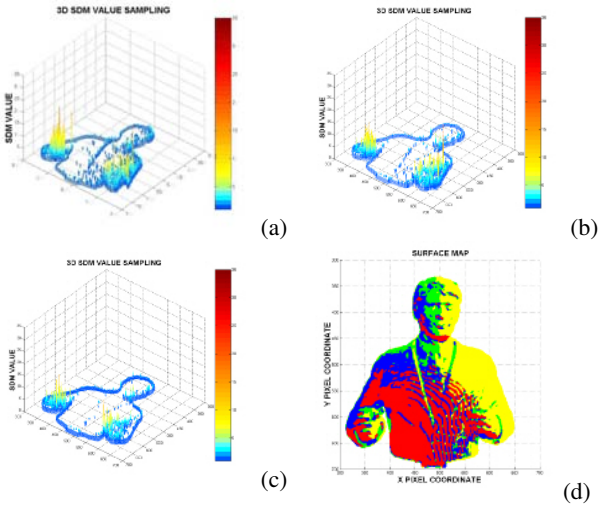


Fig. 4. (a)-(c) Incrementally improved virtual view SDM. (d) Surface Map for surface selection of (c).

Returning to our example Fig. 4(a)-(c) illustrates the virtual view MVS surface selection as reference images are incrementally added to the process. Fig. 4(a) presents the surface sampling selection when only images A and B are used. The SDM is smoother than either Fig. 2(f) or (g), those of images A and B respectively, indicating that the surface sampling is denser and therefore the virtual view surface quality has improved. Fig. 4(b) and (c) illustrate the added improvement in surface quality as images C and D, respectively, are included in the surface selection process.

3.3. Boundary Blending

Integrating the best view of each required surface implies that neighbouring view synthesis surfaces may be supplied from different images. When this occurs, specularity and other photometric differences across the reference images can cause perceptual seams to appear in the virtual view. A *Surface Map* is designed to indicate from which reference view the chosen virtual view surfaces originate. Fig. 4(d) indicates the chosen surface samples from images A, B, C and D for the virtual view defined in Fig. 4(c). Here, the colour red represents surfaces from image A, blue from image B, green from image C and yellow from D. In order to lessen the perceptual seams a weighted blending is implemented on an extended boundary region (e.g. five pixels) around the connected surfaces. This blending is designed to ensure that the following conditions are met:

- Non-boundary regions are taken exclusively from the best surface representation as dictated by the surface selection.
- Boundary regions are generated by blending surface sample values from bordering surfaces.

A fixed linear ramp blending is used in these overlapping areas to compute texture values for the final virtual view.

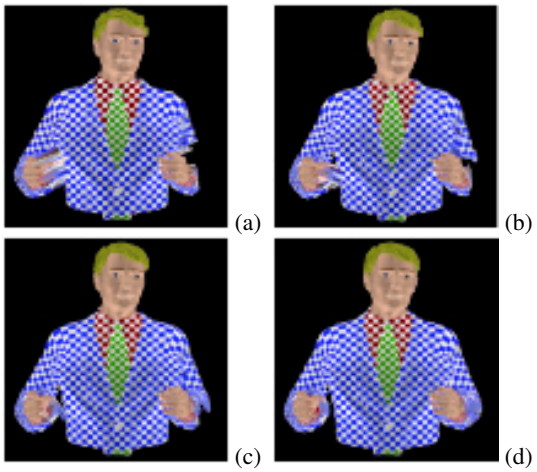
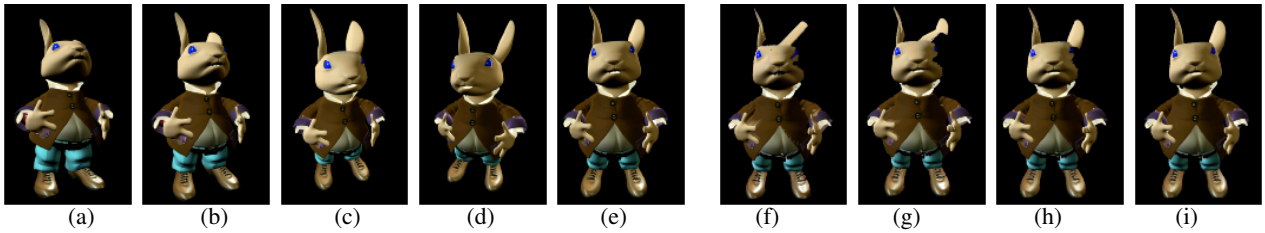


Fig. 5. MVS results corresponding to Fig. 2(f) & Fig. 4(a)-(c).

3.4. Reconstruction

The MVS view synthesis approach creates virtual views using an arbitrary number of reference images. Therefore, the surface reconstruction process must resolve issues of surface visibility and hole filling across multiple images. *Surface Visibility*: when reconstructing the 3D scene captured by an individual reference image occlusion errors due to background surfaces overwriting foreground may arise. These are resolved by implementing a back-to-front surface warping order [3]. Surface visibility issues arising



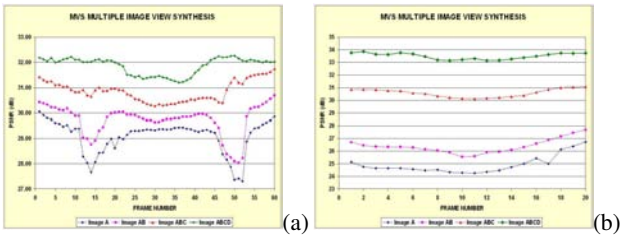
(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 6. (a)-(d) Images A,B,C and D. (e) Ground truth. (f)-(i) View synthesis results as A, B, C and D are incrementally added to MVS.

from multiple views of the same surface across the reconstructed reference images are resolved via the surface selection weighting scheme. *Hole Filling*: involves identifying areas within warped surfaces where virtual view required surface information is missing. Holes arising due to sampling gaps in continuous surfaces are filled using interpolation. While surface disclosures, which arise due to the movement of a foreground object with respect to the background, are filled using the MVS surface selection approach. This approach ensures that all the surfaces across the reference images are considered for hole filling and therefore reduces perceptual errors during surface reconstruction.

4. EXPERIMENTAL RESULTS

In order to demonstrate that the MVS algorithm provides both a scalable and flexible approach to view synthesis we now present results from two test sequences with ground truths. Reference images are incrementally added to the view synthesis process to illustrate that the approach can be used in arbitrary multi-camera FVV setups and to indicate that perceptual errors decrease as more reference images are available.



(a) (b)

Fig. 7. Improvements in PSNR as images added to MVS.

Fig. 7(a) presents the PSNR measures between the MVS constructed virtual view and the ground-truth over the 60 frame test sequence of Fig. 2. The graph details how the PSNR improves as reference images are added to the MVS process; from an average of 29dB when just one image is used to 32dB when using all four. A subjective result is provided in Fig. 5(a)-(d). Fig. 6(a)-(d) presents a frame from the four camera inputs of the second test sequence; Fig. 6(e) is the associated ground truth. Again the corresponding PSNR Fig. 7(b) details an improvement from an average of 25dB when just one image is used to an average of 33dB using all four. A subjective result is provided in Fig. 6(f)-(i).

5. CONCLUSIONS & FUTURE WORK

Current view synthesis solutions for FVV are either purely image-based, implying a large amount of information redundancy, or involve reconstructing complex 3D models of the scene. The presented multiple image view synthesis MVS algorithm provides a scalable and flexible approach that only requires camera calibration and disparity maps. Unlike other view synthesis solutions MVS identifies and selects only the best quality surface areas from the set of available reference images, thereby reducing perceptual errors in virtual view reconstruction. The approach is camera setup independent and scalable as virtual views can be created given 1 to *N* of the available video inputs. Thus, MVS provides interactive FVV applications with a means to handle scenarios where camera inputs increase or decrease over time. Experimental results were presented and verified using both objective (PSNR) and subjective comparisons. Future work includes a pre-processing step to identify from a very large number of available video input streams only those input streams that contain surfaces required for the current virtual viewpoint.

6. REFERENCES

[1] E. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. MIT Press, 1991.

[2] J. Chai, X. Tong, S. Chan, and H. Shum, "Plenoptic sampling," in *Proc of ACM SIGGRAPH*, 2000.

[3] E. Cooke, P. Kauff, and O. Schreer, "Image-based rendering for tele-conference systems," in *Proc WSCG*, 2002.

[4] E. Cooke, I. Feldmann, P. Kauff, and O. Schreer, "A modular approach to virtual view creation for a scalable immersive teleconferencing configuration," in *Proc ICIP*, 2003.

[5] E. Cooke, "Multiple image view synthesis for virtual viewpoint rendering," ISO/IEC MPEG, M11238, Oct 2004.

[6] C. Fehn, E. Cooke, O. Schreer, and P. Kauff, "3d analysis and image-based rendering for immersive tv applications," *Image Communication Journal*, vol. 17(9), 2002.

[7] K. Müller *et. al.* "Multi-texture modelling of 3d traffic scenes," in *Proc ICME*, 2003.

[8] H. Shum and S. Kang, "A review of image-based rendering techniques," in *Proc VCIP*, 2000.

[9] A. Smolic *et. al.* "Free viewpoint video extraction, representation, coding, and rendering," in *Proc ICIP*, 2004.

[10] A. Smolic and P. Kauff, "Interactive 3d video representation and coding technologies," *IEEE Special Issue on Advances in Video Coding and Delivery*, Dec 2004.

[11] A. Smolic and D. McCutchen, "3dav exploration of video-based rendering technology in mpeg," *IEEE TCSVT*, 2004.

[12] M. Tanimoto, "Free viewpoint television - ftv," in *Proc PCS*, 2004.