Scalable Virtual Viewpoint Image Synthesis for Multiple Camera Environments

E. Cooke, N. O'Connor
Centre for Digital Video Processing, Dublin City University, Dublin, Ireland
ejcooke@eeng.dcu.ie, oconnorn@eeng.dcu.ie

Abstract

One of the main aims of emerging audio-visual (AV) applications is to provide interactive navigation within a captured event or scene. This paper presents a view synthesis algorithm that provides a scalable and flexible approach to virtual viewpoint synthesis in multiple camera environments. The Multi-View Synthesis (MVS) process consists of four different phases that are described in detail: surface identification, surface selection, surface boundary blending and surface reconstruction. MVS view synthesis 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 AV applications with a means to handle scenarios where camera inputs increase or decrease over time.

1. Introduction

One of the main aims of emerging audio-visual (AV) applications is to remove the *passiveness* of viewing a captured event or scene [8]. Such systems encourage users to explore and navigate AV scenes by allowing them to specify their own viewpoint and orientation. Recently, the Moving Picture Experts Group (MPEG) of ISO/IEC formed a working group, called 3DAV, to investigate the need for standardization in this area [9].

Theoretically, all the visual information associated with a 3D scene can be specified using the 7D plenoptic function [1]. However, in practical terms it is not possible to entirely define this scene representation method. Instead, captured reference images of a scene can be regarded as a sparse set of samples of the complete plenoptic function. The more samples provided, the less additional scene geometry information required for image-based novel view creation. Image-based rendering (IBR) techniques create virtual views from a set of scene images with associated point correspondences [7]. They are most suited to an environment where the scene interaction area is well defined and can be captured using an N camera setup.

The process of identifying and combining information from captured images to create novel views is called view synthesis. A number of restrictive approaches exist for selecting which camera in a stereo setup best captures a surface and how these surfaces should be combined [2]. However, these techniques fail when migrated to an arbitrary multi-camera environment. Existing multiple image view synthesis approaches avoid surface identification and simply combine all available surface information, usually via view orientation weighting, in order to create virtual views [6]. This implies that although more original information is available, errors due to occlusions, depth mismatches etc; are incorporated into the final virtual view. In this paper we present a new view synthesis approach that identifies and selects only the best quality surface areas from available reference images, therefore 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.

The layout of the paper is as follows: a new scalable virtual view synthesis algorithm called *Multi-View Synthesis* (MVS) is described in Section 2. In Section 3 experimental results are presented and verified by comparison with ground truths. Finally, in Section 4 conclusions and future work are described.

2. Scalable Virtual Viewpoint Synthesis

Of the AV application scenarios being investigated by 3DAV the most challenging is that of *Free Viewpoint Video* (FVV) [9]. FVV applications are designed to allow unrestricted scene navigation and are captured using a multiple camera setup. Current view synthesis solutions in FVV 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 [10], or reconstructing complex 3D models of the scene from a sparse camera setup [4].

Proceedings of the Ninth International Conference on Information Visualisation (IV'05)
1550-6037/05 $20.00 © 2005 IEEE

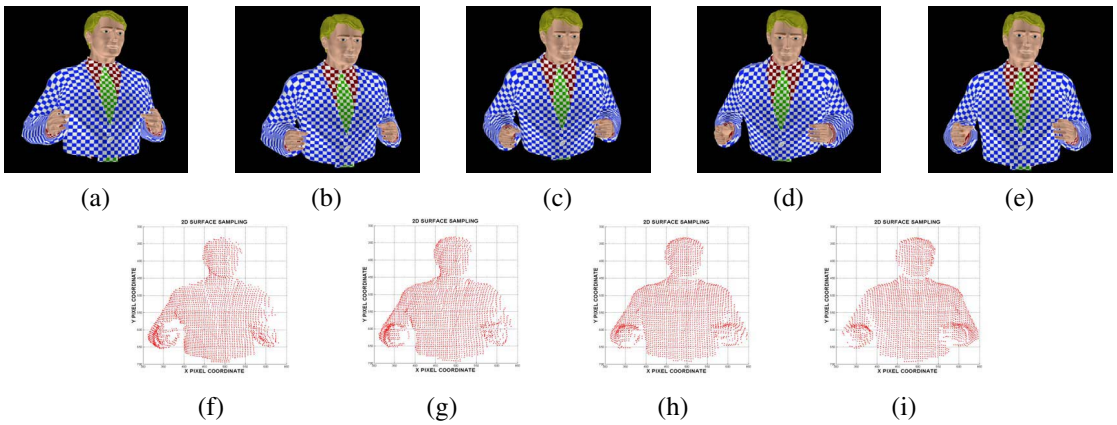


Figure 1. BUSINESS-MAN (a)-(d) Reference images A, B, C and D. (e) Ground truth. (f)-(i) 2D Surface Sampling after warping (a)-(d) respectively.

The MVS algorithm provides a scalable and flexible approach to view synthesis for FVV. MVS is scalable as it's designed to handle an arbitrary number of reference images and is therefore 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. The MVS process is a virtual view dependent approach implying that it examines the reference images at the virtual viewpoint before the surface identification and selection process occurs. It requires the camera setup to be calibrated and that each reference image has a corresponding disparity/depth map [3].

Fig. 1(a)-(d) presents four different camera views taken within a test environment that will be used to illustrate the MVS approach. Fig. 1(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.

2.1. Surface Identification

The first step in MVS surface identification is to warp the reference images to the virtual view position. After image warping, surface holes arise due to both sampling gaps and surface disclosures [2]. The 2D displacement across the surfaces is a measure of the extent of the sampling of the 3D surfaces within the reference image and can be used as criteria for virtual view surface quality recognition. Fig. 1(f)-(i) illustrates the 2D sampling of the surface points (every fifth sample) at the virtual viewpoint for reference images Fig. 1(a)-(d), respectively. Examining the 2D surface sampling we can determine the following:

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

Therefore, each reference image defines its own representation of the 3D surfaces required at the virtual view position. How we identify their surface quality is determined using the *Sampling Density Map* (SDM).

Let q be a surface sample in the reference image that is warped to position q' in the virtual view. Let $q_1 - q_8$ be the set of reference image neighbours of q defined by a 3×3 block centred at q, we call these samples the local surface of q. The displacement between a sample $q' = [x', y']$ and any of its warped reference image neighbours $q'_i = [x'_i, y'_i]$ can be determined via Eq. (1):

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

A measure of the extent of the displacement of q' with respect to its warped local surface is computed using the *sampling density function*:

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

where N is the number of reference image sample neighbours of q in the local surface. A sampling density value is computed for every pixel sample in the warped reference image and stored in the SDM. The higher the value at a sample the larger the displacement from its local surface in the warped view. Implying the surface is either undersampled in the reference image or that the sample's reference image neighbourhood lies on different sides of a depth discontinuity visible from the virtual viewpoint.

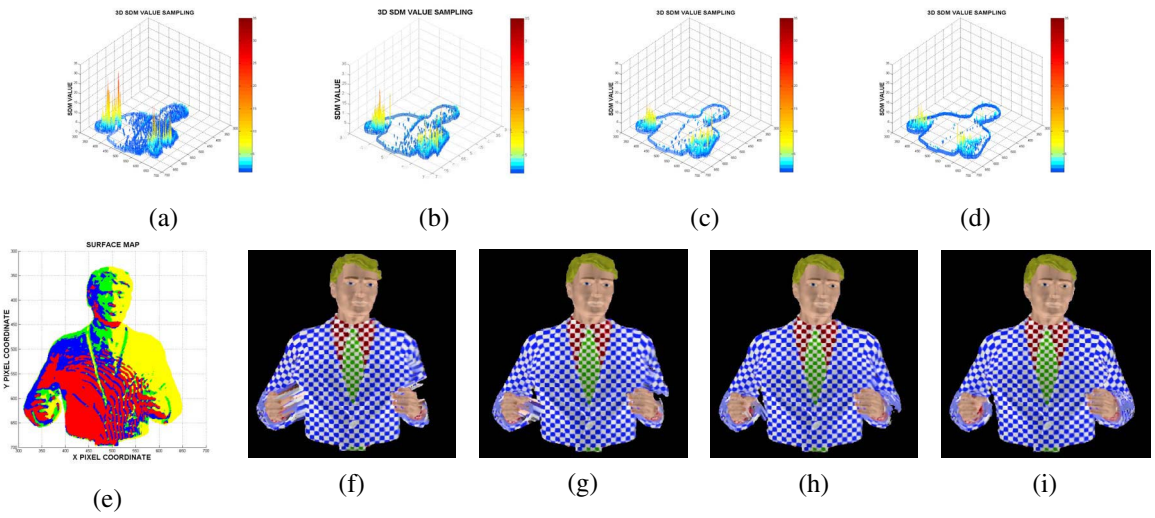


Figure 2. (a)-(d) Improvement in virtual viewpoint SDM as reference images added incrementally. Colour-bar indicates SDM values.(e) Surface Map for surface selection of (d). (f)-(i) Surface reconstruction corresponding to surface selections of (a)-(d).

2.2. Surface Selection

The MVS surface selection process determines the surfaces across the available surface representations to be used for view synthesis. This selection process is based on a sampling density weighting scheme. We have previously noted that corresponding 3D scene points or surfaces visible in more than one reference image are warped to the same position in the virtual view. Therefore, examining the SDM of two reference images A and C, we know that the surface sample at position $\delta^A_{(x,y)}$ the SDM of A and position $\delta^C_{(x,y)}$ in the SDM of C indicate the reference image's sampling density for the same virtual view position. Hence, we can compute a sampling density weight function for the surface sample Q in the SDM of image A, $\hat{\alpha}^{\delta^A}_Q$, 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. We then normalise the surface sample sampling density weight across the 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 SDMs selecting a non-processed surface with the highest α^δ weight on each pass Eq. (5):

$$\Lambda_Q = \Phi(\alpha^{\delta^k}_Q) \tag{5}$$

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. Although the approach is sample based it dynamically groups local sample neighbourhoods into surfaces of similar sampling densities. This ensures a virtual viewpoint related surface division as opposed to a strict depth or texture surface identification.

Fig. 2(a) illustrates a 3D representation of the SDM at the virtual viewpoint when surface samples are taken exclusively from the warped reference image A (Fig. 1(a)). Each SDM value is presented as a height, the higher the value the more sparse the sampling, Fig. 1(f). In Fig. 2(b) reference image B is added to the surface selection process. It can be determined that this new SDM is smoother, as the sampling density spikes have decreased, indicating that the surface sampling is denser and therefore the virtual view surface quality has improved. Fig. 2(c) and (d) illustrate the added improvement in surface quality as images C and D, respectively, are included in the surface selection process. A denser surface sampling implies an increase in surface quality at the virtual view.

2.3. Surface Boundary Blending

Integrating only the best view of each required surface implies that neighbouring surfaces in the final virtual view may be supplied from different reference images.

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

Figure 3. RABBIT (a)-(d) Reference images A, B, C and D. (e) Ground truth. (f)-(i) Improvement in view synthesis results from incrementally adding images A, B, C and D respectively to the MVS process.

Fig. 2(e) illustrates a *Surface Map* indicating the chosen surface samples from images A, B, C and D for the virtual view defined in Fig. 2(d). A surface map is designed to indicate from which reference view the chosen virtual view surfaces originate. Here, the colour red represents surfaces from image A, blue from image B, green from image C and yellow from D. At these surface boundaries, specularity and other photometric differences across the images can cause perceptual seams to appear in the virtual view. In order to lessen this effect a weighted blending is implemented on an extended boundary region (e.g. five pixels) around the connected surfaces. We use a fixed linear ramp blending in these overlapping areas to compute colour values for the final virtual view.

2.4. Surface Reconstruction

View synthesis surface reconstruction deals with two issues: those of surface visibility and hole filling during novel view creation at the virtual viewpoint.

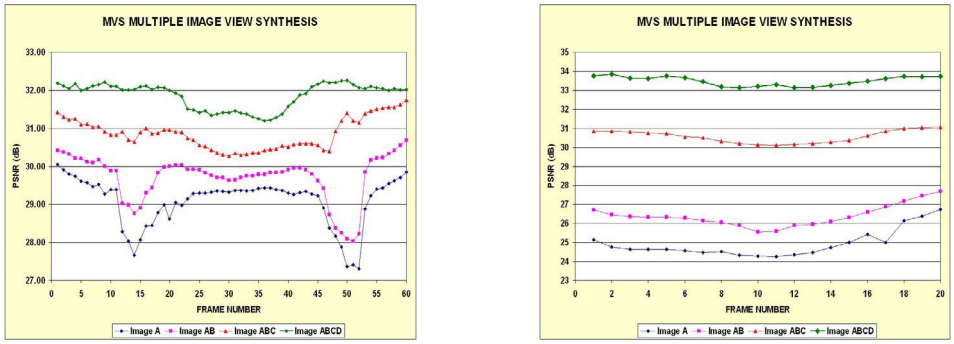
Resolving surface visibility occurs at two levels: within each warped reference image during surface identification and across the N images used during view synthesis. To solve the former issue we implement a back-to-front warp order [5] to avoid surface occlusion errors. During multiple image view synthesis the surface selection weighting scheme, which identifies only one reference image for a required surface area, resolves any visibility issues.

View synthesis hole filling involves identifying areas within warped surfaces where virtual view required surface information is missing. There are two different types of holes. Smaller 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 N reference images are considered for hole filling and therefore reduces perceptual errors during surface reconstruction.

3. Experimental Results

The MVS algorithm provides a scalable and flexible approach to view synthesis. To indicate the correctness of the approach we compare the view synthesis results from the two test sequences illustrated in Fig. 1 and Fig. 3 with their respective ground-truths. In order to demonstrate the scalability of the MVS view synthesis method we incrementally add reference images to the view synthesis process and compare all the resulting virtual views. This allows a direct comparison of the improvements in virtual view reconstruction as more reference images are added to the MVS process.

It also indicates the flexibility of the view synthesis approach as we can clearly determine that the approach can gracefully deal with both a decrease and increase in camera inputs. Fig. 4(a) presents the results of PSNR measures between the MVS view synthesis and the ground-truth over the 60 frame `BUSINESS-MAN` test sequence. The graph details how the PSNR improves as reference images are added to the view synthesis process. This improvement is from an average of 29dB for the sequence when just one image is used to an average of 32dB when using all four. A subjective result is provided in Fig. 2(f)-(i). The second test sequence is the 20 frame `RABBIT` sequence. Fig. 3(a)-(d) presents a frame from the four camera inputs, Fig. 3(e) is the ground truth. The PSNR measures between the MVS view synthesis and the ground-truth are presented in Fig. 4(b). The graph details how the PSNR improves as reference images are added to the view synthesis process. This improvement is 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. 3(f)-(i).



(a) (b)

Figure 4. Graph of scalable view synthesis improvements (PSNR) in (a) BUSINESS-MAN (b) RABBIT test sequence.

4. Conclusions and Future Work

In this paper we discussed the shortcomings of current multiple image view synthesis approaches. We then presented the MVS algorithm that provides a scalable and flexible approach to view synthesis in multiple camera environments.

The process consists of four different phases that are described in detail: surface identification, surface selection, surface boundary blending and surface reconstruction. 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 AV 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.

Ideas for future work include: 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 synthesis; and incorporating a system's currently available bandwidth and processing power into the scalability process.

5. Acknowledgments

The authors would like to acknowledge the support of Science Foundation Ireland grant number 03/IN.3/I361.

References

[1] E. Adelson and J. Bergen. The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, pages 3–20, 1991.

[2] E. Cooke, P. Kauff, and O. Schreer. Image analysis and synthesis for multiple stereo camera teleconferencing systems. *Proc WIAMIS*, pages 405–410, 2003.

[3] C. Fehn, E. Cooke, O. Schreer, and P. Kauff. 3d analysis and image-based rendering for immersive tv applications. *Image Communication Journal*, 17(9):705–715, 2002.

[4] A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE PAMI*, pages 150–162, 1994.

[5] L. McMillan. *An Image-Based Approach to Three-Dimensional Computer Graphics*. PhD thesis, University of North Carolina at Chapel Hill, 1997. Technical Report TR97-013.

[6] K. Mueller, A. Smolic, M. Droese, P. Voigt, and T. Wiegand. Multi-texture modelling of 3d traffic scenes. *Proc ICME*, 2003.

[7] H. Shum and S. Kang. A review of image-based rendering techniques. *Proc VCIP*, pages 2–13, 2000.

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

[9] A. Smolic and D. McCutchen. 3dav exploration of video-based rendering technology in mpeg. *IEEE TCSVT*, pages 348–356, 2004.

[10] M. Tanimoto. Free viewpoint television - ftv. *Proc PCS*, 2004.