# FREE-VIEWPOINT VIDEO FOR INTERACTIVE CHARACTER ANIMATION

*Jonathan Starck and Adrian Hilton*

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
j.starck,a.hilton@surrey.ac.uk

## ABSTRACT

Conventional 2D video production provides a fixed-view reproduction of a recorded event in which a director defines every moment to be viewed. Multiple camera studio systems are now being developed to capture real-world events from multiple viewpoints to replay the event in 3D as a free-viewpoint video. Free-viewpoint video provides the consumer the flexibility to interact with a scene by changing the viewpoint in replaying the event. The challenge in computer vision and computer graphics remains to capture real world events for both replay and to provide the director or consumer the freedom to synthesize new unseen content. In this paper we explore the use of free-viewpoint video in content synthesis by re-using captured scenes of people. Our work combines image based scene reconstruction and video-based animation techniques to provide a new approach that allows interactive character animation from free-viewpoint video.

## 1. INTRODUCTION

Reconstruction and rendering synthetic views of people from video images was first introduced in the 'Virtualized Reality' project by Kanade et al. [1]. The project demonstrated that a recorded event could be replayed in 3D providing an immersive viewing experience. The use of multiple camera studios for research into capture of people has since been widely reported [2, 3, 4, 5, 6, 7]. These systems produce a 3D video representation, termed free-viewpoint video, that supports the manipulation of the virtual view at the point of rendering.

Conventional 2D video production techniques provide a fixed-viewpoint reproduction of a recorded event in which a director defines every moment to be viewed. Free-viewpoint video provides the consumer the flexibility to interact with a scene by changing the viewpoint in replaying an event. This work is now being explored to create a new medium, termed 3D television (3DTV) [8, 9] or Free Viewpoint Television (FVT) [10]. The challenge in computer vision and computer graphics remains to capture real world events to both replay the event and to provide the director or consumer the freedom to synthesize new unseen content.

People form a central component of most entertainment and communication content. In this paper we explore the use of free-viewpoint video in content synthesis by re-using captured scenes of people. Our work combines image based reconstruction [7, 11] and video-based animation techniques [12, 13] to create a new form of animation termed video-based character animation [14]. A multiple camera blue-screen studio is used to capture free-viewpoint video sequences from a performer as depicted in Figure 1, and new content is synthesized by combining different motions to create an animated sequence.

**Fig. 1**. Video capture in a blue-screen studio is used to reconstruct a 3D surface representation for free-viewpoint video production.

In the following sections free-viewpoint video production is introduced and video-based character animation is presented and contrasted with a conventional approach to character animation using an articulated skeleton structure [15]. Video-based character animation allows the synthesis of highly realistic human appearance by reproducing the complex surface dynamics such as cloth motion and wrinkles captured in free-viewpoint video while also providing support to control the scene dynamics allowing a user to direct the animation of a character and produce new content.

(a) Camera image (b) 2.5D stereo

(c) visual-hull (d) silhouette + stereo

**Fig. 2**. Comparison of scene reconstruction techniques showing (b) multi-view stereo, (c) shape from silhouette and (d) global optimization to match silhouette and stereo cues.

(a) Camera image (b) 2.5D stereo

(c) visual-hull (d) silhouette + stereo

**Fig. 3**. Virtual view synthesis using view-dependent rendering from multiple camera images with inexact (b), (c) and accurate (d) surface reconstruction.

## 2. FREE-VIEWPOINT VIDEO PRODUCTION

Multiple camera studio systems have been developed to capture the appearance of people from multiple viewpoints to construct free-viewpoint video. Scene reconstruction is typically performed first to recover a 3D representation of the dynamic scene and video-based rendering is performed to synthesize a novel view by using the original video images as a set of view-dependent textures for the 3D surface.

### 2.1. Scene Reconstruction

The process of 3D scene reconstruction from 2D camera images is termed *image-based modelling*. In image-based reconstruction a 2.5D depth representation is recovered from two or more cameras through a regularized search for image correspondence. The 'Virtualized Reality' project [1] first reconstructed dynamic scenes of people using a 51 camera dome by fusing multiple 2.5D depth images into a single 3D surface. Image-based correspondence however fails with uniform appearance where matching is ambiguous, at depth discontinuities where the surface is tangential to the image plane, and surface fusion relies on accurate reconstruction of all surface regions recovered independently in the 2.5D representation. Figure 2(b) shows multi-view stereo reconstruction for a person where uniform surface appearance and inaccurate 2.5D depth estimation from wide-baseline views results in a noisy integrated 3D surface.

Volumetric reconstruction techniques have been proposed instead to derive the 3D volume that is consistent with multiple images. A volume representation allows inference of visibility and integration of appearance across multiple widely spaced camera views. Shape-from-silhouette (SFS) techniques derive the *visual-hull*, the maximal volume that is consistent with a set of foreground images. Moezzi et al. [16] first demonstrated 3D view generation of people using SFS and video-rate reconstruction has been achieved using image-based volume discretization [17]. The visual-hull, however, only provides an upper bound on the volume of the scene, concavities that are occluded in silhouettes are not reconstructed, appearance is not matched across images and phantom false-positive volumes can occur that are consistent with the image silhouettes. Figure 2(c) shows the visual-hull derived from multi-view silhouettes in which surface concavities are not represented and phantom volumes are incorporated into the recovered surface.

Space-carving [18, 19] techniques provide the *photo-hull*, the maximal volume that is photo-consistent across all visible camera images. Colour or stereo photo-consistency can be used to refine SFS to match appearance across multiple views. Vedula et al. [6] used the photo-hull to reconstruct dynamic scenes of people. Independent voxel carving however suffers either from under or over carving according to the photo-consistency criteria. Regularisation has been introduced using a level-set approach to multiple view stereo [20] and applied to spatio-temporal human shape reconstruction [21]. A level-set approach is however susceptible to local-minima and hence is dependent on the intial surface solution provided by the visual-hull. These techniques also neglect the shape constraint imposed by silhouette images.

Multiple shape cues have been combined for robust reconstruction using iterative local-surface optimization tech-

(a) Captured camera images (0°,45°,90°)

(b) Rendered mid view images (22.5°,67.5°)

(c) Real cameras not used (22.5°,67.5°)

**Fig. 4.** Comparison of video-based rendering using global surface optimization [11] against real images from the same view which were not used in the reconstruction

niques. The visual-hull [22, 23, 7] has been used to provide an initial surface model which is then optimized to match both silhouette contours and to maximize photo-consistency. Surface optimization is however subject to local minima, the surface is constrained to represent only those structures that are defined in the initial surface and retains any incorrect structures. Initialization using the visual-hull retains the phantom structures illustrated in Figure 2(c). Only limited work exists [24, 11] that integrates multiple shape cues for robust scene reconstruction without constraint to a deformation with respect to the visual-hull surface. A global optimization approach [11] without restriction to the shape of the visual-hull is illustrated in Figure 2(d).

### 2.2. Video-based Rendering

The problem of synthesising novel views from camera images is termed *image-based rendering*. In the purest form view synthesis is performed directly from the camera images without the requirement for an explicit 3D scene reconstruction. The images are treated as a set of samples of the light-field in the scene [25, 26] and image samples are interpolated to generate the light-field arriving at a novel viewpoint. Highly realistic view synthesis can be achieved at the cost of requiring dense image sampling to avoid artefacts in interpolating camera images without the exact registration between the images provided by the underlying scene geometry.

Image-based modelling and image-based rendering have been combined to synthesize novel views from a sparse set of cameras by using scene geometry to provide the correspondence between images. Debevec et al. [27] introduced the concept of *view-dependent texturing* in which camera images are used as a set of view-dependent textures for a reconstructed scene. In rendering a virtual view only a subset of cameras that are closest to the virtual camera are used as textures with a weighting defined according to the relative distance to the virtual viewpoint. The advantage of view-dependent rendering is that it can overcome inaccuracies in geometric scene reconstruction by reproducing the change in surface appearance that is sampled in the original camera images. A unified framework for image-based rendering that extends light-field rendering to incorporate a geometric representation for view-dependent texturing has now been presented [28].

Video-based rendering simply extends image-based rendering to the temporal domain by rendering from multiple video streams using the reconstructed surface at each time frame. Free-viewpoint video production has been demonstrated using multiple-view stereo reconstruction [1], shape from silhouette [2, 3, 4], photo-consistency [6], area-based stereo [5] and surface optimization [7]. Figure 3 illustrates the advantage of view-dependent rendering with inexact shape reconstruction using multiple-view stereo, Figure 3(b), shape from silhouette, Figure 3(c) in comparison with exact shape reconstruction using global optimization [11], Figure 3(d). Even with a poor scene reconstruction the virtual view approximates the true scene by reprojecting the appearance sampled in the closest camera views.

Video-based rendering reproduces the detailed surface appearance in the camera images. However, with only an approximate scene reconstruction and potentially inexact camera calibration, the correspondence provided between camera images will be inaccurate. Without exact sub-pixel accurate correspondence, view-dependent rendering will blend different scene points between images causing blurring and double exposure effects in the synthesized views as seen in Figure 3(b),(c). In previous work [7] a view-dependent approach to image-based modelling was introduced to optimize scene geometry for view-dependent rendering, overcoming ambiguities in reconstruction and inexact camera calibration by providing the scene geometry that optimizes image correspondence in the subset of cameras used in rendering. With accurate camera calibration and scene reconstruction in the studio environment, a single optimized scene geometry can now be used to achieve view synthesis approaching the quality of the original video images as shown in Figure 3(d). Figure 4

**Fig. 5**. Model-based reconstruction from camera images instruments a reconstructed scene with a skeletal animation structure to allow synthesis of new animated content.

shows a comparison of view synthesis against a set of camera images not used in reconstruction.

## 3. ANIMATION SYNTHESIS

Free-viewpoint video techniques make no assumption on the structure of the scene to be captured. The advantage of this approach lies in the ability to handle any scene. The disadvantage is that the structure is not consistent over time. The scene is typically captured as an unstructured sequence of triangulated surface meshes in which the surface sampling, geometry, topology and mesh connectivity changes at each time frame. This poses two problems: firstly, there is no consistent parameterization to efficiently represent the large amounts of geometry and appearance data; and secondly there is no structure to manipulate the data to synthesize new content.

### 3.1. Model-based Scene Representation

In previous work [15] a model-based approach to scene reconstruction was introduced to recover a structured representation for scenes of moving people. Model-based reconstruction techniques take a prior model and perform model fitting to recover the shape that best fits the camera images. This approach introduces prior knowledge to regularize scene reconstruction in the presence of visual ambiguities such as noise, sparse or outliers in the data. A prior model also provides a single consistent scene structure to represent the changing surface geometry and appearance. The captured surface can

**Fig. 6**. Captured free-viewpoint video is decomposed into a move-tree of motions for a character. Interactive character control is achieved by allowing a user to change the state of a character in the move tree.

also be animated by instrumenting the prior model with an animation structure. Figure 5 illustrates model-based reconstruction from camera images in which a skeletal structure enables animation of the reconstructed surface.

Model-based scene reconstruction provides the structure necessary to represent free-viewpoint video to allow the synthesis of new dynamic events. Model-based techniques are however restricted to scenes that are consistent with the prior model. The humanoid model depicted in Figure 5 cannot be used for example to recover the motion of a skirt or coat that is not represented. The model must also be registered at each time-frame prior to surface reconstruction requiring articulated motion tracking across all frames. The challenge for computer vision and computer graphics lies in representing multiple view video to support synthesis of new animated content with the freedom to handle any scene content.

### 3.2. Video-based Character Animation

Example-based approaches to animation have demonstrated synthesis of highly realistic content from captured real-world events. Human skeletal motion has been captured using marker-based motion capture technology and represented as a database of motion segments. This database can be compiled into a graph structure, termed a motion graph [29], that defines the feasible transitions between motions. Novel animations can then be synthesized by concatenating segments along paths in the graph [30, 29, 31]. Similarly, resampling video sequences of the face [32, 12] and dynamic scenes [33, 13] has demonstrated video-quality animation synthesis for a fixed viewpoint using a database of video sequences. In previous work [14, 34] example-based animation of people was proposed, extending video-based animation synthesis to an arbitrary virtual viewpoint using motions captured as free-viewpoint video without restriction to the shape of a prior model.

Animation by example is a technique that has already found

**Fig. 7.** Highly realistic animation synthesis using free-viewpoint video motions captured in an 8 camera high-definition studio.

wide-spread use in computer games where segments of skeletal motion data are captured from real people using motion capture technology. A library of articulated motion is constructed for a digital character termed a *Performance Library* and a graph structure termed a *Move Tree* is constructed to define the flow of motion for a character through the library. Character control is achieved at run-time by controlling the state of the character in this move tree. Figure 6 illustrates this concept of character control through a graph of animation states in a move tree.

Interactive character control is achieved using a library of free-viewpoint video recorded from a person in a multiple camera studio. The person is asked to perform a series of predefined motions that form the building blocks for animation synthesis. As traditionally performed in computer games the performance library is manually constructed by defining the start and end-points for motion clips and a move tree is constructed by defining the transition points between different motions. This new form of animation representation reproduces the complete appearance of an actor rather than simply the skeletal motion. Animation synthesis then recreates the detailed surface motion dynamics recorded in the original video images to produce a highly realistic digital character. Figure 7 demonstrates the level of realism that can now be achieved in character animation using free-viewpoint video.

## 4. CONCLUSION

Free-viewpoint video production in a multiple camera studio has been introduced as a means to construct highly realistic digital character animation. A character is constructed as a move-tree of free-viewpoint video sequences of human motion. Compared to a conventional model-based approach with a skeletal animation structure, free-viewpoint video based animation has the advantage of being able to model any animated content that can be recorded in video images. This provides a new representation for animation that recreates the complex surface dynamics recorded in the original video images to create highly realistic characters while providing support to interactively control animation.

## 5. REFERENCES


[1] T. Kanade, P.W. Rander, and P.J. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE Multimedia*, vol. 4, no. 1, pp. 34–47, 1997.

[2] W. Matusik, C. Buehler, and L. Mcmillan, "Polyhedral visual hulls for real-time rendering," *Proceedings of Eurographics Workshop on Rendering,*, pp. 115–126, 2001.

[3] J. Carranza, C. M. Theobalt, M. Magnor, and H.P. Seidel, "Free-viewpoint video of human actors," *ACM Transactions on Graphics (ACM SIGGRAPH 2003)*, vol. 22, no. 3, pp. 569–577, 2003.

[4] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara, "Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 393–434, 2004.

[5] C.L. Zitnick, S. B. Kang, M. Uyttendaele, S. A. J. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics (ACM SIGGRAPH 2004)*, vol. 23, no. 3, pp. 600–608, 2004.

[6] S. Vedula, S. Baker, and T. Kanade, "Image-based spatio-temporal modeling and view interpolation of dynamic events," *ACM Transactions on Graphics*, vol. 24, no. 2, pp. 240–261, 2005.

[7] J. Starck and A. Hilton, "Virtual view synthesis of people from multiple view video sequences," *Graphical Models*, vol. 67, no. 6, pp. 600–620, 2005.

[8] M. Price, J. Chandaria, O. Grau, G.A. Thomas, D. Chatting, J. Thorne, G. Milnthorpe, P. Woodward, L. Bull, E-J. Ong, A. Hilton, J. Mitchelson, and J. Starck, "Real-time production and delivery of 3D media," *Proceedings of the International Broadcasting Convention*, 2002.

[9] W. Matusik and H. Pfister, "3D TV: A scalable system for real-time acquistion, transmission and autostereoscopic display of dynamic scenes," *ACM Transactions on Graphics (ACM SIGGRAPH 2004)*, vol. 23, no. 3, pp. 814–824, 2004.

[10] M. Tanimoto, "Free viewpoint television FTV," *Proceedings of the Picture Coding Symposium*, pp. 15–17, 2004.

[11] J. Starck, G. Miller, and A. Hilton, "Volumetric stereo with silhouette and feature constraints," *British Machine Vision Conference (BMVC)*, vol. 3, pp. 1189–1198, 2006.

[12] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*, pp. 388–398, 2002.

[13] A. Schodl and I. Essa, "Controlled animation of video sprites," *ACM Symposium on Computer Animation*, pp. 121–127, 2002.

[14] J. Starck, G. Miller, and A. Hilton, "Video-based character animation," *ACM Symposium on Computer Animation*, pp. 49–58, 2005.

[15] J. Starck and A. Hilton, "Model-based multiple view reconstruction of people," *IEEE International Conference on Computer Vision (ICCV)*, pp. 915–922, 2003.

[16] S. Moezzi, L.C. Tai, and P. Gerard, "Virtual view generation for 3D digital video," *IEEE Multimedia*, vol. 4, no. 1, pp. 18–25, 1997.

[17] W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan, "Image-based visual hulls," *ACM Transactions on Graphics (ACM SIGGRAPH 2000)*, pp. 369–374, 2000.

[18] K. Kutulakos and S. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, 2000.

[19] G. Slabaugh, W. B. Culbertson, T. Malzbender, M. Stevens, and R. Schafer, "Methods for volumetric reconstruction of visual scenes," *The International Journal of Computer Vision*, vol. 57, no. 3, pp. 179–199, 2004.

[20] O. Faugeras and R. Keriven, "Variational principles, surface evolution, pde's, level set methods and the stereo problem," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 336–344, 1998.

[21] B. Goldluecke and M. Marcus, "Space-time isosurface evolution for temporally coherent 3d reconstruction," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. I, pp. 350–355, 2004.

[22] J. Isidoro and S. Sclaroff, "Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1335–1342, 2003.

[23] C. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.

[24] S. Sinha and M. Pollefeys, "Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation," *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp. 349–356, 2005.

[25] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," *ACM Transactions on Graphics (ACM SIGGRAPH 1996)*, vol. 30, pp. 43–54, 1996.

[26] M. Levoy and P. Hanrahan, "Light field rendering," *ACM Transactions on Graphics (ACM SIGGRAPH 1996)*, vol. 30, pp. 31–42, 1996.

[27] P. Debevec, Y. Yu, and G. Borshukov, "Efficient view-dependent image-based rendering with projective texture-mapping," *Proceedings of Eurographics Workshop on Rendering,*, pp. 105–116, 1998.

[28] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," *ACM Transactions on Graphics (ACM SIGGRAPH 2001)*, pp. 425–432, 2001.

[29] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*, pp. 473–482, 2002.

[30] O. Arikan and D.A. Forsyth, "Synthesizing constrained motions from examples," *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*, pp. 483—490, 2002.

[31] J. Lee, J. Chai, P.S.A. Reitsma, J.K. Hodgins, and N.S. Pollard, "Interactive control of avatars animated with human motion data," *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*, pp. 491–500, 2002.

[32] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio"," *ACM Transactions on Graphics (ACM SIGGRAPH 1997)*, pp. 1–8, 1997.

[33] A. Schodl, R. Szeliski, D.H. Salesin, and I. Essa, "Video textures," *ACM SIGGRAPH*, pp. 489–498, 2000.

[34] J. Xu, T. Yamasaki, and K. Aizawa, "Motion editing in 3d video database," *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2006.