Surface Capture for Performance-Based Animation

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

Abstract

Digital content production traditionally requires highly skilled artists and animators to first manually craft shape and appearance models and then instill the models with a believable performance. Motion capture technology is now increasingly used to record the articulated motion of a real human performance to increase the visual realism in animation. Motion capture is limited to recording only the skeletal motion of the human body and requires the use of specialist suits and markers to track articulated motion. In this paper we present surface capture, a fully automated system to capture shape and appearance as well as motion from multiple video cameras as a basis to create highly realistic animated content from an actor's performance in full wardrobe. We address wide-baseline scene reconstruction to provide 360 degree appearance from just 8 camera views and introduce an efficient scene representation for level of detail control in streaming and rendering. Finally we demonstrate interactive animation control in a computer games scenario using a captured library of human animation, achieving a frame rate of 300fps on consumer level graphics hardware.

Index Terms

Image-based modelling and rendering, Video-based character animation

I. Introduction

Creating realistic animated models of people forms a central task in digital content production. Traditionally highly skilled artists and animators have to construct shape and appearance models for a digital character. The motion is then defined at each time frame or specific key-frames in

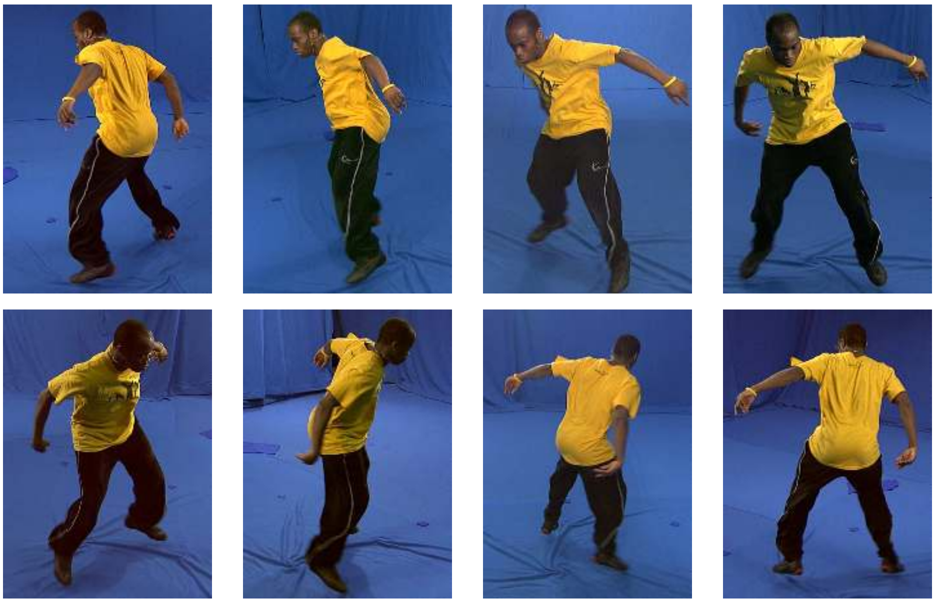


Fig. 1. Surface capture (SurfCap) from multiple video images records the complete appearance for an actor without the requirement for specialist suits and markers used in conventional motion capture. Eight camera views are used in our studio, providing 360° coverage from wide-baseline camera positions at 45° intervals.

a motion sequence to create a digital performance. Motion Capture (MoCap) technology is now increasingly used to record animations from an actor's performance. This technology reduces the time for animation production and captures natural movements to create a more believable production. However, motion capture requires the use of specialist suits and markers and is limited to recording only the skeletal motion of the human body. This technology lacks the detailed secondary surface dynamics of cloth and hair that provides the visual realism of a live performance.

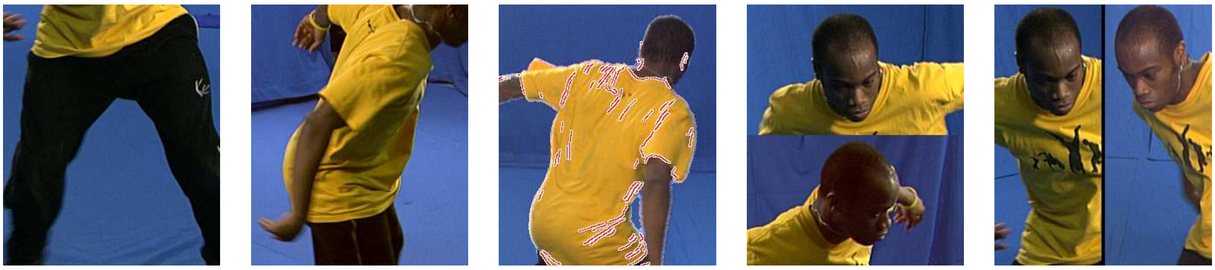
Over the past decade we have investigated studio capture technology with the objective of creating models of real people which accurately reflect the time-varying shape and appearance behaviour of the whole body with clothing [1]. Figure 1 shows a typical frame recorded from a performer in our multiple camera studio. We introduce a system for surface motion capture (SurfCap) that unifies the acquisition of shape, appearance and motion of the human body from multiple-view video. Our system captures a shape and appearance model for the performer in full wardrobe as well as the model animation and surface dynamics to create highly realistic digital content from the performance, unencumbered by a specialist suit.

Performance capture requires the solution of two key problems, scene capture from a limited number of camera views and efficient scene representation for visualization. Three-dimensional (3D) scene reconstruction from camera images must firstly give accurate alignment between views to avoid visual artefacts. We have developed a technology for accurate scene reconstruction from wide-baseline camera positions that allows the recovery of 360° scene structure from just 8 camera views. Secondly, we show how the massive amount of data from multiple view video capture can be efficiently represented for visualization. An efficient representation is introduced for the scene dynamics suitable for level-of-detail control in streaming and rendering. We present fully automated pipelines to first reconstruct and then represent surface motion sequences from multiple view video footage. Finally we demonstrate the application of surface capture for animation synthesis from a real human performance. The rendered performance reproduces the actor's motion and appearance complete with secondary surface dynamics such as cloth motion at a frame rate of 300fps on consumer-level graphics hardware. The multiple-view video and scene reconstruction in this work is available as a resource to the computer graphics research community (*http://www.ee.surrey.ac.uk/cvssp/vmrg/surfcap*).

II. Related Work (sidebar refs [2]-[16])

Manually creating visually realistic digital models of people is a huge task made all the more difficult by our knowledge about how people appear and move in the world around us. Approaches to achieving realism in computer graphics cover a spectrum from pure physical simulation of surface geometry and light interaction in a scene through to direct observation and replay from the real world. Over the past decade there has been a convergence of computer graphics and vision techniques aimed at achieving realism by using observations from camera images. This process of synthesis by example greatly simplifies the complex task of physical simulation and can produce stunning results simply by reusing real-world content.

Reconstruction and rendering images of people from multiple camera views was first popularized by Kanade et al. [2] who coined the term 'Virtualized Reality'. A 5m dome was used with 51 cameras to capture an actor's performance and replay the event in 3D to create an immersive, virtualized view. Rendering virtual views of moving people from multiple cameras has since received considerable interest. Systems for multiple view reconstruction and video-based rendering have been developed [3], [4], [5], [6], [7], [8], [9]. These techniques create a



(a) Uniform (b) Occlusions (c) Features (d) Specularity (e) Baseline

Fig. 2. Whole-body images present several important challenges with (a) uniform surface appearance, (b) self occlusions, (c) sparse features, (d) non-lambertian surfaces and (e) large distortions between views with wide baseline camera positions.

3D video, also called free-viewpoint video, at a quality that now approaches the original video images [10], [11].

Whole-body images of people present several important challenges for conventional computer vision techniques in free-viewpoint video production. These problems are illustrated in Figure 2.

- **Uniform appearance** Extended areas of uniform appearance for skin and clothing limit the image variation to accurately match between camera views to recover surface shape.
- **Self occlusions** Articulation leads to self-occlusions that makes matching ambiguous with multiple depths per pixel, depth discontinuities and varying visibility across views.
- **Sparse features** Features such as clothing boundaries must be matched to recover appearance without discontinuities or blurring but provide only sparse cues in reconstruction.
- **Specularities** Non-lambertian surfaces such as skin cause the surface appearance to change between camera views making image matching ambiguous.
- **Wide-baseline** With only a restricted number of cameras a wide baseline configuration is required for 360° coverage leading to large distortions in appearance between views.

Reconstruction algorithms are based either in the 2D domain and search for image correspondence to triangulate 3D position, or work in the 3D domain to derive the volume that projects consistently into the camera views. Image-based correspondence forms the basis for conventional stereo-vision in which pairs of camera images are matched and a surface recovered [2]. However, image-based correspondence fails where matching is ambiguous and requires fusion of surfaces from stereo pairs which is susceptible to errors in the individual surface reconstruction. A volumetric approach on the other hand allows inference of visibility and integration of appearance across all camera views without image correspondence. Shape-from-silhouette (SFS) techniques [3] derive the *visual-hull*, the maximal volume that reproduces a set of foreground silhouettes in the cameras. This is refined in space-carving techniques [7] which provide the *photo-hull*, the maximal volume that has a consistent foreground colour across all visible camera images. Silhouette and colour cues have now been combined for robust shape reconstruction using iterative shape optimisation techniques [11], [12], [13].

Shape reconstruction for free-viewpoint video [2]-[11] simply allows the replay of a recorded event in 3D without the structure necessary to create or manipulate content for animation production. Model-based shape reconstruction techniques [4], [14] have been developed which fit a generic humanoid model to multiple-view images, providing a model structure to enable motion editing and retargeting. However, model-based techniques are limited by the predefined model structure and cannot be applied to complex scenes with large changes in structure for example where loose clothing causes large changes in the surface geometry. On the other hand, data-driven techniques with no prior model have demonstrated highly realistic synthesized animations in the 2D domain by replaying sequences from example video clips. Resampling video sequences of simple dynamic scenes [15] has achieved video-quality animation for a single fixed viewpoint. In related work [16] we proposed animation by example using free-viewpoint video of human motions to provide a complete 3D digital representation. In this paper we present a complete system for surface capture and a free-viewpoint video representation suitable for use in animation production from a recorded human performance.

III. Surface Capture

We record the performance of an actor in a dedicated multiple camera studio with controlled lighting conditions and a chroma-key background. Due to cost such systems have typically been developed using machine vision rather than professional cameras which do not give the colour quality required for production work in broadcast or film. The aim of our work is to demonstrate the potential of this technology for high-quality entertainment content. We have therefore developed a studio system using a limited number of professional film quality High-Definition (HD) cameras.

In our studio, eight HD cameras are equally spaced around a circle of 8m diameter at a height

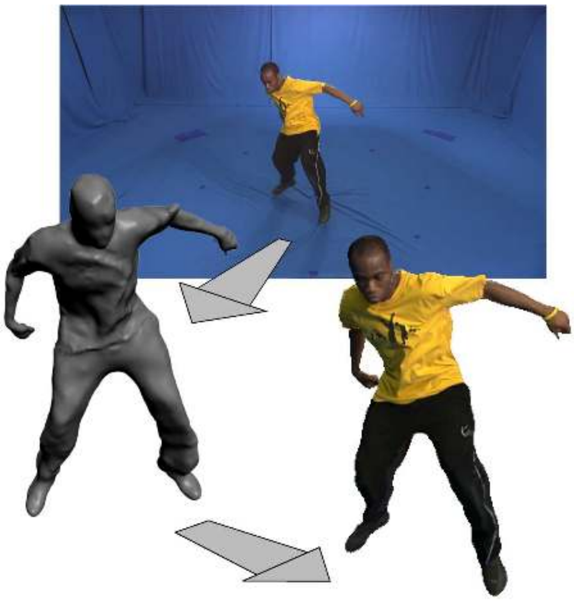


Fig. 3. Video capture is used to reconstruct a surface model of an actor for 3D visualization.

of 2m above the studio floor. This gives a performance volume of $4 \times 4 \times 2$m with a wide-baseline $45°$ angle between adjacent camera views as shown in Figure 1. Performances are captured using Thomson Viper cameras in HD-SDI 20-bit 4:2:2 format with 1920×1080 resolution at 25Hz progressive scan. Synchronized video from all eight cameras are recorded uncompressed direct to disk with eight dedicated PC capture boxes using DVS HD capture cards. This studio setup enables high-quality performance capture from wide-baseline camera positions.

A. Studio Calibration

Recording a performance from multiple video cameras provides the appearance of the actor from a fixed set of viewpoints. Geometric surface reconstruction is required to provide a complete 3D model for interactive visualization as shown in Figure 3. In order to recover geometric shape from camera images the camera parameters must be calibrated. Camera calibration is based on imaging a known object and deriving the camera transformation that reproduces the object in a set of captured images. Techniques typically use a planar grid of known geometry that is visible from all cameras and a number of public domain tools are available to achieve this (*http://www.vision.caltech.edu/bouguetj/calib_doc*). This approach is suitable for camera systems where the images have an overlapping field of view but is impractical in studio where cameras surround the capture volume. For studio production where cameras may be reconfigured many times in one day with changes in location and zoom, a simple and quick method of calibration is of key practical importance.

To achieve rapid and flexible calibration of a multiple camera studio we developed a wand-based calibration technique. Wand-based calibration uses two spherical markers at a known distance apart on a rigid rod. By acquiring video sequences of the moving wand, it is possible to build up large sets of point correspondences between views in a short time. These replace the role of a planar grid and have the advantage that markers are simultaneously visible from opposing camera views. A calibration algorithm using wand markers has been developed to estimate both the intrinsic (focal length, centre-of-projection, radial distortion) and extrinsic (pose, orientation) camera parameters. This approach allows calibration of studio camera systems in less than 10 minutes with an accuracy comparable to grid-based calibration. In addition there is no requirement for all cameras field of view to overlap allowing flexible calibration of studio camera systems with extended capture volumes. A public domain implementation of the wand calibration together with further technical details is available at (*http://www.ee.surrey.ac.uk/cvssp/vmrg/wandcalibration*).

B. Surface Geometry Reconstruction

Once the camera system is calibrated, a performance can be recorded from multiple viewpoints for reconstruction. Scene reconstruction addresses the problem of recovering a 3D model for a scene that places the appearance sampled in the camera images in correspondence, a process termed *image-based modelling*. Our reconstruction algorithm is specifically tailored to provide robust shape reconstruction from wide-baseline camera views without loss of visual detail such as creases in clothing. This is based on the observation that a number of shape cues in camera images constrain the geometry. Firstly, with a fixed chroma-key backdrop foreground silhouettes can be reliably extracted to constrain the outline of the geometry. Secondly, image features at appearance discontinuities such as clothing boundaries provide a dominant cue that can be reliably matched across wide-baseline images to constrain the surface position. Finally, conventional appearance matching based on surface colour or intensity provides only a weak cue with wide-baseline cameras that will define the dense surface geometry. We combine all shape cues within a single

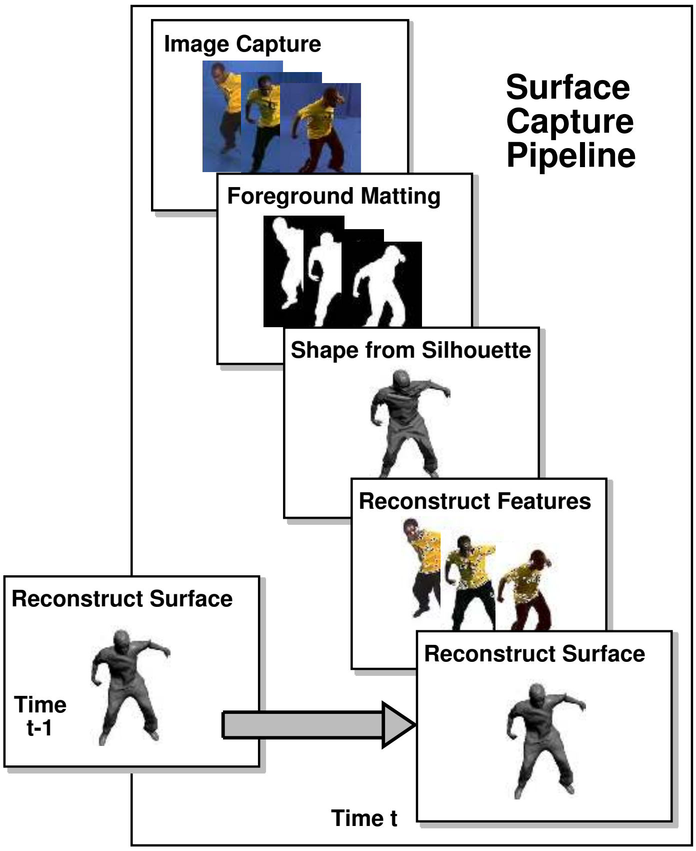


Fig. 4. Overview of the automated surface capture pipeline.

framework that recovers the surface with a maximum appearance consistency between camera images while constrained to match extracted foreground silhouettes and feature correspondences.

Our surface reconstruction pipeline is divided into several distinct steps outlined in Figure 4. A multiple view video sequence of a performance is first recorded from an actor. Foreground silhouettes are then extracted from the camera images by performing chroma-key matting where the foreground pixels in the images are separated from the known background colour. An alpha matte is extracted for each image that defines the foreground opacity at each pixel and the foreground colour where pixels in the original image are mixed between foreground and background. The silhouettes are then used to derive the *visual-hull*, the maximal volume in the scene that reproduces the silhouettes in the camera images. The visual-hull defines an upper-bound on the true volume of the scene and so constrains the feasible space for surface reconstruction. Wide-baseline feature matching is then performed between the cameras to extract contours on the underlying surface inside the visual-hull that produce feature lines in the images. The scene is finally reconstructed as the surface within the visual-hull that passes through the surface features while reproducing the silhouette images and maximizing the consistency in appearance between views.

Visual-Hull Reconstruction: Shape reconstruction from silhouettes, or shape-from-silhouette (SFS), is a popular technique for scene reconstruction due to the simplicity of the reconstruction algorithm and the robust nature of shape recovery in the studio setting where the foreground can be reliably and consistently extracted from a known fixed background. However, SFS 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 7(a) shows the reconstructed visual-hull from multiple silhouette images in which surface concavities are not represented and phantom volumes are incorporated into the recovered surface. Multiple shape cues have been used previously in reconstruction by iteratively refining the shape of the visual-hull. Local shape optimization is however subject to local minima and the surface retains the phantom structures in the visual-hull. We extend recent work on global optimization techniques [17] to integrate multiple shape cues for robust wide-baseline reconstruction without restriction to a deformation with respect to the visual-hull surface [18].

Feature Matching: Once the extent of the scene is defined by reconstructing the visual-hull, surface features are matched between views to derive constraints on the location of the scene surface. Surface features correspond to local discontinuities in the surface appearance and candidate features are extracted in the camera images using a Canny-Deriche edge detector. Each feature contour in an image is then matched with the appearance in an adjacent camera view. Correspondence is constrained first to satisfy the camera epipolar geometry defining the relationship between observations in pairs of cameras. Each feature pixel corresponds to a ray in space that connects the scene surface and the camera centre of projection. This ray in turn projects to a line of pixels called an epipolar line in an adjacent camera view and the feature pixel can only match along this line. Correspondence is further constrained by intersecting this ray with the feasible volume defined by the visual-hull giving a set of line segments in the adjacent camera view. The connected set of pixel correspondences are then derived in the adjacent view that maximizes the image correlation for the feature contour. Correspondence is verified by

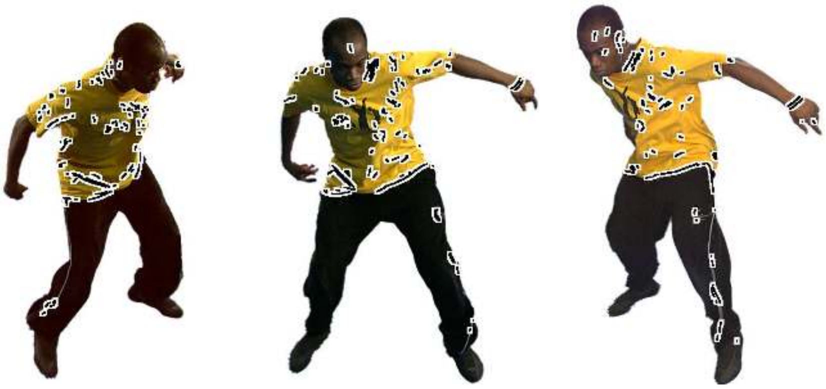
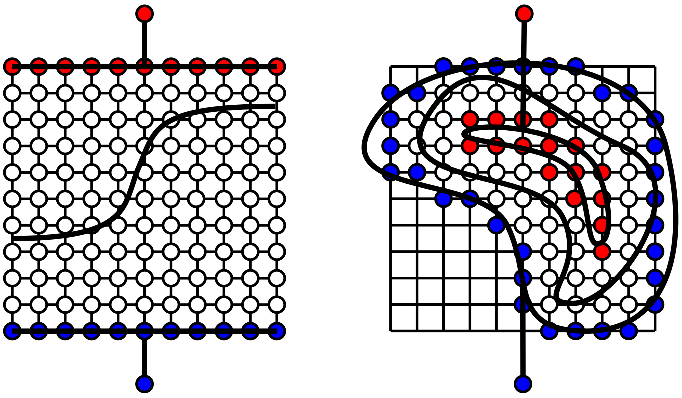


Fig. 5. Extracted feature matches in camera images. Feature lines are matched between adjacent views and pruned to a *left-right* consistent set by enforcing reciprocal correspondence. The reconstructed surface contours are subsequently used to to constrain surface reconstruction at the features.

enforcing left-right consistency between views such that a feature pixel in one camera is required to match a feature pixel in an adjacent camera with a reciprocal correspondence. Figure 5 shows the set of left-right consistent features derived for a set of camera images.

Dense Reconstruction: Feature reconstruction provides only a sparse set of 3D line segments that potentially lie on the scene surface. Dense surface reconstruction is then performed inside the volume defined by the visual-hull. We adopt a global optimization approach by discretizing the volume and treating reconstruction as a maximum-flow / minimum-cut problem on a graph defined in the volume. Surface reconstruction as network flow problem on a graph is illustrated in Figure 6. Each discretized element of the volume, termed a voxel, forms a node in the graph with adjacent voxels connected by graph edges. Edges are weighted by a cost defined by the consistency in appearance between camera images. The maximum flow on the graph saturates the set of edges where the cost is minimized and the consistency is maximized. The final surface can then be extracted as the set of saturated edges cutting the graph. Efficient optimization methods exist using graph-cuts, providing the global optimum that maximizes the correlation between views on the final surface [19]. In our approach we adapt graph-cut optimization to derive a surface that passes through the reconstructed feature contours where possible and reproduces the initial set of silhouette images. We further constrain the reconstruction to be temporally consistent by minimizing the distance between surfaces at subsequent time frames.

Surface Extraction: The surface for the scene is finally extracted from the volume recon-



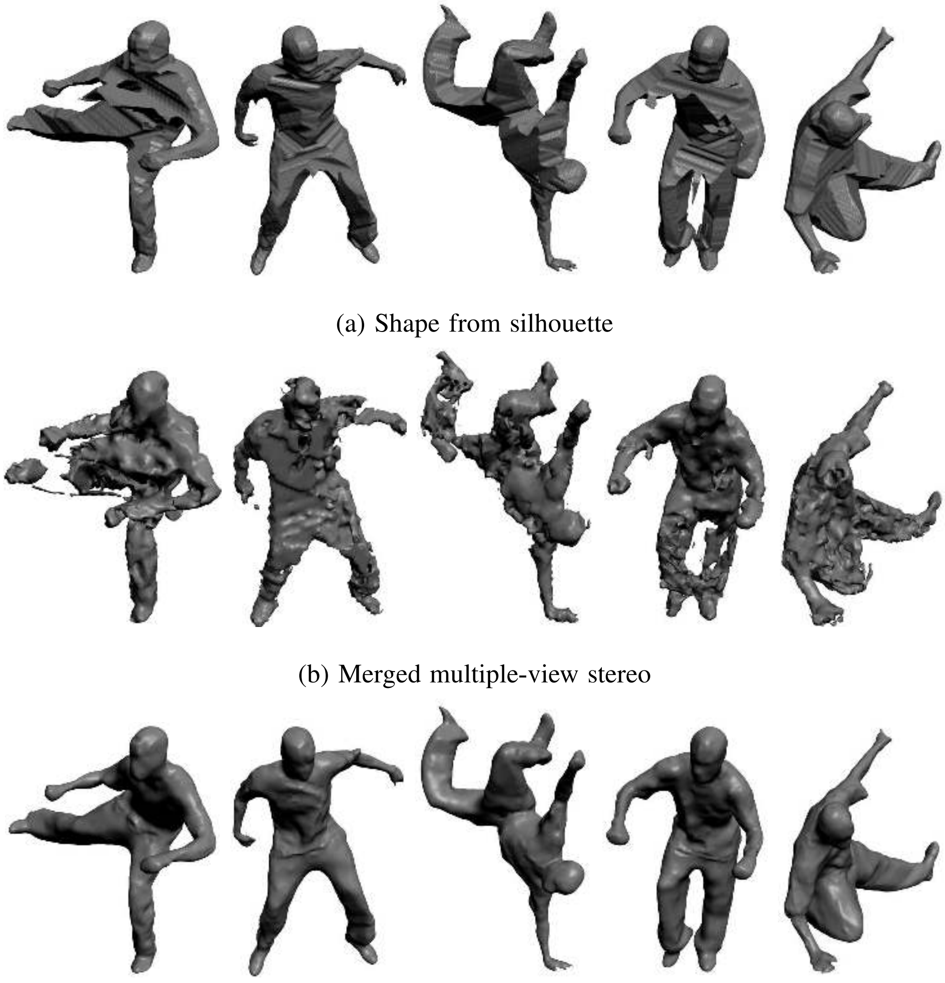
(a) Graph for a planar scene (a) Graph for a generalized scene

Fig. 6. Surface reconstruction as a cut on a discrete volumetric graph shown here in cross-section. The maximum-flow on the graph between a source (blue) and sink (red) node saturates the set of edges where the edge weight is minimized and consistency between the camera images is maximized, corresponding to the scene surface.

struction as a triangulated mesh. Mesh vertices are derived to sub-voxel accuracy using a local search to maximize image consistency across all visible cameras. Figure 7 shows the result of surface reconstruction in comparison with conventional reconstruction techniques. The visual-hull alone, Figure 7(a) provides only an approximate estimate of the scene geometry that incorporates phantom volumes. Reconstruction by matching camera pairs, termed *multi-view stereo* Figure 7(b), illustrates the inherent ambiguity in matching the appearance in adjacent views with wide-baseline cameras leading to a noisy 3D surface estimate. Global surface optimization using multiple shape cues, Figure 7(c), combines robust shape reconstruction from silhouettes with appearance matching across camera views.

IV. Scene Representation

Studio capture records multiple video streams of a human performance from a specific set of viewpoints. Our system for studio calibration and scene reconstruction enables high-quality recovery of the 3D time-varying surface during a performance. Even with a limited number of camera views this capture represents a huge overhead in terms of the stored data. With 8 high-definition images this equates to approximately 50MB per frame or 75GB per minute. This massive amount of video and geometry data must be represented efficiently to allow streaming for real-time rendering.



(a) Shape from silhouette

(b) Merged multiple-view stereo

(c) Surface capture fusing silhouette, feature and stereo cues

Fig. 7. Improved surface shape reconstruction is achieved from 8 wide-baseline camera views by combining silhouette, feature and stereo cues in comparison with multiple-view stereo and shape from silhouette.

Our pipeline for constructing a structured surface representation is outlined in Figure 8. Surface capture initially provides a time-varying sequence of triangulated surface meshes in which the surface sampling, geometry, topology and mesh connectivity changes at each time frame for a 3D object. We transform this unstructured representation to a single consistent mesh structure such that the mesh topology, connectivity and texture domain is fixed and only the geometry changes over time. This is achieved by mapping each mesh onto the spherical domain and remeshing as a fixed subdivision sphere. Praun and Hoppe [20] introduce a robust approach to parameterization of genus-zero surfaces in the spherical domain. Surface genus is defined as the maximum number

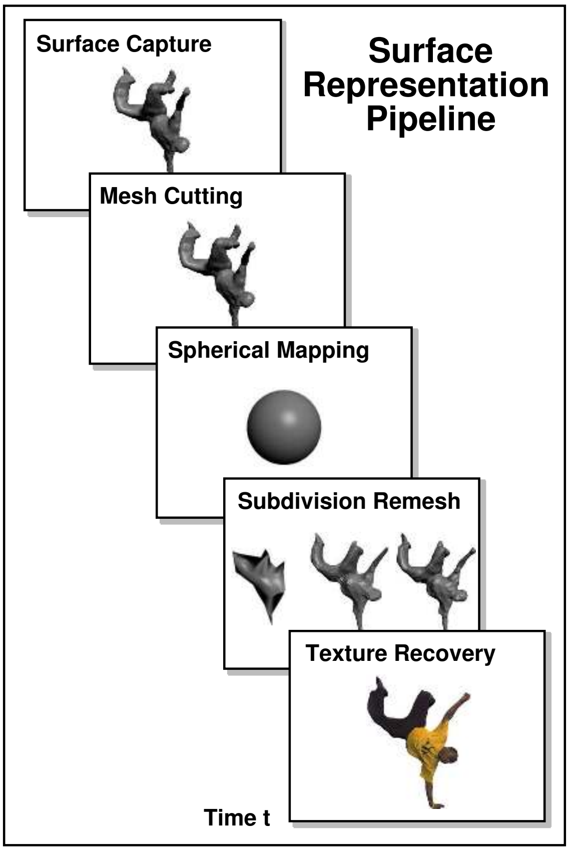


Fig. 8. Overview of the automated surface representation pipeline.

of cuts that can be made before a surface becomes disconnected and genus-zero surfaces are topologically equivalent to a sphere. Our surface remeshing algorithm extends this technique to handle genus-N surfaces and incorporates adaptive resampling [21] to handle the mapping of highly deformed surfaces such as the human body onto the spherical domain. The final step is to combine the appearance from the original camera images as a single time-varying texture map for the subdivision surface.

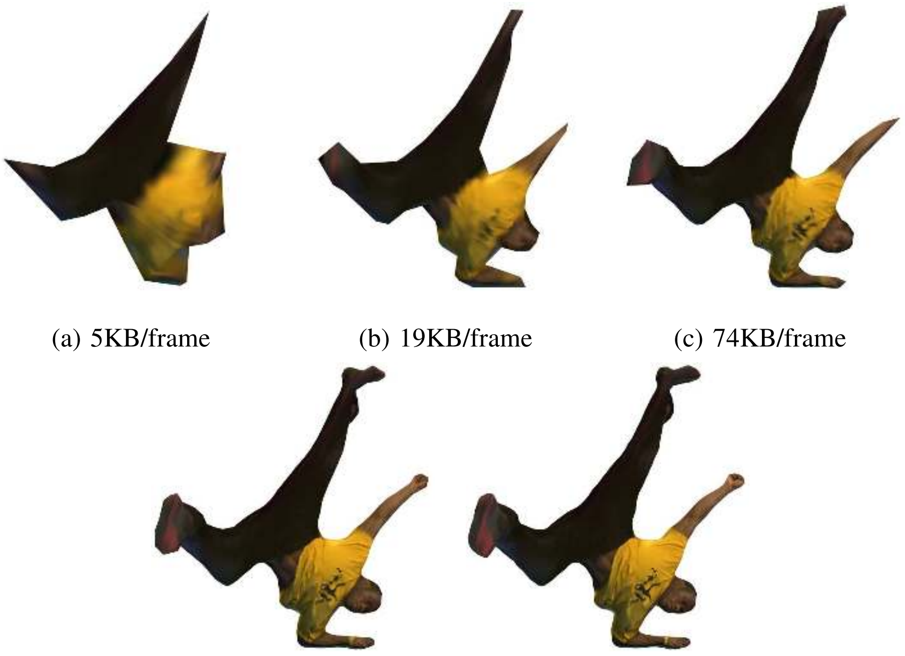
Mesh Cutting: Captured sequences of people often have non genus-zero topology and we first transform a genus-N surface using mesh cutting [22]. A closed genus-zero mesh M is constructed by iteratively inserting a series of cuts on a reconstructed surface. The algorithm first creates a topological graph of iso-curves on the mesh. Branch points are identified and loops in the graph are then cut along the shortest isocurve in the loop. The mesh becomes topologically equivalent to a trimmed sphere and the holes are triangulated to create a closed genus-zero surface.

Spherical Parameterization An embedding is constructed for the mesh on the unit sphere $\hat{S}$, $(M \rightarrow \hat{S})$ [20]. The mesh M is simplified to a tetrahedron by iteratively removing vertices from the mesh. The tetrahedron is then mapped to the unit sphere and the vertex removal reversed, reinserting the vertices onto the spherical domain. A 1-to-1 map is achieved as vertex removal maintains an embedding during mesh simplification and vertices are inserted into the kernel of their neighborhood on the sphere, maintaining the embedding in the spherical domain.

Adaptive Remeshing The spherical mesh $\hat{S}$ is resampled onto a regular quaternary subdivision of a unit octahedron S by constructing a map $(S \rightarrow \hat{S})$. Embedding a complex genus-zero surface such as a person on the unit sphere requires a high degree of mesh deformation resulting in a highly distorted parameterization. To accurately represent complex geometry during mapping we optimize the mesh to match the vertex sampling density [21]. Given the mapping $M \rightarrow \hat{S}$ and $S \rightarrow \hat{S}$, we can finally resample the attributes of the original mesh M onto the uniform domain of the subdivision surface S.

Texture Recovery Surface appearance is resampled in the texture domain from the camera view with the greatest surface sampling rate, retaining the highest resolution appearance in the texture. We group the assignment of mesh facets to camera images to create maximal contiguous regions, minimizing the boundary on the surface between images. Multiple resolution blending [23] with spherical surface continuity is then used to construct a single seamless texture. This multi-resolution approach ensures that the extent of texture blending corresponds to the spatial frequency of the features in the image, preserving the higher frequency detail that can become blurred with simple linear texture blending techniques. Blending at low-frequencies can compensate for discrepancies in the colour balance between views, although in practise colour calibration should be performed prior to image capture.

The unstructured surface motion sequence is now transformed into a single mesh structure as a subdivision sphere S and the time-varying surface geometry is represented as a single time-varying vertex buffer with a predefined overhead. The subdivision connectivity in the mesh allows for level of detail control to manipulate this overhead in the geometric representation. Figure 9 illustrates the representation at different levels in the sub-division hierarchy for the surface. Uncompressed surface capture at 50 MB/frame can now be represented as a structured surface with control of the overhead in storage and streaming allowing for an uncompressed



(a) 5KB/frame (b) 19KB/frame (c) 74KB/frame

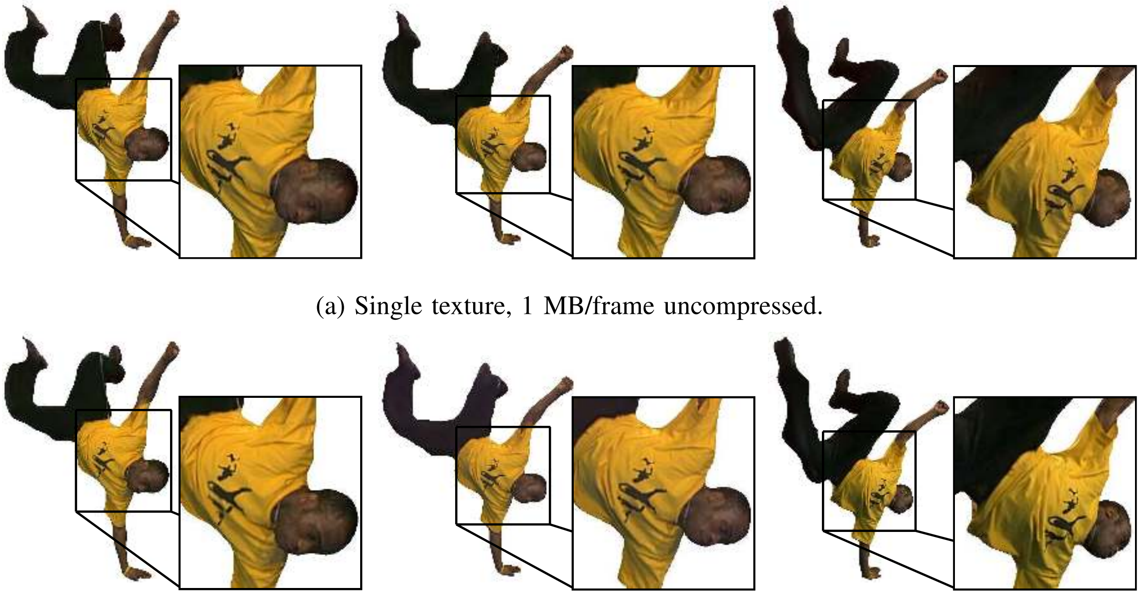
(d) 291KB/frame (e) 1158KB/frame

Fig. 9. Remeshing as a constant topology subdivision surface reduces the geometry overhead and provides level of detail control. The *uncompressed* overhead for the geometry and texture is illustrated here at different levels for comparison with the raw video capture at 50MB/frame.

cost of 5Kb/frame, 19Kb/frame, 74KB/frame, 291KB/frame and 1MB/frame at different levels of detail.

Our process of texture recovery differs from current techniques for free-viewpoint rendering of human performance where the original video images are typically used as texture maps in rendering. In free-viewpoint video *view-dependent texturing* is performed in which only a sub-set of cameras that are closest to the virtual camera are used as textures with a weight defined according to the relative distance to the virtual viewpoint. By using the original camera images this retains the highest-resolution appearance in the representation and incorporates view dependent lighting effects such as surface specularity. The disadvantage lies in the large overhead in storing, streaming and rendering from all camera images at 50MB/frame.

It is interesting to note that view-dependent rendering is often used simply to overcome problems in surface reconstruction by reproducing the change in surface appearance that is sampled in the original camera images. However, resolution and view-dependent reflectance



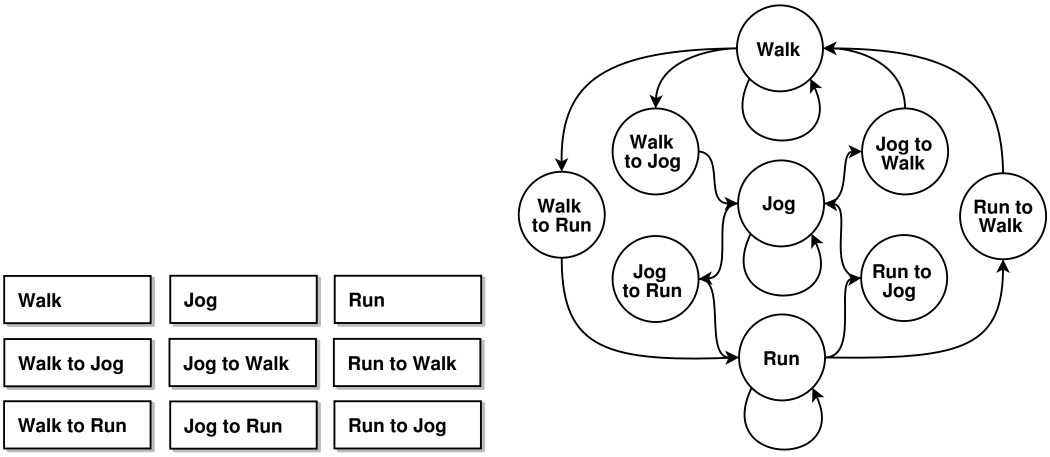
(a) Single texture, 1 MB/frame uncompressed.

(b) View-dependent rendering from camera images, 50 MB/frame uncompressed.

Fig. 10. With accurate surface reconstruction a single surface texture can be extracted reducing the overhead in comparison to conventional free-viewpoint video where the original camera images are used as a set of view-dependent textures.

effects are only retained where the geometry is exact such that the alignment of the camera images is correct in blending the images. Blending with incorrect alignment produces image blurring and double-exposure effects in a synthesized view. Our surface capture technique instead optimizes the alignment of surface geometry between camera images such that a single texture can be recovered from all cameras without visual artefacts, significantly reducing the overhead in representing appearance.

Rendering with a single texture is compared with view-dependent rendering in Figure 10. This demonstrates an equivalent visual quality for both techniques. In fact a small amount of blurring can be seen in view-dependent rendering where exact sub-pixel alignment cannot be achieved across a 45° camera baseline. This blurring is removed in our texture resampling stage by recovering surface appearance from the camera images with the highest sampling rate and using a non-linear multiple resolution blend between the appearance in each camera.



(a) Captured Performance Library

(b) Character Move Tree

Fig. 11. A performance library of captured motions is transformed into a move-tree that defines the feasible transitions for a digital character. Run-time character control is achieved by controlling the state of a character in the move tree.

V. Animation by Example

Surface capture and representation provide the data necessary to replay a human performance in 3D with control of the overhead to allow streaming and real-time rendering. This provides a 3D video representation, termed free-viewpoint video, where the user now has control over the camera viewpoint. The goal of our work is to use free-viewpoint video for content production rather than simply replaying a fixed event as a virtualized reality. In this section we consider the application of surface capture for animation production using the paradigm of synthesis by example.

Example based techniques for animation have been proposed previously to re-use content captured using conventional motion capture technology for skeletal motion synthesis [24]. These techniques use a library of captured performance and concatenate motion segments to produce new content. Animation by example is a technique that has already found wide-spread use in computer games where motion capture is reused to synthesize character actions. Skeletal animation for a digital character is typically compiled into a library of motions termed a *Performance Library*. A graph structure termed a *Move Tree* is then 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 11 illustrates this concept of character control through a graph of animation states in a move tree.

Surface capture records human performance as a temporal sequence of surface shape and appearance. We use our surface capture system to record a performance library from an actor and construct a move tree for interactive character control. The actor is asked to perform a series of predefined motions such as walking, jogging and running 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. A move tree is then manually constructed by defining the transition points between different motions. This new form of animation representation using surface capture reproduces the complete appearance of an actor rather than simply the skeletal motion and recreates the detailed surface motion dynamics recorded in the original video images to produce a highly realistic digital performance.

Results are shown for a Streetdancer performing a variety of movements wearing loose fitting clothing. The performance included body popping, kicks, jumps, handstands and freeform moves. Figure 12 shows the captured surface motion sequences from views not aligned with the original camera views. These results demonstrate firstly that a high visual quality is obtained and secondly that the rendered sequences for novel views reproduce a life-like appearance for the digital character. Results of rendering with interactive animation and viewpoint control in a public domain games engine (*http://axiomengine.sourceforge.net*) are shown in Figure 13. Figure 14 illustrates animation control for a second character wearing a long coat where a user directs the transition between walk and run motions. This demonstrates the capture and reproduction of loose clothing for an actor in full wardrobe. The rendering performance for the representation achieved 300 frames per second on a nVidia 6600GT graphics card with an uncompressed overhead of 1MB/frame. Video sequences captured from the game-engine are available from (*http://www.ee.surrey.ac.uk/cvssp/vmrg/surfcap*) demonstrating that surface motion capture reproduces the natural dynamics of loose clothing and achieves video-quality rendering in animation that is comparable in resolution and detail to conventional video.

Animation production is currently limited by the manual construction of a character motion tree and the transitions between animation states within this tree. In the ideal case an actor would perform an arbitrary set of motions within a multiple camera studio. The captured surface sequences would then be automatically transformed into a representation that allows any human motion to be recreated from the performance library. This represents a challenging task for future work. Firstly, surface correspondence is required between arbitrary body poses so that

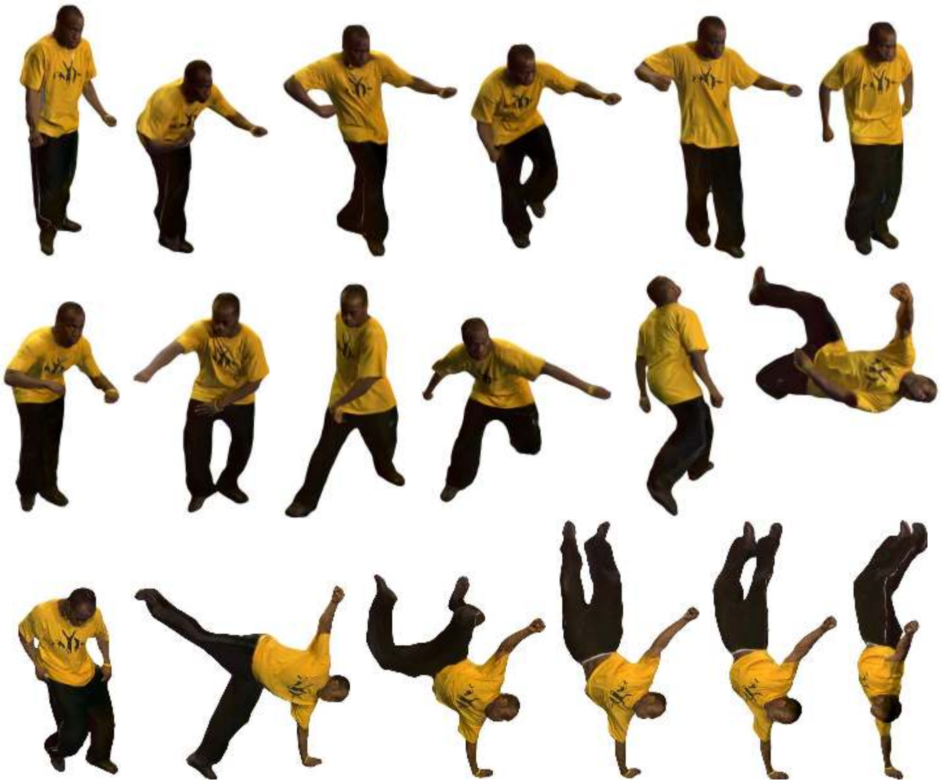


Fig. 12. Captured surface animation sequences: lock (top), kick (middle) and hand (bottom) sequences.

the correspondence between motion segments can be determined to connect and then seamlessly blend segments. To date only limited work has addressed whole-body correspondence and this has only been applied to blend similar surface shapes. Secondly, artistic control is required to direct and edit the content produced. Currently only high-level control can be achieved through the animation state of a character. An artist requires the ability to interactively manipulate the character animation at a finer level to synthesize new content that was not recorded in the original surface capture.

VI. Conclusion

We have presented a fully automated framework for the reconstruction and representation of surface motion sequences of people recorded in a multiple camera studio. The advantage of our system lies in the robust method of shape reconstruction from wide-baseline camera views and the efficient representation of geometry and appearance allowing for level-of-detail control in

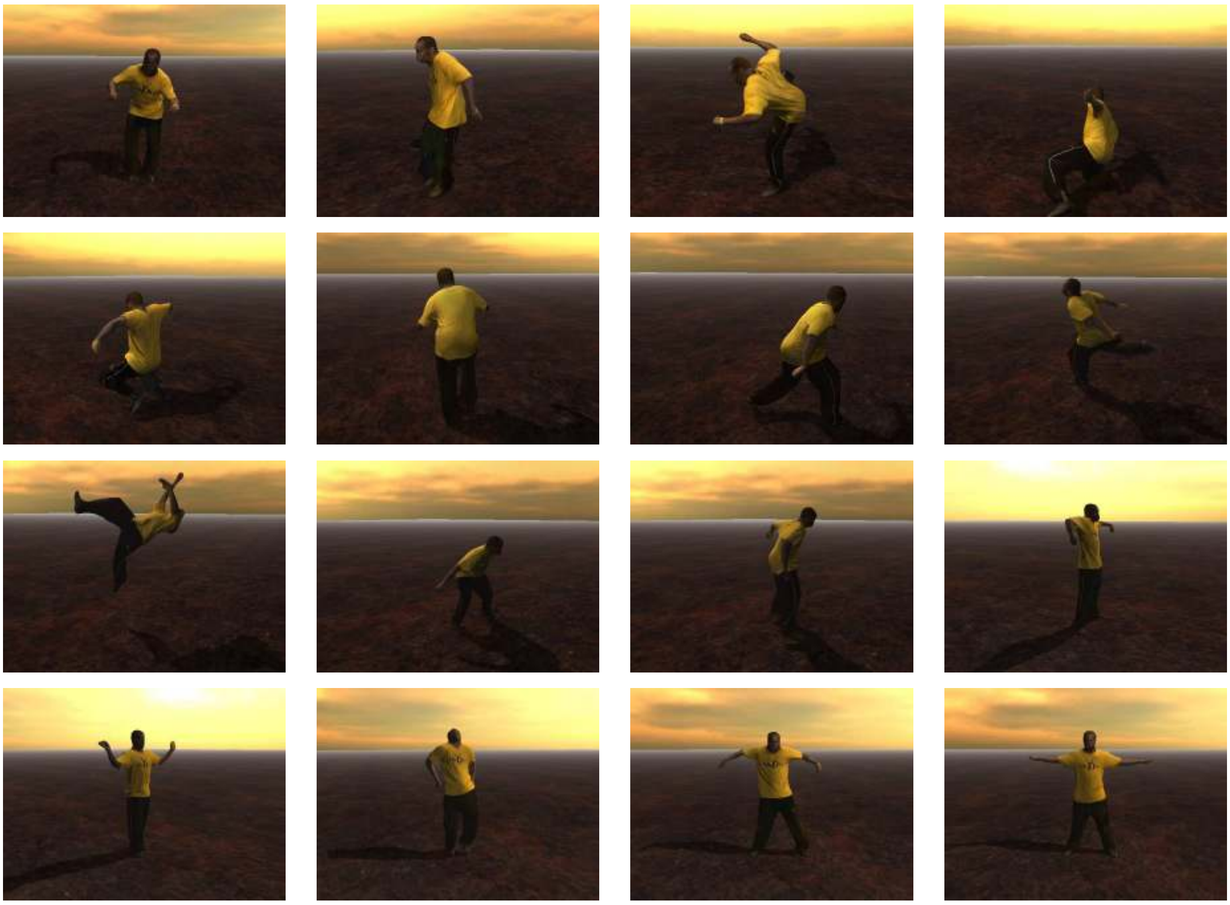


Fig. 13. Transition between lock and pop using a kick sequence, rendered interactively from 360° views in the public domain Axiom games engine.

storage and streaming. We are able to capture a complete surface with 360° appearance from just 8 high-definition film quality cameras and represent the geometry and appearance data for real-time rendering.

The application of surface capture for animated content production has been demonstrated. A library of animated motions is first captured from a real human performance and represented as a move-tree traditionally used in computer games for skeletal character animation. Interactive character control is achieved in an open-source game engine with real-time rendering and control of complex movements visualized from 360° views. Surface capture allows the introduction of a new form of animation synthesis using real-world content that creates a highly realistic character reproducing the surface detail in clothing wrinkles and facial expressions recorded in multiple

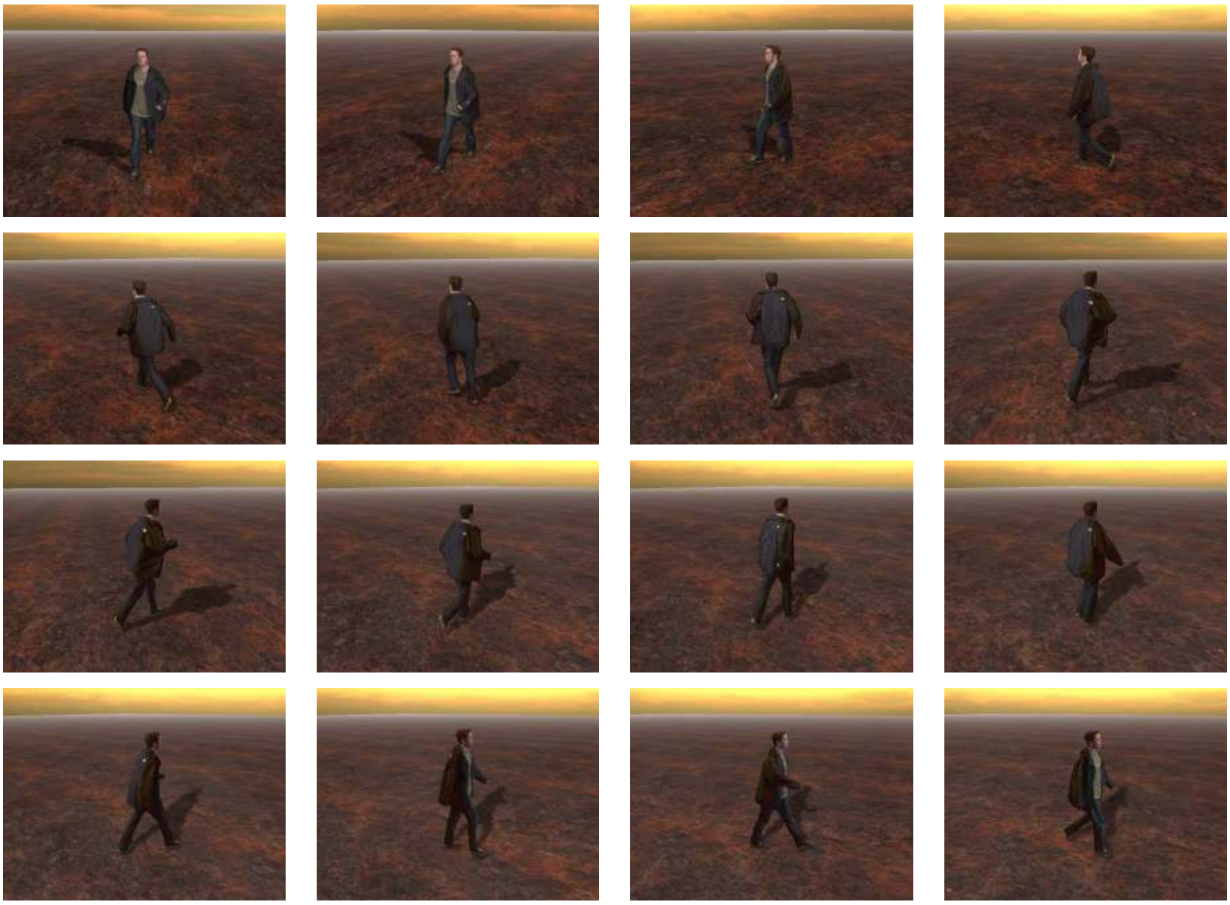


Fig. 14. Interactive animation control with a character changing between walk and run motions. Synthesised views demonstrate the reproduction of the cloth motion recorded in the actor's coat.

view video footage.

References

[1] A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun, and J. Illingworth, "Whole-body modelling of people from multi-view images to populate virtual worlds," *The Visual Computer: International Journal of Computer Graphics*, vol. 16, no. 7, pp. 411–436, 2000.

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

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

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

[5] O. Grau, “Studio production system for dynamic 3D content,” *Visual Communications and Image Processing. Proceedings of the SPIE.*, vol. 5150, pp. 80–89, 2003.

[6] 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.

[7] 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.

[8] M. Waschbsch, S. Wrmlin, D. Cotting, F. Sadlo, and M. Gross, “Scalable 3D video of dynamic scenes,” *The Visual Computer: International Journal of Computer Graphics*, vol. 21, no. 8-10, pp. 629–638, 2005.

[9] J. Allard, J.-S. Franco, C. Ménier, E. Boyer, and B. Raffin, “The grimage platform: A mixed reality environment for interactions,” *Proceedings of the 4th IEEE International Conference on Computer Vision Systems*, p. 46, 2006.

[10] C. 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.

[11] 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.

[12] C. Hernandez 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.

[13] Y. Furukawa and J. Ponce, “Carved visual hulls for image-based modeling,” *European Conference on Computer Vision (ECCV)*, pp. I: 564–577, 2006.

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

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

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

[17] G. Vogiatzis, P. Torr, and R. Cipolla, “Multi-view stereo via volumetric graph-cuts,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 391–398, 2005.

[18] 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.

[19] Y. Boykov and V. Kolmogorov, “Computing geodesics and minimal surfaces via graph cuts,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 26–33, 2003.

[20] E. Praun and H. Hoppe, “Spherical parameterization and remeshing,” *ACM Transactions on Graphics (ACM SIGGRAPH 2003)*, pp. 340–349, 2003.

[21] K. Zhou, H. Bao, and J. Shi, “3D surface filtering using spherical harmonics,” *Computer-Aided Design*, vol. 36, pp. 363–375, 2004.

[22] D. Steiner and A. Fischer, “Cutting 3D freeform objects with genus-n into single boundary surfaces using topological graphss,” *ACM Symposium on Solid Modeling and Applications*, pp. 336—343, 2002.

[23] P. Burt and E. Adelson, “A multiresolution spline with application to image mosaics,” *ACM Transactions on Graphics*, vol. 2, no. 4, pp. 217–236, 1983.

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

PLACE
PHOTO
HERE

Jonathan Starck is a Senior Research Fellow in multiple view scene analysis and studio based 3D content production at Surrey University, UK. He studied Engineering at Cambridge University, graduating with double-starred first honours and received the BMVA Sullivan thesis award for his PhD at Surrey University. His research interests include image based modelling for computer graphics and animation.

PLACE
PHOTO
HERE

Adrian Hilton is Professor of Computer Vision and Graphics at the University of Surrey, UK. His research interest is robust computer vision to model and understand real world scenes for entertainment and communication. Contributions include the first hand-held 3D scanner and automatic modelling of people. He is an area editor of CVIU and member IEE, IEEE and ACM.