Model-Based Multiple View Reconstruction of People

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

Abstract

This paper presents a framework to reconstruct a scene captured in multiple camera views based on a prior model of the scene geometry. The framework is applied to the capture of animated models of people. A multiple camera studio is used to simultaneously capture a moving person from multiple viewpoints. A humanoid computer graphics model is animated to match the pose at each time frame. Constrained optimisation is then used to recover the multiple view correspondence from silhouette, stereo and feature cues, updating the geometry and appearance of the model.

The key contribution of this paper is a model-based computer vision framework for the reconstruction of shape and appearance from multiple views. This is compared to current model-free approaches for multiple view scene capture. The technique demonstrates improved scene reconstruction in the presence of visual ambiguities and provides the means to capture a dynamic scene with a consistent model that is instrumented with an animation structure to edit the scene dynamics or to synthesise new content.

1. Introduction

The challenge of achieving computer generated scenes that are visually indistinguishable from the real world is leading to an increasing use of real-world observations to create and animate models in computer graphics. This is demonstrated by the wide spread use of 3D scanning techniques to reconstruct detailed character models from clay maquettes and the use of marker based motion capture to produce believable movement in character animation. An exciting new area of research lies in capturing dynamic 3D models from real world events using multiple cameras. This has the potential to allow for the rapid creation of 3D content with the visual realism of conventional video, unencumbered by the restricted capture environment of 3D scanning systems and the invasive use of markers required for motion tracking.

Three-dimensional production from multiple view video, or 3D video, was first popularised by Kanade et. al. [9, 15] who coined the term "Virtualized Reality". Conventional video provides only a 2D view of a scene in a linear form defined by a director. Presenting an event in 3D allows visualisation in the same way as virtual reality, providing an immersive 3D experience. Systems for multiple view reconstruction have been developed [9, 14, 21] and applied to reconstruct sequences of moving people. These techniques make no assumptions on the structure of the captured scene and produce a new scene model for each frame of a sequence. The advantage of this approach is that there are no restrictions on the dynamic content of the scene. The disadvantage is that the structure of the scene is not consistent over time and reconstruction is not robust to visual ambiguities. Inconsistencies in the models at different time frames become apparent when viewed as a sequence. There is also no consistent structure to edit or reuse the dynamic content limiting the techniques to replaying the captured event.

In this paper we introduce model-based reconstruction of people using a generic model of human shape and kinematic structure. This approach enables anatomically correct estimation of whole-body shape and imposes a consistent structure for sequences of reconstructed models of a moving person. The model-based approach provides prior knowledge of human shape to overcome visual ambiguities inherent in multiple view reconstruction. Reconstruction of models with a common underlying structure provides temporal correspondence between models for a dynamic sequence. The use of a model animation structure allows for reuse of captured models to synthesise new dynamic events or the editing of the dynamic content of a captured sequence.

A novel model-based reconstruction algorithm is introduced to optimise the triangulated surface mesh of a generic humanoid model to simultaneously match stereo, silhouette, and feature data across multiple views. Stereo correspondence is used to optimise surface shape to sub-pixel accuracy for recovery of colour texture. The shape of the model is used to constrain the search for stereo correspondence in a coarse-to-fine framework that enables shape recovery from

noisy stereo data. Multiple shape cues are incorporated in the framework to complement the stereo data which can fail in regions of poor image texture and occlusion boundaries. The surface mesh is treated as a shape constrained surface to regularise optimisation and preserve the surface parameterisation with respect to kinematic structure. Preservation of surface parameterisation is important for subsequent animation using the predefined kinematic structure of the model. A shape constraint is presented for an arbitrary triangular mesh, a common representation used to model surface shape in computer graphics. The resulting surface model is optimised to sub-pixel accuracy and constrained by the generic humanoid shape where visual ambiguities arise between multiple camera views.

2. Related Work

Acquisition of visually realistic models of objects and scenes from images has been a long standing problem in computer vision and computer graphics. This has resulted in formulation of the projective geometry for reconstruction of shape from multiple view images of unknown scenes. Common techniques for shape reconstruction include: voxel carving from image silhouettes to obtain the maximal bounding volume or *visual hull* for a given set of segmented foreground images [11]; photo-consistency between views which reconstructs the maximal volume inside the visual hull which is consistent with the observed image colour or *photo hull* [18, 10]; and multiple view stereo [9, 15] where local surface appearance is used to estimate correspondence between views.

Previous research has achieved reconstruction of 3D shape, appearance and movement of a person from multiple view video sequences using image silhouettes [14], photo-consistency [21] and stereo [9]. These techniques reconstruct a sequence of independent surface representations of a person which are then textured to achieve realistic appearance. However, due to the inherent visual ambiguity such techniques may fail to accurately reconstruct complex self-occluding objects with areas of limited surface texture such as people. Object-centred and model-based techniques have been introduced to overcome visual ambiguities by making use of approximate scene geometry to constrain reconstruction.

Object-centred surface reconstruction was introduced by Fua and Leclerc [5] in which an initial surface model is derived from stereo data and then optimised to match stereo and shading cues between images, accounting for geometric distortions and occlusions between views. Vedula et. al. [22] proposed a model-enhanced stereo system where an initial reconstructed scene model is used to refine the search range for stereo correspondence to improve stereo matches for reconstruction. Faugeras and Keriven [4] present a volumetric reconstruction technique posed in a level-set framework where the estimated scene geometry from the evolving surface of the level-set is used to account for geometric distortions and occlusions between camera views. These techniques make use of reconstructed geometry to improve the estimation of image correspondence. However they remain susceptible to problems such as lack of image texture that makes correspondence ambiguous.

Model-based techniques use prior knowledge of the scene geometry to constrain shape recovery in the presence of visual ambiguities and can reduce the influence of noisy, sparse or outlier data in shape estimation. Debevec et. al. [2] describe a model-based stereo system, in which manually defined planar sections of architectural scenes are used to recover dense geometric detail from stereo. Kakadiaris and Metaxas [8] infer a segmented body model for a person from the deforming contour of image silhouettes and construct 3D shape of body parts from orthogonal views. Hilton et. al. [7] present model-based shape from silhouette to recover whole-body models of people. Plankers and Fua [16] adopt a model consisting of implicit volume primitives to recover the gross upper-body shape and pose from both stereo and silhouette data.

In this paper a model-based technique for reconstruction is introduced that recovers the shape of a whole-body animated human model. The model-based framework integrates shape information from silhouette, stereo and feature cues across multiple views and uses prior knowledge of human shape in the model to constrain shape recovery. This new model-based approach provides the integration of different visual cues in a coarse-to-fine optimisation with the use of a novel local shape constraint for arbitrary triangular surface meshes. This is the first example of a model-based framework integrating multiple visual cues for whole-body human modelling in an arbitrary pose.

3. Model-Based Reconstruction

The model-based approach to scene reconstruction introduced in this paper takes a generic animated humanoid model and matches it to available shape, appearance and feature data. The generic model is first registered with the images using a sparse set of manually defined feature points. The shape of the model is then optimised to match the images in a coarse-to-fine, model-based framework, in which the model shape is used to constrain and regularise the multiple view reconstruction process. The model is then textured to generate a final representation that matches the shape and appearance in the multiple view images.

The generic model used in this work consists of a single seamless mesh defining the surface shape of the body, connected to a skeleton structure for animation. The mesh contains 8000 polygons and the skeleton 17 articulated joints,

providing the gross shape and pose of the human body. A single texture map is generated for the generic model providing an intuitive representation of appearance that can be easily edited. A vertex weighting scheme is used for model animation.

3.1. Data acquisition

A multiple camera studio is used for data acquisition. A backdrop is used for foreground segmentation and diffuse light sources are used to provide an even ambient lighting environment. The studio contains 9 cameras, 8 of which form 4 stereo pairs that are positioned to give 360^o coverage of the subject at the centre of the studio with 1 camera placed overhead to increase the intersection angle between views for shape from silhouette. One additional camera is also placed to provide a viewpoint for comparison and is not used in reconstruction or texture recovery. Sony DXC-9100P 3-CCD colour cameras are used, providing PAL-resolution progressive scan images. Intrinsic and extrinsic camera parameters are calibrated using the Camera Calibration Toolbox for Matlab from MRL-Intel [1].

3.2. Model Registration

The generic model is aligned with the subject in the studio using a manual interface. A user first defines a set of feature points consisting of skeleton joints and mesh vertices on the model. The corresponding image locations are then selected on the captured images and the 3D feature locations are reconstructed from the multiple views. The skeleton pose and dimensions are finally updated to register the model with the features, aligning the model with the images. Feature points are typically specified at the articulated joints and the facial definition points such as the eyes, ears, nose and mouth.

The alignment process is performed in a least-squares framework widely used in human pose estimation [19]. The generic model is parameterised with 6 degrees of freedom (DOF) for global position and orientation, 9 DOF for the skeletal bone-lengths, with left and right-sides constrained to be symmetric, and 25 DOF for the articulation of the skeletal joints [19]. These model parameters, ϕ, are estimated to minimize the squared error between the model features, $\underline{x}_f(\phi)$, and the reconstructed manually defined feature locations, $\underline{o}_f$.

$$\arg\min\phi \sum_f \|\underline{o}_f - \underline{x}_f(\phi)\|^2 \tag{1}$$

The problem is non-linear, requiring close initial values for convergence to the correct solution [19]. The pose of the model is therefore initialised using an analytical solution for the position and orientation of the trunk given the locations of the hip and shoulder joints, and an analytical solution for the limb rotations given three joint locations on each limb, such as the hip, knee and ankle. The model parameters are then optimised using an iterative bound-constrained solver, allowing for inclusion of constraints on joint rotations and bone-lengths [19].

3.3. Multiple View Shape Reconstruction

Once the generic model is aligned to match the subject in the studio, the shape of the model is optimised to satisfy the appearance in each of the captured images. The objective function for optimisation is formulated with three data terms, the cost of fitting to silhouette data E_V, fitting stereo data E_S and fitting the features E_F, and a regularisation term E_R across the surface of the model governed by the factor α. The function is discretized at the vertices of the mesh, $\underline{x}_v$, and energy minimization is performed using gradient descent. The deformation of the model mesh is then given by:

$$E = E_V + E_S + E_F + \alpha E_R \tag{2}$$

$$\frac{d\underline{x}_v}{dt} = -\frac{dE}{d\underline{x}_v} = -\left(\frac{dE_V}{d\underline{x}_v} + \frac{dE_S}{d\underline{x}_v} + \frac{dE_F}{d\underline{x}_v} + \alpha\frac{dE_R}{d\underline{x}_v}\right) \tag{3}$$

Each data fitting term in the objective function is defined in terms of a squared 3D error between the vertex location and the corresponding reconstructed vertex position from the data, giving a least-squares solution in data fitting. Gradient-descent optimisation is performed in a model-based coarse to fine fashion. Optimisation starts at the coarsest resolution corresponding to the initial expected error in the shape of the model, and vertex positions are reconstructed for the stereo, silhouette and feature data up to the expected error in the shape. The model vertices are then deformed by gradient-descent to match the data. Optimisation and data reconstruction is then scheduled to progress to finer levels of resolution, and finishes at the reconstruction accuracy of the cameras. The advantage of this coarse-to-fine model-based approach is that the reconstructed vertex locations are iteratively updated as the model deforms, allowing the reconstructed vertex positions to converge to a solution in the presence of noisy data and incorrect matches.

3.3.1. Model-based stereo

Stereo matching is used between camera images to provide the shape data that aligns the appearance in the images. The stereo energy term is defined as the squared error between a vertex and the reconstructed location obtained by stereo matching between adjacent cameras in the studio. A model-based approach to stereo is adopted in which the shape of

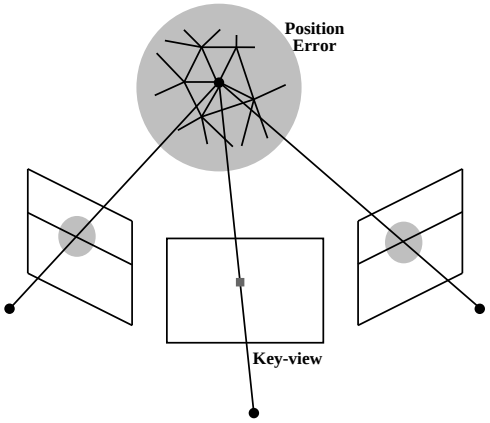

Figure 1. Stereo matching between a key view for a vertex and adjacent cameras, showing the search region along each epipolar line in adjacent views.

the model is used to constrain the search range for matches between cameras.

For each mesh vertex we first determine the key view with the greatest surface visibility according to the camera with the closest viewpoint to the direction of the vertex normal. We then recover the disparity in each camera adjacent to the key view that forms a stereo pair as illustrated in Figure 1. Here we make the simplifying assumption of a fronto-parallel surface at each vertex and match along epipolar lines in rectified camera images [6]. The search range along an epipolar line is defined by the expected error in the position of the model vertex, and the search range perpendicular to an epipolar line is defined by the maximum reprojection accuracy of the cameras. We use a normalised cross-correlation matching score to allow for a linear change in intensity between images with non-lambertian surfaces or inexact intensity matched cameras.

For each vertex we obtain a reconstructed 3D position from the disparity in each camera c that forms a stereo pair with the key view. We define our stereo matching energy term at each vertex as the squared error between the vertex position and each of the reconstructed 3D positions $\underline{z}_{v,c}$ as follows:

$$E_S = \sum_v \frac{1}{n_v^c} \sum_{c=0}^{n_v^c-1} \|\underline{z}_{v,c} - \underline{x}_v\|^2 \tag{4}$$

$$\frac{dE_S}{d\underline{x}_v} = -\frac{1}{n_v^c} \sum_{c=0}^{n_v^c-1} \left(\underline{z}_{v,c} - \underline{x}_v\right) \tag{5}$$

It is important to account for occlusion in stereo matching to remove unfeasible matches between occluded regions. Stereo matching is therefore only performed between camera images in which a vertex is unoccluded. Here we use the visibility algorithm introduced by Debevec et. al. to determine vertex visibility [3]. Hardware accelerated OpenGL rendering is used to render the model to each camera view with mesh polygons colour coded for identification. A vertex is then defined as visible in a camera view if it is unoccluded by the rendered polygon at the projected location in the camera.

3.3.2. Shape from silhouette

The visual-hull gives a robust shape estimate, complementing the stereo data that can be noisy where image texture is lacking or where there is a significant distortion in appearance between images. Here we use a volumetric voxel-carving technique to generate the visual-hull from the segmented image silhouettes. The capture volume is divided into a discrete set of voxel elements at a 1cm resolution. All voxels that project outside the foreground silhouettes are then carved, leaving the visual-hull. The surface voxels are extracted as the set of foreground elements with one or more faces connected to a background voxel.

The visual-hull data fitting term E_V is defined as the squared error between each vertex position and the closest surface element on the visual hull $\underline{y}_v$.

$$E_V = \sum_v \|\underline{y}_v - \underline{x}_v\|^2 \tag{6}$$

$$\frac{dE_V}{d\underline{x}_v} = -\left(\underline{y}_v - \underline{x}_v\right) \tag{7}$$

3.3.3. Surface feature matching

In model registration, exact feature locations are defined for the model. In fitting we seek to satisfy these constraints on the shape of the model. The feature fitting term E_F is defined as the squared error between each vertex position and the feature correspondence for the vertex, $\underline{o}_v$.

$$E_F = \sum_v \|\underline{o}_v - \underline{x}_v\|^2 \tag{8}$$

$$\frac{dE_F}{d\underline{x}_v} = -\left(\underline{o}_v - \underline{x}_v\right) \tag{9}$$

Here we use sparse data interpolation with radial basis functions to generate the 3D error term, $(\underline{o}_v - \underline{x}_v)$, at the vertices where no feature correspondence f is defined. The interpolated error term is given by Equation 10 using the the 3D thin-plate spline function $\psi_f(\underline{x}) = \|\underline{x} - \underline{x}_f\|^3$. The interpolation includes an explicit affine basis $R, \underline{t}$ to account for any global error terms. The parameters $\underline{\lambda}_f$ are obtained by the solution to the linear system given in Equation 11, in terms of the known error terms at the defined feature points, under an additional set of constraints that remove the affine contribution from the radial basis functions, Equation 12.

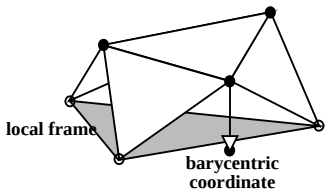

Figure 2. Local parameterization of triangle vertices in terms of barycentric coordinates and height in the frame defined by the vertices edge-connected to the triangle.

$$\underline{o}_v = \underline{x}_v + \sum_f \underline{\lambda}_f \psi_f(\underline{x}_v) + R\underline{x}_v + \underline{t} \tag{10}$$

$$\left(\underline{o}_f - \underline{x}_f\right) = \sum_{f'} \underline{\lambda}_{f'} \psi_{f'}(\underline{x}_f) + R\underline{x}_f + \underline{t} \tag{11}$$

$$\textstyle\sum_{f'} \underline{\lambda}_{f'} = \underline{0} \quad \sum_{f'} \underline{\lambda}_{f'}^T \underline{x}_{f'} = 0 \tag{12}$$

3.3.4. Shape regularisation

The internal energy of the model is designed to preserve the prior shape information of the generic model in order to regularise shape fitting in the presence of noisy or irregular data. Regularisation energy, E_R, is defined as the squared error between each vertex and the reconstructed position of the generic vertex location in terms of a local parameterization that is invariant to position and orientation. Here we adopt a polygon based parameterization that can describe the local shape on an irregular triangular mesh [19]. This local shape constraint preserves the local parameterisation of the surface mesh and hence the animation structure for the surface in relation to the underlying skeletal control structure.

A vertex on a triangle can be defined in terms of the barycentric coordinates in a local coordinate frame given by the vertices edge-connected to the triangle, as depicted in Figure 2. A vertex position can then be reconstructed in each of the local frames for the triangles sharing the vertex. The error metric for regularisation is defined as the squared error between the vertex location and the reconstructed location in each local frame, Equation 13, where $(\alpha_{v,j}, \beta_{v,j}, h_{v,j})$ are the barycentric coordinates (α, β) and height offset h in the j^{th} triangle-based frame for a vertex with valence N_v on the generic model.

$$E_R = \sum_v \frac{1}{N_v} \sum_j \left\| \underline{x}(\alpha_{v,j}, \beta_{v,j}, h_{v,j}) - \underline{x}_v \right\|^2 \tag{13}$$

$$\frac{dE_R}{d\underline{x}_v} = -\frac{1}{N_v} \sum_j \left(\underline{x}(\alpha_{v,j}, \beta_{v,j}, h_{v,j}) - \underline{x}_v \right) \tag{14}$$

3.4. Texture Recovery

For texture recovery each polygon in the model is assigned to the camera image with the greatest surface visibility, according to the camera with the closest viewpoint to the direction of the polygon normal. Occlusion is accounted for in camera selection using the mesh visibility algorithm to ensure that a polygon is not textured from a camera view for which it is occluded [3]. For each camera image in turn the model texture is recovered for all the visible polygons. Stereo correspondence between camera images in optimisation provides sub-pixel accurate image locations for the model vertices. Texture resampling is performed here using OpenGL rendering to the model texture from a camera image. The textures derived from each camera are then composited onto the single texture map for the model using image masks corresponding to the polygons selected for texturing from each camera. Texture blending is required to ensure a smooth transition between regions of texture recovered from different camera images. Blending is performed using a multi-resolution spline [12] ensuring that the extent of blending between images corresponds to the spatial frequency of the image features.

4. Results and Discussion

4.1. Shape reconstruction

Model-based scene reconstruction is first compared with current model-free techniques including shape from silhouette [14], voxel colouring [21], and stereo vision [9]. The visual-hull is derived using a volumetric voxel carving technique as describe in section 3.3.2. Colour carving is performed using the *Voxel-Coloring* algorithm [18] with the exception that RGB colour values are normalised by intensity in testing colour consistency to allow for intensity variations between viewpoints with non-lambertian surfaces in the scene. A triangular surface model is generated from the volume representation through iso-surface extraction. Stereo depth-maps are derived from camera pairs using a maximal surface technique [20] with the added constraint that the search range for disparity is restricted to lie within the visual-hull, removing outliers in stereo correspondence. Multiple depth maps are fused into a single surface model using volumetric fusion and iso-surface extraction [15].

Figure 3 shows the reconstructed geometry for several captured frames. The visual-hull provides a good estimate of shape as the human body has few concavities. Inaccuracies tend to arise where there is self-occlusion with articulation of the limbs leading to extraneous sections in the reconstructed volume. Colour carving provides only limited improvement on shape due to the quality of reconstruction from silhouettes and the consistent colour across many

sections of the body. Stereo vision provides a noisy estimate of shape due to the inherent ambiguity in establishing stereo correspondence across uniform regions of colour texture on the body. The model-based technique demonstrates improved shape recovery, integrating silhouette and stereo data, and making use of prior knowledge to regularise shape where data is ambiguous. The disadvantage here is that the prior shape of the model constrains the set of feasible shapes that can be recovered and the detailed geometry at the hands and the hair is missing.

4.2. View generation

The goal of multiple view scene capture is not necessarily accurate surface reconstruction, instead it can be the generation of visually realistic virtual views. View-dependent rendering techniques have been developed [3, 13, 17] to synthesise realistic views with only approximate scene geometry. We therefore compare the reconstructed models for the synthesis of a virtual view. View-dependent rendering is performed by texturing the models from the camera images adjacent to the virtual view. A view-dependent vertex weighting [17] is used that favours the closest camera views to smoothly blend between the texture derived from each camera. Multi-pass texturing is used to blend the weighted texture derived for each camera image in rendering to the virtual view.

Figure 4 shows view-dependent rendering to the tenth camera image not used in reconstruction for the geometry shown in Figure 3. This camera is placed equidistant between two adjacent stereo pairs in the studio. The triangulated surfaces for the visual-hull and voxel-colouring generate a "blocky" rendered view. This arises due to the discrete nature of the surface normals leading to non-smooth view-dependent weighting of texture. The surface of the voxel-coloured volume is also noisy which further degrades the visual-quality of the rendered view. The merged stereo depth-maps provide an improved result due to the smooth nature of the merged surface geometry. However, visual artifacts are apparent where stereo matching fails leading to missing regions of the geometry. The model-based approach provides a complete surface model that can be used to interpolate the camera views in view-dependent rendering and demonstrates improved visual quality in comparison with the model-free techniques.

4.3. Motion editing and synthesis

One of the advantages of the model-based approach is that we are free to instrument the model with an animation structure to control the captured scene dynamics. An example is given in which a series of models is constructed with texture for a sequence in which a subject jumps in the air. The motion of the limbs are then retargeted to different locations using inverse kinematics and the models rendered as shown in Figure 6. We also demonstrate that models captured from a single frame can subsequently be used to synthesise new motion sequences as shown in Figure 5.

Figure 5. Synthesis of new motions, showing the reconstructed and texture-mapped human model on the left with three novel poses.

5. Conclusions

In this paper we have presented a model-based technique for multiple view scene reconstruction. A humanoid computer graphics model is used to reconstruct static frames and dynamic sequences of people in a multiple camera studio. The model is animated to match the pose of a subject at each time frame. The model shape is then optimised in a regularised coarse-to-fine stereo framework in which the search range for stereo matches is gradually reduced to the calibration accuracy of the camera system, enabling convergence in the presence of noisy stereo data. The model-based framework incorporates silhouette and feature cues to complement the stereo data. Regularisation is performed using a local constraint that preserves the shape and parameterisation of the surface mesh of the model. This approach provides improved shape reconstruction when compared to model-free techniques in the presence of visual ambiguities by making use of prior shape information in the model and multiple shape cues.

The principal drawbacks of the technique are the requirement for manually defined feature points in order to pose the model, and the restriction on the content of the scene

(a) Camera image (b) Visual-hull (c) Voxel-colouring (d) Merged stereo (e) Model-based

Figure 3. Shape reconstruction.

(a) Camera image (b) Visual-hull (c) Voxel-colouring (d) Merged stereo (e) Model-based

Figure 4. View-dependent rendering in comparison to camera view not used in reconstruction.

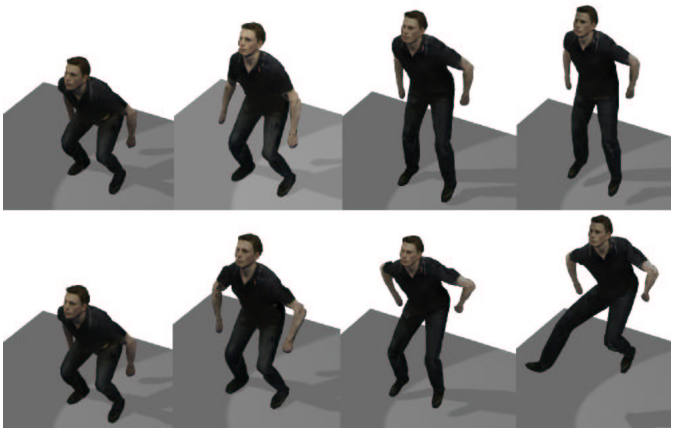

Figure 6. Editing a captured sequence (top row) using the animation structure of the model (bottom row).

imposed by the shape of the generic model. This can however form the basis to analyse, edit or synthesise new data from a limited set of captured frames. Model-free techniques for multiple view reconstruction allow the capture of arbitrary scenes. A model-based approach provides the means to deal with visual ambiguities in scene reconstruction and provides a temporally consistent scene model for analysis of dynamic sequences.

Acknowledgements

This work was supported by EPSRC Grant GR/M88075 and sponsored by the BBC and BT Exact.

References

[1] J.-Y. Bouguet. Camera calibration toolbox for matlab: www.vision.caltech.edu/bouguetj/calib-doc. Technical report, MRL-INTEL, 2003.

[2] P. Debevec and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Proceedings of ACM SIGGRAPH*, pages 11–20, 1996.

[3] P. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. *9th Eurographics Rendering Workshop*, pages 105–116, 1998.

[4] O. Faugeras and R. Keriven. Variational principles, surface evolution, pde's, level set methods and the stereo problem. Technical Report 3021, INRIA, 1996.

[5] P. Fua and Y. Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16:35–56, 1995.

[6] A. Fusiello, E. Trucco, and A. Verri. Rectification with unconstrained stereo geometry. *British Machine Vision Conference*, pages 400–409, 1997.

[7] A. Hilton, D. Beresford, T. Gentils, R. Smith, W. Sun, and J. Illingworth. Whole-body modelling of people from multiview images to populate virtual worlds. *The Visual Computer*, 16(7):411–436, 2000.

[8] I. Kakadiaris and D. Metaxas. Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, 30(3):191–218, 1998.

[9] T. Kanade, P. Rander, and P. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.

[10] K. Kutulakos and S. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[11] A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[12] W. Lee and N. Magnenat-Thalmann. Head modelling from pictures and morphing in 3d with image metamorphosis based on triangulation. *Modelling and motion capture techniques for virtual environments (Lecture Notes in Artificial Intelligence 1537)*, pages 254–268, 1998.

[13] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. *Proceedings of ACM SIGGRAPH*, pages 369–374, 2000.

[14] S. Moezzi, L. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE Multimedia*, 4(1):18–25, 1997.

[15] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. *IEEE International Conference on Computer Vision*, pages 3–10, 1998.

[16] R. Plankers and P. Fua. Articulated soft objects for video-based body modeling. *IEEE International Conference on Computer Vision*, pages 394–401, 2001.

[17] K. Pulli, M. Cohen, T. Duchamp, H. Hoppe, L. Shapiro, and W. Stuetzle. View-based rendering: Visualizing real objects from scanned range and color data. *Eurographics workshop on Rendering*, pages 23–34, 1997.

[18] C. Seitz and C. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):1–23, 1999.

[19] J. Starck, G. Collins, R. Smith, A. Hilton, and J. Illingworth. Animated statues. *Machine Vision and Applications, Special Issue on Human Modeling, Analysis, and Synthesis*, 2002.

[20] C. Sun. Fast stereo matching using rectangular subregioning and 3d maximum-surface techniques. *International Journal of Computer Vision*, 47(1/2/3):99–117, 2002.

[21] S. Vedula, S. Baker, and T. Kanade. Spatio-temporal view interpolation. *Eurographics Workshop on Rendering*, pages 1–11, 2002.

[22] S. Vedula, P. Rander, H. Saito, and T. Kanade. Modeling, combining, and rendering dynamic real-world events from image sequences. *Proceedings of Virtual Systems and Multimedia*, pages 323–344, 1998.