Virtual view synthesis from multiple view video sequences

J.Starck and A.Hilton

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, GU2 7XH, UK.

Abstract

This paper addresses the synthesis of virtual views of people from multiple view image sequences. We consider the target area of the multiple camera "3D Virtual Studio" with the ultimate goal of capturing video-realistic dynamic human appearance. A mesh based reconstruction framework is introduced to initialise and optimise the shape of a dynamic scene for view-dependent rendering, making use of silhouette and stereo data as complementary shape cues. The technique addresses two key problems: (1) robust shape reconstruction; and (2) accurate image correspondence for view dependent rendering in the presence of camera calibration error. We present results against ground truth data in synthetic test cases and for captured sequences of people in a studio. The framework demonstrates a higher resolution in rendering compared to shape from silhouette and multiple view stereo.

Key words: Visual scene reconstruction, View dependent rendering

PACS:

1 Introduction

The challenge of creating realistic computer generated scenes is leading to a convergence of computer graphics and computer vision technology. Where computer graphics deals with the complex modelling of objects and simulation of light interaction in a virtual scene to generate realistic images, computer vision offers the opportunity to capture and render such models directly from the real-world with the visual realism of conventional video images, illustrated in Figure 1.

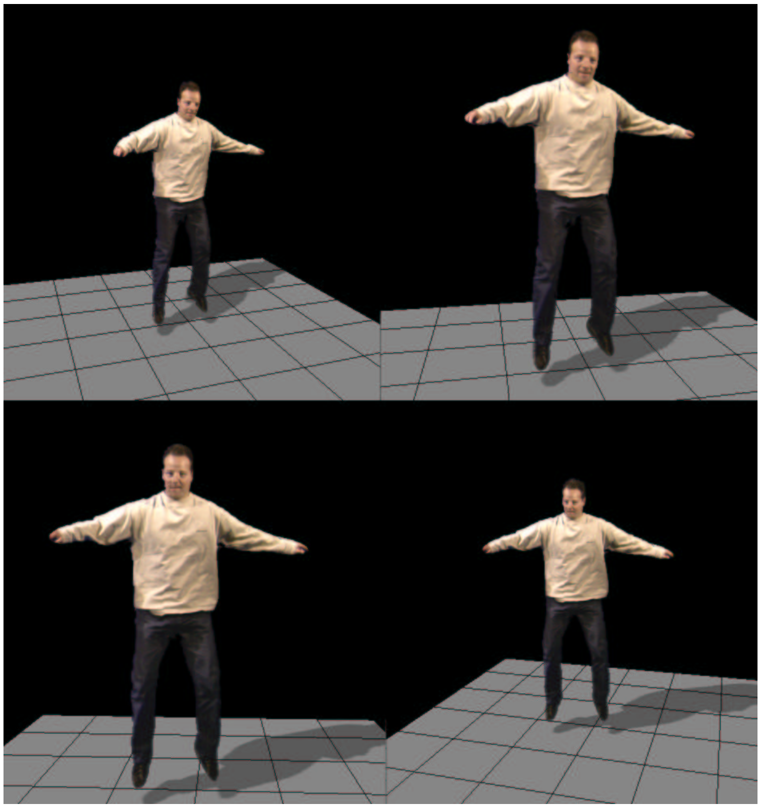


Fig. 1. Panning a virtual camera for a frame of a multiple view video sequence captured in an 8 camera 3D Virtual Studio.

One of the key challenges lies in the creation of realistic human models, a central component of most visual media. In recent years the problem of generating visually realistic graphical models of people has been addressed through computer vision techniques. Three-dimensional production from camera images was popularised by Kanade et al. [12] who coined the term "Virtualized Reality". Kanade et al. [12] demonstrated the ability to reconstruct dynamic scenes of people from multiple view video sequences, allowing a recorded event to be visualised in 3D as an immersive viewing experience.

The concept of using multiple cameras for 3D production of human actors is now being explored in the broadcast industry [22,9]. Traditionally the use of computer graphics in broadcast has centred on the virtual studio in which a camera films live action against a constant background such as a blue screen. Virtual studio technology has now developed to the point where the camera footage can be overlaid with real or synthetic video in real-time [9]. The use of multiple cameras in this setting opens up the potential for the "3D Virtual Studio" in which the dynamic shape and appearance of a person can be captured as a 3D computer graphics model.

There are two strands of research in computer vision that will potentially lead to the realisation of image based human appearance capture in a 3D Virtual Studio. Firstly marker-free visual motion capture in which the motion of a person is tracked from multiple cameras. Secondly accurate scene reconstruction to capture the dynamic shape and appearance of a person moving in the studio. In this paper we address the problem of reconstructing the geometric shape and rendering the appearance of a person from multiple view video sequences. There are several important considerations in our approach to the problem:

(1) Robust 3D shape estimation in the presence of reconstruction ambiguities

such as self occlusions between camera images and a limited variation in surface appearance; and

(2) Recovering sub-pixel image correspondence in the presence of inexact camera calibration for rendering with the visual quality of the original images.

In this work we seek to render a virtual view of a person with the greatest fidelity to the original camera images captured in a multiple camera studio. We present and discuss a new technique to reconstruct geometry for view-dependent rendering of appearance. A robust initial estimate of the scene is first derived from multiple view silhouette images. We then update geometry to recover the correspondence for rendering with a view-dependent appearance. The geometry is optimised in a coarse to fine framework, integrating both stereo and silhouette data. The stereo matches are then used to provide the sub-pixel accurate image correspondence to render the virtual view. This approach is compared to current methods that use either silhouette or stereo data alone without consideration of image correspondence with inexact camera calibration. We demonstrate that the technique provides a more robust estimate of geometry and greater visual fidelity to the camera images in rendering virtual views of people from multiple view video sequences.

2 Related work

The acquisition and rendering of visually realistic images of real objects and scenes has been a long standing problem in both computer graphics and computer vision. There are two contrasting approaches to the problem: *image-based modelling* and *image-based rendering*. In modelling from images, a 3D surface model is constructed for a scene and texture maps are extracted from the images. The advantage of this approach is that it allows the model to be rendered and manipulated in a conventional computer graphics pipeline. The disadvantage lies in the quality of the geometric reconstruction that can be achieved from images and the fixed appearance given by the model texture. Image-based rendering on the other hand synthesises novel views directly from the original images rather than through explicit reconstruction of scene geometry. This provides greater visual fidelity to the original data at the cost of requiring dense sampling of the scene for view synthesis. In this work we use 8 fixed cameras to capture dynamic sequences in a studio, making image based rendering unfeasible without a restrictive range of virtual viewpoints.

2.1 Image based modelling

Geometric modelling from images is a central problem in computer vision and techniques have been developed to automate the process of scene reconstruction [10]. The classical approach to 3D reconstruction developed first in photogrammetry attempts to jointly estimate 3D structure and camera viewing parameters through a process termed *bundle-adjustment* [31]. In visual scene reconstruction this can be simplified by calibrating the viewing parameters of the cameras. The problem is then to solve for the 3D shape of the scene that reproduces the images. Techniques for shape estimation from multiple cameras include reconstruction of volume from image silhouettes, termed the *visual hull* [15], volume from colour consistency between images, termed the *photo hull* [25,14], and surface recovery from stereo correspondence between pairs of camera images [12,20].

Multiple camera systems have been developed to reconstruct dynamic sequences of people, Moezzi et al. [19] demonstrated the use of the visual hull, Vedula et al. [32] made use of the photo hull, and Kanade et al. [12] fused multiple stereo depth-maps into a surface model of a person. There are several important limitations for these techniques for visual reconstruction. The visual hull provides only a bounding approximation to a scene from silhouettes [15]. Matching colour between images in the photo hull can refine the estimated shape, however colour consistency techniques suffer from holes or false cavities in the volume of a scene where consistency cannot be correctly estimated between views, and the fattening of areas of the scene where there is insufficient colour information in the images to distinguish different surfaces [6]. Finally stereo correspondence can fail in regions of poor image texture or occlusion boundaries and can produce noisy depth estimates with inexact matches between images.

In this paper we present a technique to integrate multiple visual cues in scene reconstruction to provide improved reconstruction in the presence of visual ambiguities. Multiple shape cues have been ysed previously for reconstruction in computer vision. Terzopoulos [30] introduced a model-based approach to visual reconstruction in which a prior model is deformed to satisfy multiple constraints on shape. Model-based reconstruction of whole-body human models from silhouettes has been presented by Hilton et al. [11] and from silhouettes, stereo and feature data by Starck and Hilton [27]. A model-based approach relies on a prior scene model and in the case of human shape reconstruction this model must be articulated to match the pose at each frame of a video sequence [27].

Fua and Leclerc [7] introduced an object-centred approach to reconstruction in which an initial surface estimate rather than a prior model is optimised to match multiple shape cues. The technique used an initial surface derived from stereo and updated the surface to match both stereo and shading between images. Vedula et al. [33] used a similar technique to reconstruct human shape where an initial surface derived from multiple view stereo is used to refine the search range for stereo correspondence. These techniques make use of reconstructed geometry to improve the estimation of image correspondence and remove the need for a prior scene model. In this work we adopt an object-centred approach using the visual hull as a robust initial estimate of shape. The geometry is then optimised to match both silhouette and stereo data as complementary shape cues. The shape of the model is also used to constrain the search for stereo correspondence in a coarse-to-fine framework giving a wider range of convergence compared to local optimisation techniques [7].

2.2 View-dependent rendering

Hybrid techniques have been introduced to combine an image-based representation of appearance with geometric reconstruction. These techniques provide the visual quality of image based rendering, making use of reconstructed scene geometry to interpolate from a sparse set of camera images. Debevec et al. [4] introduced the concept of *view-dependent texturing* from photographs and demonstrated the visual realism that can be achieved in rendering using the camera images as multiple texture maps. Pulli et al. [23] introduced the concept of both view-dependent geometry and texture for general view-dependent rendering. These techniques provide the visual quality of the captured images closest to the novel viewpoint in rendering and highly realistic virtual views of people have been demonstrated [24,18,17,32].

Several problems remain for virtual view synthesis from multiple view video sequences: (1) robust shape reconstruction; and (2) accurate image correspondence for view dependent rendering. Current techniques for view generation rely on a single reconstructed scene model and render novel views from the camera images using the scene geometry. While this has enabled the capture and rendering of realistic 3D scenes, ambiguities in reconstruction can lead to incorrect scene geometry and errors in correspondence due to either incorrect geometry or inexact camera calibration. These become apparent as a misalignment and blurring in the rendered images, as shown later in Figure 7.

In this paper we present a technique to obtain a robust estimate of surface geometry using shape from silhouette and then to optimise estimated surface geometry to match the appearance across the camera images used in view-dependent rendering. The technique provides the shape and sub-pixel accurate image correspondence for rendering a view-dependent appearance. This gives increased resolution by correctly aligning image texture from multiple views, reducing the blur and misalignment of features compared to approaches that make the assumption that an estimated surface is in correspondence between images [19,32,12].

3 Image based reconstruction

In this section we describe the multiple camera 3D Virtual Studio for the acquisition of multiple view video sequences. Algorithms are then presented for reconstruction of shape from image silhouettes and stereo correspondence as a basis for comparison with the object-centred optimisation technique introduced in this paper.

3.1 Data acquisition

Video sequences are recorded from 8 cameras in a dedicated studio. Sony DXC-9100P 3-CCD colour cameras are used, providing PAL-resolution progressive scan images at 25Hz. The cameras are synchronised by an external trigger and the RGB analogue output is converted to a time-stamped digital SDI stream. The video is stored to disk using multiple frame grabbers on a PC network. The studio is equipped with a lighting grid to provide controlled lighting conditions and a blue curtain for background segmentation. All cameras are colour calibrated by white-balancing the RGB output with a white reference object. The studio set-up provides 8 channels of synchronised broadcast standard digital video capture.

The cameras are positioned to provide a frontal ring of 7 cameras, giving 6 pairs for stereo matching. The final camera is mounted on the ceiling to increase the intersection angle between views for reconstruction of the visual hull. The cameras provide a capture volume of approximately 2.5m × 2.5m × 2.5m with a frontal viewing range in the order of 120^o surrounding the volume. The intrinsic and extrinsic camera parameters are calibrated using the Camera Calibration Toolbox for Matlab from MRL-Intel [1]. The source code for the implementation of the toolbox is available in the Open Source Computer Vision library distributed by Intel [2]. Camera calibration provides a worst case reprojection error of 1.6 pixels averaged across the cameras, equivalent to a reconstruction error in the order of $5mm$ to $10mm$.

An image silhouette describes an occluding contour that encloses the projected shape of an observed scene. The visual hull is reconstructed through the volume intersection of the occupied region of 3D space represented by multiple image silhouettes [15]. Techniques for volumetric reconstruction of the visual hull in general use a discrete representation of space as a set of volume elements or voxels [6,26]. The voxels corresponding to the visual hull are extracted by intersecting the visual cones for the silhouettes. This intersection test, also called the *voxel occupancy problem*, is performed by projecting voxels to each image in turn and testing the overlap with the silhouettes [29].

The visual hull reconstruction algorithm used here is outlined in Algorithm 1. A voxel grid is first defined in the studio capture space. The set of occupied voxels that lie inside the visual hull are then derived by testing the projected overlap of each voxel with the silhouettes. If the projected shape of a voxel overlaps all the silhouettes it is set as occupied, otherwise if a voxel falls outside any silhouette the voxel is set as unoccupied. The image region corresponding to a voxel is simplified as the rectangular region enclosing the projected corners of a voxel. These image regions can be pre-computed to speed up the procedure. The surface voxels for a scene are finally extracted as the set of occupied voxels that are adjacent to unoccupied voxels. This discrete representation is converted to a surface mesh by iso-surface extraction using a variation on the *Marching Cubes* algorithm [16].

Algorithm 1: Visual hull reconstruction

```
set (all voxels = occupied)
for (each voxel)
    for (each image)
        project (each voxel corner to image)
        set (image region containing voxel corners)
        if (no silhouette pixels in image region)
            set (voxel = unoccupied)
for (each voxel)
    if (voxel = occupied)
        if (connected voxel = unoccupied)
            set (surface voxel)
extract (iso-surface for occupied voxels)
```

3.3 Shape from stereo correspondence

Surface reconstruction from stereo is performed by extracting a 2.5D stereo depth-map for each camera pair in the studio. Here we use a two-stage dynamic programming technique proposed by Sun [28] to extract a surface that maximises the stereo correspondence between images and enforces continuity in the depth-map. We use a normalised cross-correlation metric to allow for linear changes in intensity between images with non-Lambertian surfaces or inexact intensity matched images. We also add the constraint that the disparity range for stereo correspondence lies within the visual hull extracted from image silhouettes. This follows the model-enhanced stereo paradigm proposed by Vedula et al. [33] and removes outliers in stereo correspondence.

Multiple 2.5D depth-maps are fused into a single 3D surface representation using volumetric fusion similar to that proposed by Narayanan et al. [20]. The fusion technique outlined in Algorithm 2, averages the depth to the surface of the scene at a discrete set of points on a volumetric grid and extracts the shape of the scene as the zero-distance surface inside the volume. Here the discrete volume defined by the visual hull is used. At each occupied voxel in the visual hull a 3D depth value is derived. The depth value is calculated by projecting the voxel to all the depth maps and searching for the closest 3D surface point in each view. An average is then taken for the depth to each 3D point within a set tolerance of the closest surface point across all views. The tolerance is automatically set as the size of the voxels used in volumetric fusion in order to average the surfaces that fall within each voxel. A signed distance function is constructed by assigning positive distance values where the depth of the voxel to the camera view-point is less than the distance in the depth-map and a voxel lies outside the surface. A negative depth value is otherwise assigned where a voxel falls inside the surface. The surface of the scene is then extracted as the zero-valued iso-surface of the distance function using the marching cubes algorithm [16].

Algorithm 2: Fuse multiple stereo depth-maps

```
for (each voxel)
    set (closest distance as undefined)
    if (voxel inside visual hull)
        for (each 2.5D depth image)
            project (each voxel corner to image)
            get (average depth inside voxel)
            set (distance = depth to voxel - average depth)
            if (|| distance || < closest)
                set (closest = || distance ||)
    if (closest is undefined)
        set (voxel distance = positive)
    else
        average (distances within tolerance of closest)
extract (iso-surface from signed distances)
```

4 Surface optimisation for rendering

In this section we describe the object-centred multiple view optimisation algorithm. We start with an initial estimate of shape using a mesh generated for the visual hull as outlined in Algorithm 1. We then update the mesh to minimize the error in fitting to both stereo and silhouette data. Shape optimisation is performed in a regularised coarse-to-fine framework. The result is a regularised mesh that fits the data, aligning available texture between camera images and returning sub-pixel accurate image locations for rendering.

The surface optimisation technique follows the physically based deformable model framework proposed by Terzoupoulos [30]. A cost-function is constructed consisting of a potential energy term derived from the fit of the model to the data, and an internal energy term that penalises the deviation from the desired model properties. The model is then deformed to minimize the total energy function, hence minimizing the error between the model and the data while the internal energy regularises the model deformation. In data fitting we use the cost of fitting to stereo data E_S and matching the shape from silhouette provided by the visual hull E_V. The trade-off between these data terms is governed by a weighting β, and the influence of model regularisation, E_R, is governed by α.

$$E = \beta E_S + (1 - \beta) E_V + \alpha E_R \tag{1}$$

We discretize the energy function at the vertices of our mesh $\underline{x}_i$ and use gradient descent for minimization. In terms of physics-based deformable models this is equivalent to a zero mass dynamic system. The deformation of the mesh vertices is then given as.

$$\frac{d\underline{x}_i}{dt} = -\nabla E = -\left(\beta \nabla E_S + (1 - \beta) \nabla E_V + \alpha \nabla E_R\right) \tag{2}$$

4.1 Silhouette data

The visual hull is used to provide an initial estimate of the scene geometry for optimisation, we then seek to update the estimated geometry to match stereo data. Stereo matching can however fail where there is a limited variation in image appearance or where there is significant distortion in appearance between views due to projective distortion or occlusion boundaries. Silhouette data is therefore incorporated by fitting the volumetric visual hull as well. A data energy term for the visual hull, $E_V(\underline{x}_i)$, is defined as the squared error between the vertex position and the closest surface voxel on the visual hull $\underline{y}_i$ derived using Algorithm 1.

$$E_V = \sum_i (1 - \beta(\underline{x}_i)) \|\underline{y}_i - \underline{x}_i\|^2 \tag{3}$$

4.2 Stereo data

In stereo matching we use a direct search for stereo correspondence between the images used in view dependent rendering. For each mesh vertex we first determine the key view, from the views used in rendering, that has 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 stereo pair that uses the key view. Here we make the simplifying assumption of a fronto-parallel surface at each vertex and use area-based normalized cross-correlation between rectified camera images [8]. For each offset image in each stereo pair we locate the sub-pixel match to the key image with the highest correlation score. We define the search range along the epipolar line in each rectified offset image according to the expected error in the shape of the mesh.

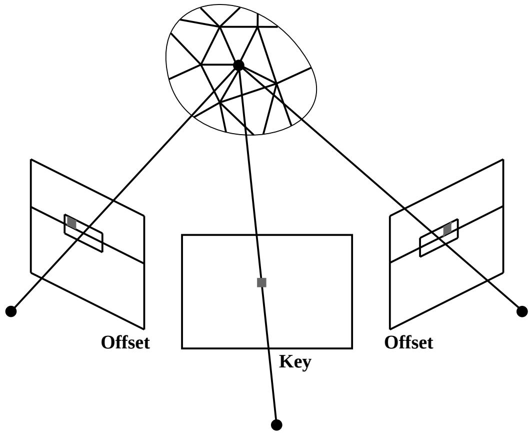


Fig. 2. Stereo matching between key and offset views, showing the search region along each epipolar line allowing for off-axis matches with inexact camera calibration.

We also match up to a specified pixel error perpendicular to each epipolar line according to the expected accuracy of the camera calibration as illustrated in Figure 2. In this work we assume an expected shape error of 200mm and a 2 pixel reprojection error in all cases.

For each vertex we derive a sub-pixel correspondence in each offset image and a reconstructed 3D position. The stereo energy term at each vertex, $E_S(\underline{x}_i)$, is defined as the squared error between the vertex position and the reconstructed 3D position $\underline{z}_{i,c}$ for each offset camera c. We therefore seek a least-square error fit to the matched vertex positions across the whole mesh, as given in Equation 4. The energy term is weighted according to the quality of the stereo matches as given by the correlation score $w_{i,c} \in [0,1]$, Equation 5. This enables a trade-off between fitting stereo data where good matches are obtained and fitting silhouette data where matching is poor. The weighted stereo term and the corresponding vertex weight are defined as follows where n_i^c indicates the number of offset cameras for which a 3D position is derived.

$$\beta E_S = \sum_i \frac{1}{n_i^c} \sum_{c=0}^{n_i^c-1} w_{i,c} \|\underline{z}_{i,c} - \underline{x}_i\|^2 \tag{4}$$

$$\beta(\underline{x}_i) = \frac{1}{n_i^c} \sum_{c=0}^{n_i^c-1} w_{i,c} \tag{5}$$

In stereo matching it is important to account for self-occlusions to prevent incorrect matches between occluded and visible regions. We deal with self-occlusions by checking the visibility of each mesh vertex in each camera image and only matching between unoccluded views. Here we use the visibility algorithm introduced by Debevec et al. [5] that uses hardware accelerated OpenGL rendering. To test the visibility in a camera, the mesh is rendered to the camera viewpoint with a unique colour ID assigned to each polygon. For each front-facing vertex we can then retrieve the polygon at the projected location in the camera and check for occlusion against the polygon.

4.3 Surface regularisation

The standard approach to shape regularisation is to treat a deformable model as a thin-plate material under tension [30]. Here we use membrane tension for regularisation. The membrane functional for E_R across a parameterised surface $\underline{x}(u,v)$ is given in Equation 6 and the variational derivative is given by the Laplacian $\triangle(\underline{x})$. Under the simplifying assumption of a regular mesh parameterisation, the Laplacian at a mesh vertex is given by the "umbrella-operator" in Equation 7 where the index v spans the 1-neighbourhood $\underline{x}_{i,v}$ of a vertex $\underline{x}_i$ [13].

$$E_R = \int \int \left(\|\underline{x}_u\|^2 + \|\underline{x}_v\|^2 \right) dudv \tag{6}$$

$$\frac{dE_R}{d\underline{x}_i} = -\frac{1}{n_i^v} \sum_{v=0}^{n_i^v - 1} (\underline{x}_{i,v} - \underline{x}_i) \tag{7}$$

The effect of the umbrella operator is to pull vertices towards the centroid of the 1-neighbourhood. Intuitively the role of regularisation is to maintain a smooth, even parameterisation of the mesh surface during deformation. The weighting parameter α introduced in Equation 1 provides user defined control over the degree of regularisation required in data fitting. The exact value of the parameter required will depend both on the accuracy of the data used in fitting, as well as the triangulation of mesh due to the scale dependence of the umbrella operator. Throughout this work a fixed value of $\alpha = 10.0$ is used to compare results.

4.4 Surface optimisation

The shape optimisation process is performed in a coarse-to-fine framework in order to deal with noisy stereo matches. The geometry is initialised as the surface shape of the visual hull, this surface mesh is then optimised to match the appearance between the images used in view-dependent rendering. We start at an initial expected error for the surface mesh and locate the stereo matches up to the error estimate in all the cameras pairs for which a camera is to be used in rendering. We then update vertex locations to minimize the energy function. The expected error is successively reduced and the model again deformed to minimize the energy. The convergence criteria at each error level is set to the error estimate multiplied by the error reduction factor such that the assigned data remains within the matching range at the next iteration of mesh deformation. Optimisation is stopped when the error level reaches the reconstruction accuracy of the camera set-up. The advantage of the coarse-to-fine matching and model deformation is that it allows the mesh to converge to a solution, allowing for noisy stereo matches and an increased range of convergence compared to local optimisation techniques [7]. The technique also provides sub-pixel accurate image correspondence for the mesh vertices up to the expected calibration accuracy of the camera system for subsequent rendering.

Algorithm 3: Coarse to fine surface optimisation

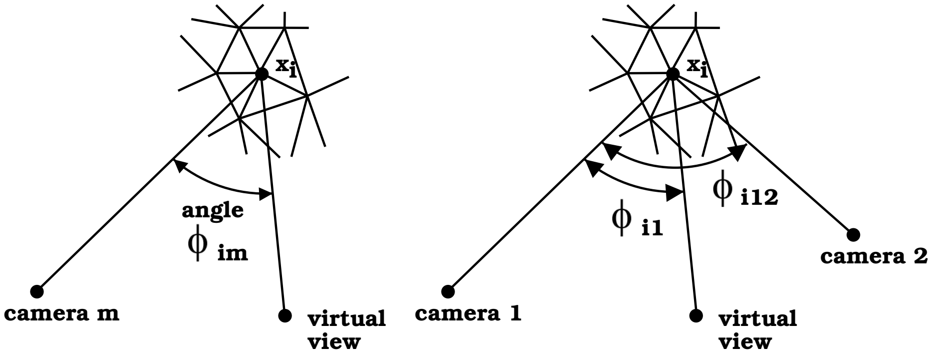
```
(*extract surface mesh for visual hull*)
set (*camera views for rendering*)
set (*initial expected shape error*)
while (*error* $>$ *reconstruction accuracy*)
    set (*mesh visibility in camera images*)
    for (*each mesh vertex* $\underline{x}_i$)
        set (*closest surface voxel* $\underline{y}_i$)
        for (*each camera used in rendering*)
            set (*stereo matches* $\underline{z}_{i,c}$)
    while ($\|\frac{d\underline{x}_i}{dt}\| \geq \eta \times error$)
        set ($\frac{d\underline{x}_i}{dt} = -\beta \frac{dE_S}{d\underline{x}_i} - (1-\beta)\frac{dE_V}{d\underline{x}_i} - \alpha \frac{dE_R}{d\underline{x}_i}$)
    set ($error = \eta \times error$)
```

5 View-dependent rendering

In view-dependent rendering the original camera images are used as a set of texture maps for a surface mesh and blended dynamically according to the view-point used. The input to view-dependent rendering is the image plane correspondence for each model vertex in the camera views. A triangle-centred vertex weight is defined to blend between the texture for each camera. A vertex-centred weight is also defined in order to derive surface colour where the mesh cannot be textured. The mesh is then rendered in multiple passes, first using a view-dependent colour and then texturing from the camera images.

In rendering we use the camera images closest the virtual view to provide the view-dependent appearance. A blend weight b_{im} is calculated at each vertex i for each image m. The proximity of a camera to the virtual view is defined as the cosine of the angle from the camera viewing direction to the viewing direction of the virtual view, $b_{im} = \cos\phi_{im}$ [23,21] as shown in Figure 3. The vertex weights are normalised to sum to one across all visible views $\hat{b}_{im} = v_{im}b_{im}/\sum_m v_{im}b_{im}$. The colour at each model vertex $\underline{I}_i$ is defined as the weighted average of the image colour in each image, $\underline{I}_i = \sum_m \hat{b}_{im}\underline{I}_{im}$. Some vertices may be occluded in all camera views, in which case a vertex colour cannot be derived. Each vertex with no colour assignment is therefore iteratively assigned an average of the adjacent vertex colours to give a complete description of the surface appearance.

Techniques for view-dependent texturing make use of the subset of the available camera images closest to the rendered viewpoint [5,18,23]. In the general case where cameras are located at arbitrary positions in space, camera se-



(a) Vertex colour weighting (b) Vertex texture weighting

Fig. 3. The virtual viewing angles used to define (a) view-dependent vertex colour, and (b) view-dependent texture weighting.

lection has been based on the three closest cameras surrounding the virtual viewpoint. In our studio, cameras are located in a circle in order to surround a person from a limited set of views. The two closest cameras to the desired virtual view are therefore selected for view-dependent texturing. A view-dependent weight is derived at the triangle vertices of the mesh to define the relative influence of these two closest views in texturing each triangle.

The view-dependent vertex weight b_{imf} for each vertex i on each triangle facet f is again defined by the proximity of the camera viewing direction to the virtual view given by the angle ϕ_{im}. The blend weight is now defined as $b_{imf} = \cos\phi_{im} - \cos\phi_{i12}$, where ϕ_{i12} is the angle between the two viewing directions to the cameras used for texturing at the vertex as proposed by Pighin et al. [21], shown in Figure 3. Blending now favours the original camera views exactly when the virtual viewing direction is coincident with a camera viewing direction. The view-dependent vertex weight is set to zero, $b_{imf} = 0$, if any of the vertices are not visible in the camera view or if any of the vertices project to the segmented background region of the image. The vertex weights are finally normalised to sum to one across the two texture views to give $\hat{b}_{imf}$.

The virtual view is generated using hardware accelerated OpenGL rendering. The mesh is first rendered with the vertex colours $\underline{I}_i$. Multi-pass texturing is then used to render the mesh from each camera image with the texture modulated by the blend weights $\hat{b}_{imf}$ at each polygon vertex. In the first instance of texturing a polygon, blending replaces the colour rendered mesh and subsequent passes add modulated texture.

6 Results

In this section we present and discuss the performance of the proposed technique for the estimation and rendering of view dependent geometry and appearance from multiple view video sequences. The technique is first compared against ground truth data for a synthetic test case. Results are then presented for dynamic sequences of people captured in the 3D Virtual Studio.

6.1 Comparison with ground truth

We now consider the idealised problem of reconstructing the shape and appearance of a simple cube object located within the capture volume for the studio. Idealised camera views are rendered for the cube as illustrated in Figure 4. A colour photograph is texture mapped onto the cube to provide a surface appearance for stereo matching. This test case provides ground truth data to test the reconstruction and rendering, allowing an objective evaluation of the proposed technique in comparison with shape from silhouette and multiple view stereo.

Volumetric reconstruction of the visual hull is performed with a voxel size of

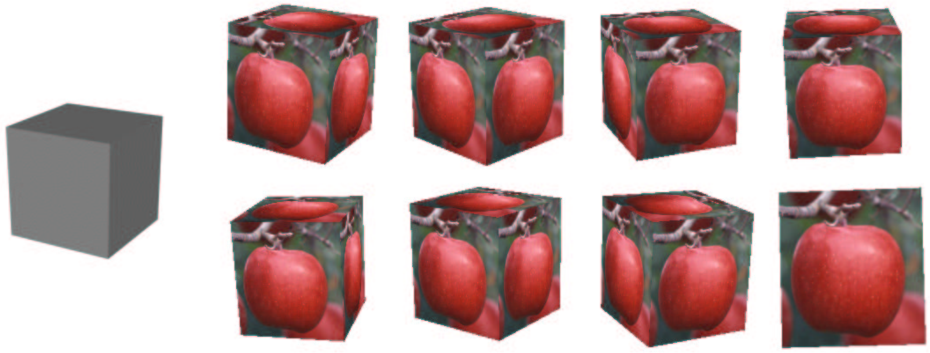
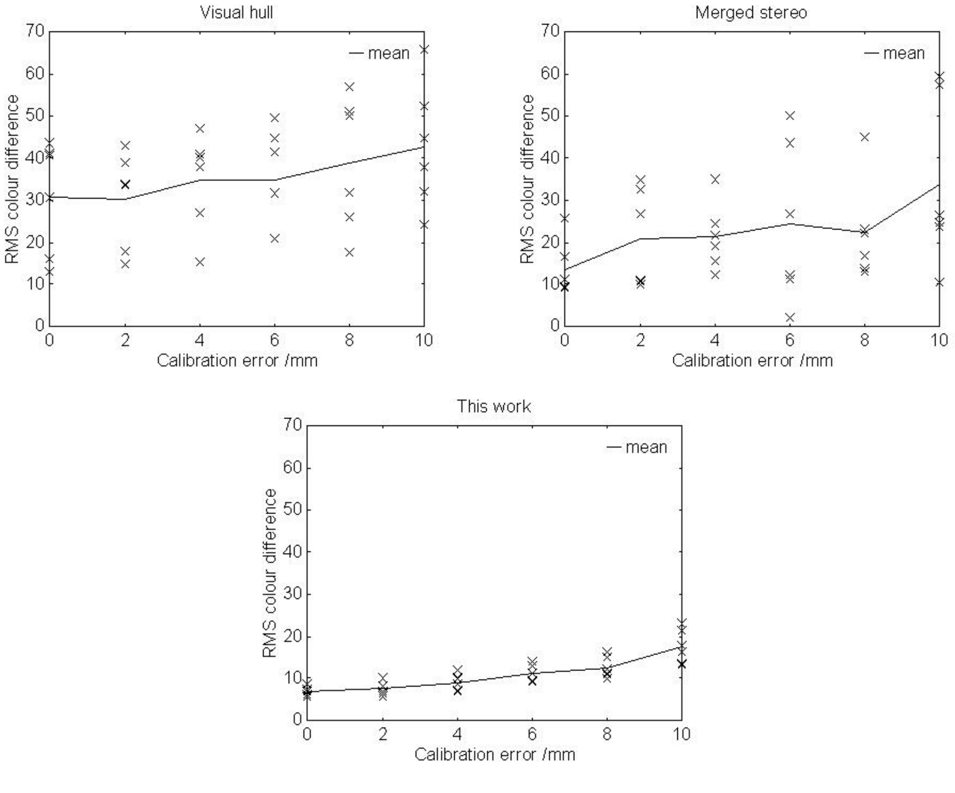


Fig. 4. Synthetic views rendered with an ideal camera model for a 1m cube located at the centre of an 8 camera studio.

10mm to encompass the reprojection error in the real camera data. Camera calibration error is simulated in the images by adding a fixed displacement of up to 10mm in a random direction to the cube when rendering the synthetic camera views shown in Figure 4. The image correspondence for rendering the visual hull and merged 2.5D stereo depth-maps is obtained using the projected image plane location of the mesh vertices.

An objective measure is proposed to assess the quality of the rendered images. It is not possible to make a comparison directly with views synthesised from the original cube as we cannot expect the camera images to reproject to the same original position in the presence of calibration error. Instead we note that the quality of the rendered images is degraded by the misalignment of the multiple camera textures used in view-dependent rendering. This produces the blurring and double exposure effects shown later in Figure 7. We therefore define an error metric that measures the difference in the RGB colour between the two camera textures used in rendering. The camera textures are first rendered independently to the virtual view-point without modulation. The root mean square RMS difference between the RGB colour is then computed across all pixels for which the colour is defined in both rendered images. For the exact colour balanced cameras used in the synthetic images we can expect that the RMS colour difference will be minimised where the camera textures are in alignment.

Six virtual viewpoints are considered, positioned midway between the 7 cameras forming the frontal ring in the studio. The RMS difference in rendering is shown in Figure 5, for a range of simulated calibration errors. The graphs quantify the misalignment in the camera textures used in view-dependent rendering for the visual hull, merged 2.5D stereo depth-maps and the optimised shape of the visual hull derived in this work. The reconstructed shape for one virtual view is shown in Figure 6 and the rendered images shown in Figure 7 for simulated errors of 0mm, 5mm and 10mm.


Visual hull

Merged stereo

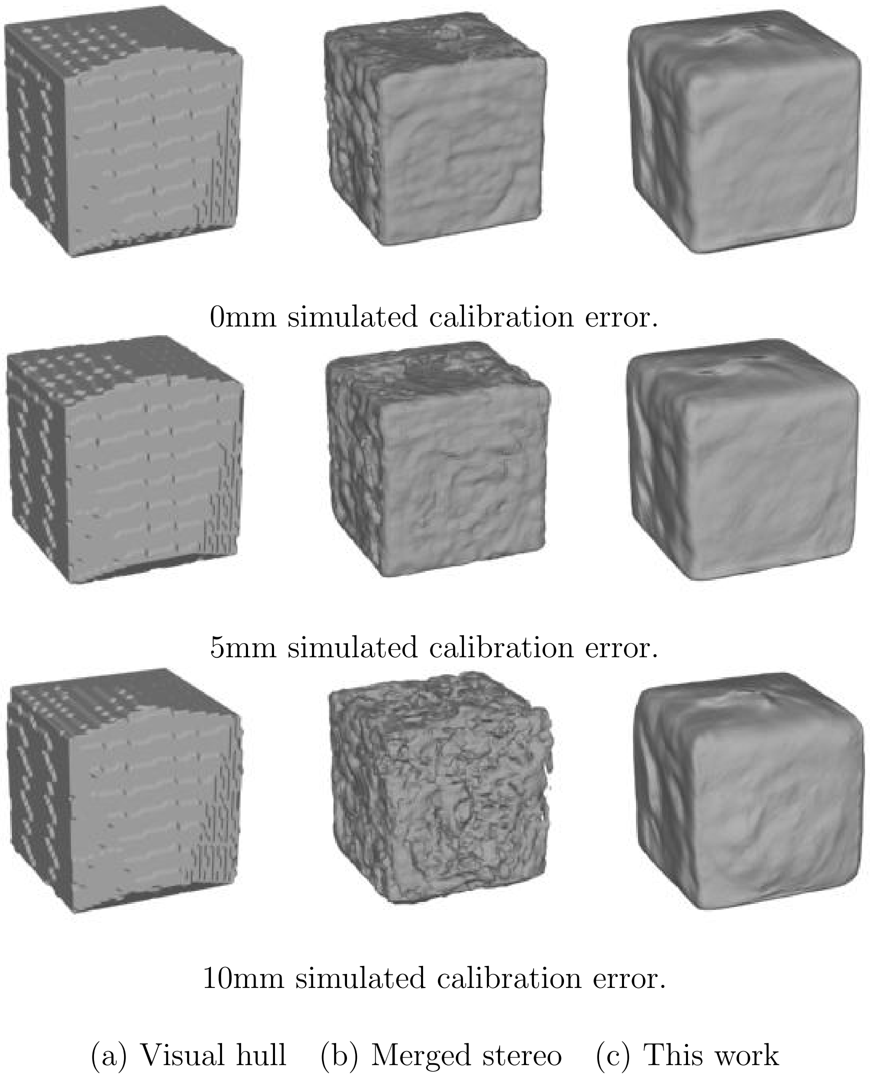
This work

Fig. 5. The RMS RGB colour difference between textures used in view-dependent rendering for a range of simulated calibration errors.

The quantified colour difference shown in Figure 5 illustrates the performance of the proposed technique in comparison with the visual hull and multiple view stereo. The visual hull represents an upper bound on the shape of the cube and so the surface will not be in correspondence even with exact camera calibration. A closer approximation to the underlying geometry can be obtained by making use of a greater number of image silhouettes, however the visual hull cannot reproduce any concavities in a scene which will always be out of correspondence. Stereo matching derives geometry by directly considering image correspondence and provides an improved estimate of geometry in this test case where there is a sufficient appearance variation in the images for matching. However, the accuracy degrades with the simulated error as the technique does not consider the reprojection error in the images and different stereo pairs will also provide different geometry estimates. In this work both silhouette and stereo data are used as complementary shape cues and the image correspondence is derived in the presence of calibration errors. This provides a more robust estimate of geometry and reduces the misalignment of camera images in rendering, providing an improved visual quality in virtual view synthesis.

6.2 Reconstructing and rendering people

The technique is now applied to reconstruct the shape and render virtual views of people from multiple view video sequences. Reconstruction and rendering is first compared with the visual hull and multiple view stereo. Figure 8 shows a rendered viewpoint along with the corresponding geometry derived for each technique. A voxel size of 20mm is now used to merge the noisy stereo depth-maps over a larger region. The visual hull, Figure 8(b) shows the blurring effect



0mm simulated calibration error.

5mm simulated calibration error.

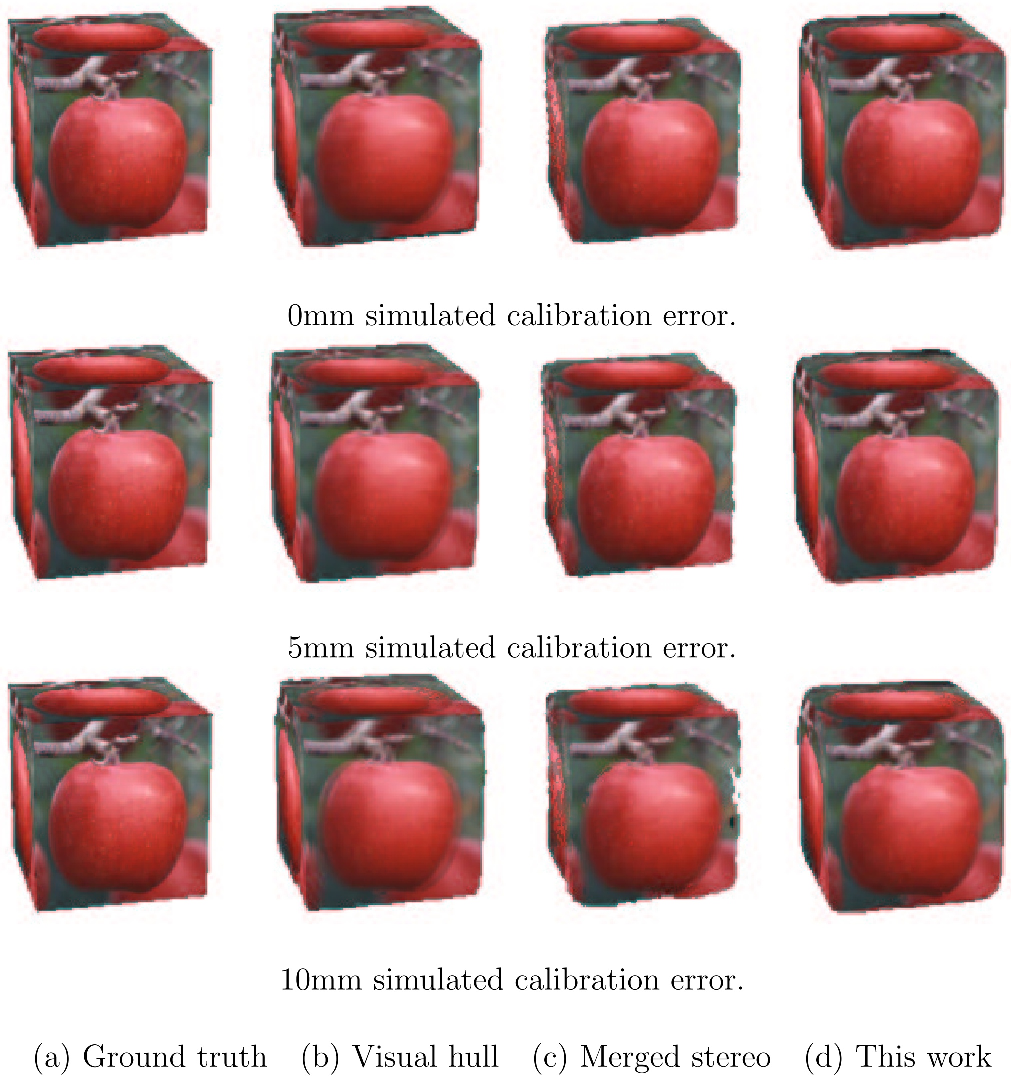
10mm simulated calibration error.

(a) Visual hull (b) Merged stereo (c) This work

Fig. 6. Reconstructed geometry for rendering a virtual viewpoint at 0mm, 5mm and 10mm simulated calibration error.

with incorrect geometry. The merged stereo, Figure 8(c), shows improved resolution but suffers from missing and incorrect sections of geometry due to the lack of appearance variation in the camera images for stereo matching. Figure 8(d) shows the optimised shape of the visual hull using both silhouette and stereo shape cues, and demonstrates the highest resolution with the recovered sub-pixel correspondence.

Sequences of rendered views are now shown Figures 9 for virtual view-points



0mm simulated calibration error.

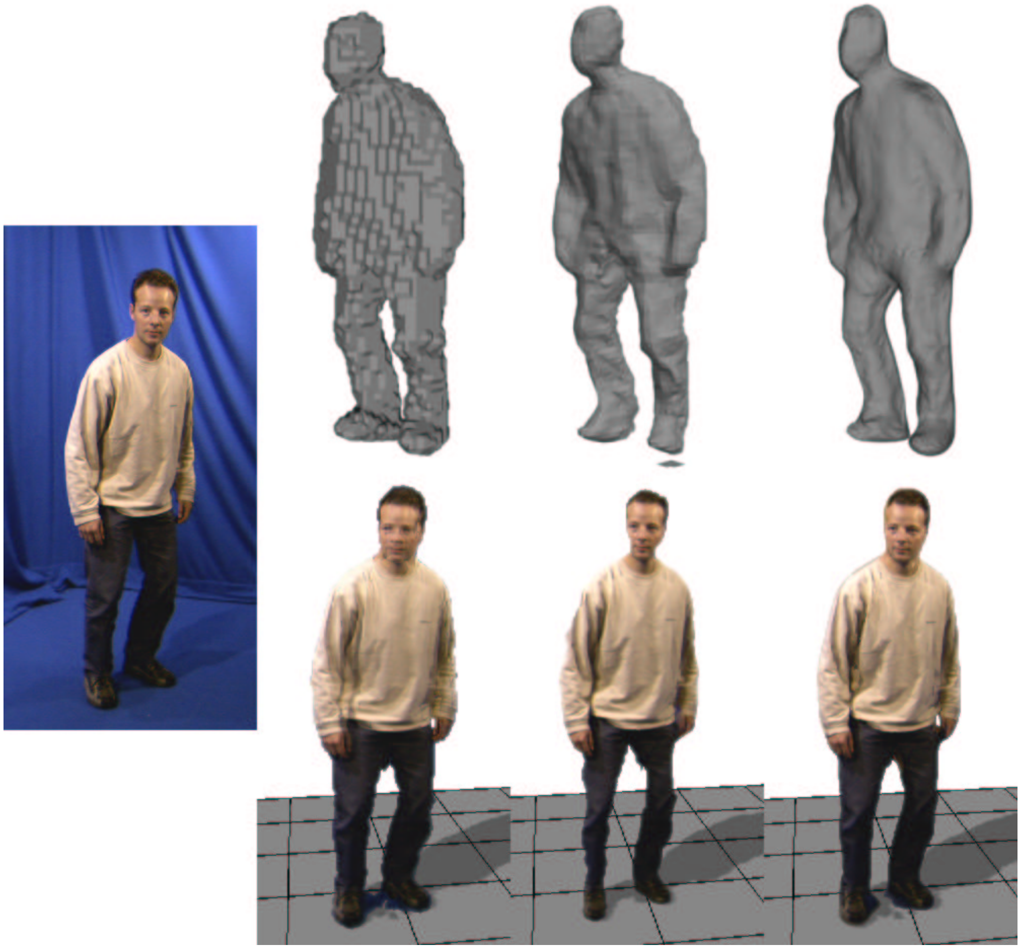
5mm simulated calibration error.

10mm simulated calibration error.

(a) Ground truth (b) Visual hull (c) Merged stereo (d) This work

Fig. 7. Rendered virtual view-point for the visual hull, merged 2.5D stereo depth-maps and this work shown in comparison with the ground truth image at a 0mm, 5mm and 10mm simulated calibration error.

scripted to move and pan around several different dynamic scenes. This demonstrates the flexibility in viewpoint control that is given by the 3D description of the scene. The virtual views approach the resolution of the original camera images and the dynamic appearance of the clothing wrinkles produces a video-realistic result. The corresponding movie sequences can be viewed at [3].



(a) Camera image (b) Visual hull (c) Merged stereo (d) This work

Fig. 8. Rendering a virtual view mid-way between two cameras.

7 Concluding remarks

In this paper we have demonstrated a technique to reconstruct the shape and render the appearance of people from multiple view video sequences captured in a 3D Virtual Studio. The technique considers two important problems, robust scene reconstruction and the recovery of image correspondence for rendering.

The framework makes use of multiple shape cues through shape from silhouette and stereo matching. A robust initial estimate of geometry is derived from

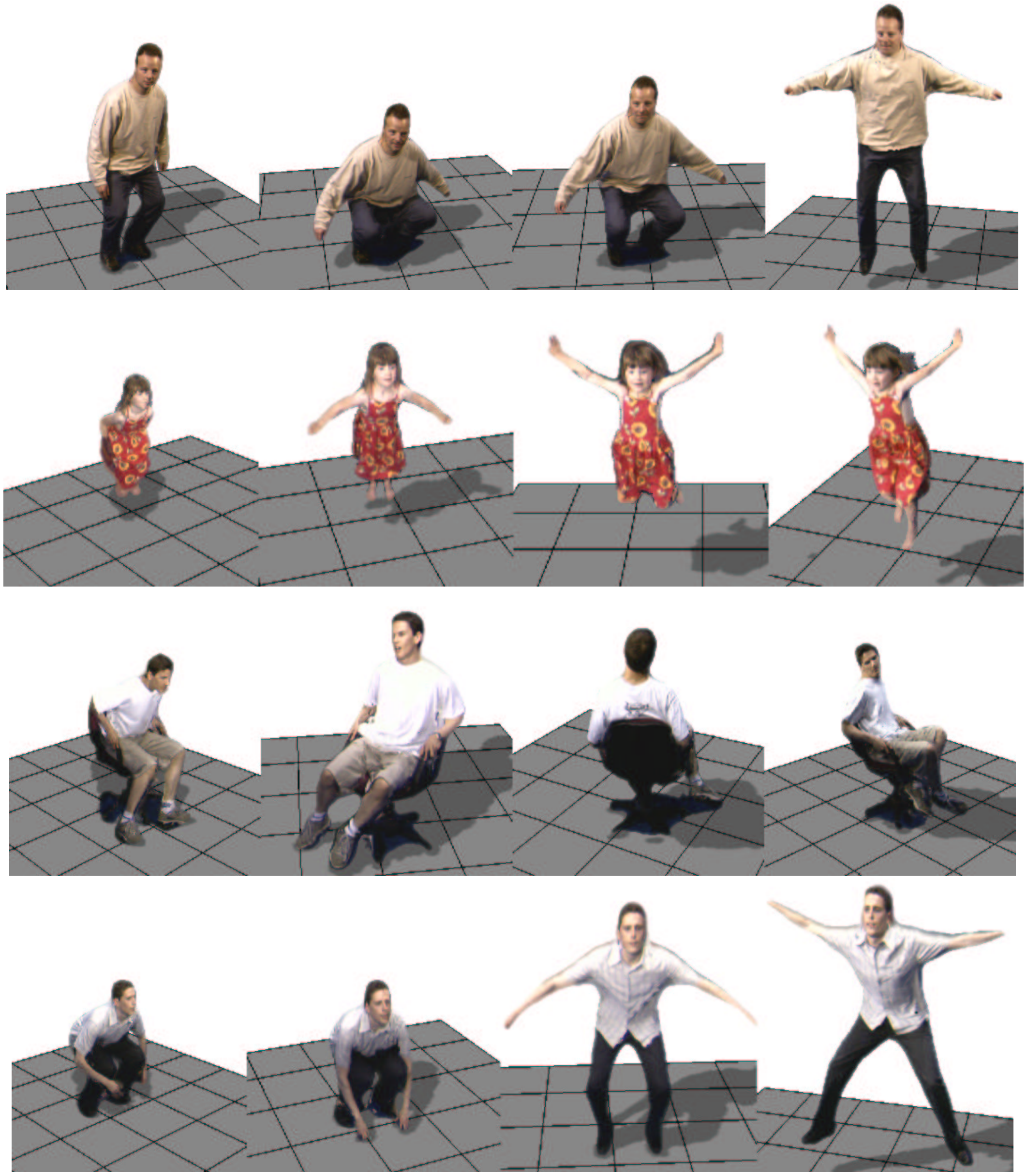


Fig. 9. Virtual views rendered from multiple view image sequences of people.

the visual hull which is then updated to match available stereo and silhouette data as a deformable mesh. The mesh is optimised in a coarse to fine algorithm in which the search range for stereo matches is gradually reduced to the calibration accuracy of the camera system. Surface optimisation provides the means to deform the shape derived from image silhouettes to satisfy stereo matching between views for the recovery of image correspondence. The framework incorporates silhouette data where stereo matches are poor due to lack of appearance variation in the images or occlusion boundaries. This can demonstrate improved shape reconstruction compared to the use of silhouette or stereo data alone.

Current techniques for view generation rely on rendering a novel view using reconstructed scene geometry under the assumption that the scene model is in correspondence between views. Errors in correspondence can arise either due to inaccuracies in reconstruction or inexact camera calibration. This becomes apparent as a misalignment and blurring of texture in rendering. The surface optimisation technique presented in this work returns sub-pixel accurate correspondence for view-dependent rendering in the presence of camera calibration errors. The technique demonstrates improved resolution in rendering virtual views compared to shape from silhouette and multiple view stereo.

We have demonstrated that it is possible to render virtual views from multiple video sequences with a resolution approaching that of the original captured images. This is achieved by optimising the geometry used in rendering to match the texture between the views used for view-dependent texturing and by recovering image correspondence up to the expected calibration accuracy of the camera images. The iterative optimisation technique runs as an off-line process and so can be considered as a potential post-production tool to improve the resolution in virtual view synthesis from multiple view video.

One limitation of the framework currently lies in the number of cameras required for reconstruction. In this work short baseline stereo matching is used to derive image correspondence, limiting the relative position of the cameras in the studio. A viewing range of 120^o is achieved with 8 cameras, however this would require 19 cameras to achieve full 360^o coverage. Future research should address the problem of recovering image correspondence from wide baseline camera positions allowing for a greater viewing range for a limited set of cameras.

The framework is also limited to considering each time frame of the multiple view video sequences independently. There is therefore no structure in the temporal sequence of surface meshes to edit or re-purpose the recorded event. Future research should address spatio-temporal image correspondence to deform a surface mesh over time as well as space, or the use of a model based approach to give a consistent structure with a prior surface model.

Acknowledgements

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

References

[1] *Camera Calibration Toolbox.* www.vision.caltech.edu / bouguetj / calib-doc.

[2] *Open Source Computer Vision Library.* www.intel.com / research / mrl / research / opencv.

[3] *Visual Content Production.* www.ee.surrey.ac.uk / Research / VSSP / VMRG / VCPhuman.html.

[4] P. Debevec, C. Taylor, and J. Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. *SIGGRAPH*

Conference Proceedings, pages 11–20, 1996.

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

[6] C. Dyer. Volumetric Scene Reconstruction from Multiple Views. In Davis, L.S (ed.). *Computational Models of Visual Processing*, pages 469–489, Kluwer, Boston. 2001.

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

[8] A. Fusiello, E. Trucco, and A. Verri. Rectification with unconstrained stereo geometry. *8th British Machine Vision Conference*, pages 400–409, 1997.

[9] O. Grau. G. Thomas. 3D image sequence acquisition for TV & film production. *1st International Symposium on 3D Data Processing Visualization and Transmission*, pages 320–326, June 2002.

[10] R. Hartley and A. Zisserman. *Multiple-View Geometry in Computer Vision.* Cambridge University Press, Cambridge, UK, 2000.

[11] 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.

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

[13] L. Kobbelt, S. Campagna, J. Vorsatz, and H.P. Seidel. Interactive multi-resolution modeling on arbitrary meshes. *SIGGRAPH Conference Proceedings*, pages 105–114, August 1998.

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

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

[16] W.E. Lorensen and H.E. Cline. Marching cubes: a high resolution 3d surface reconstruction algorithm. *SIGGRAPH Conference Proceedings*, 21(4):163–169, 1987.

[17] T. Matsuyama and T. Takai. Generation, visualization, and editing of 3d video. *1st International Symposium on 3D Data Processing Visualization and Transmission*, pages 234–245, June 2002.

[18] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. *Eurographics Workshop on Rendering*, pages 115–125, 2001.

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

[20] P.J. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. *6th IEEE International Conference on Computer Vision*, pages 3–10, 1998.

[21] F. Pighin, J. Hecker, D. Lischinski, and R. Szeliski. Synthesizing realistic facial expressions from photographs. *Proceedings of ACM SIGGRAPH*, pages 75–84, 1998.

[22] 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*, pages 348–356, 2002.

[23] 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. *8th Eurographics workshop on Rendering*, pages 23–34, 1997.

[24] J. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade. Appearance-based virtual view generation of temporally-varying events from multi-camera images in the 3d room. *2nd International Conference on 3-D Digital Imaging and Modeling*, pages 516–525, 1999.

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

[26] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer. A survey of methods for volumetric scene reconstruction from photographs. *Proceedings of the Joint IEEE TCVG and Eurographics Workshop*, pages 81–100, 2001.

[27] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. *IEEE International Conference on Computer Vision*, 2003.

[28] 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.

[29] R. Szeliski. Rapid octree construction from image sequences. *Computer Vision, Graphics and Image Processing: Image Understanding*, 58(1):23–32, 1993.

[30] D. Terzopoulos. The computation of visible-surface representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):417–438, 1988.

[31] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - a modern synthesis. *Vision Algorithms: Theory and Practise. Triggs W., Zisserman A, and Szeliski, R. (Eds.)*, pages 298–375, 2000.

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

[33] 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.