Volumetric stereo with silhouette and feature constraints

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

Abstract

This paper presents a novel volumetric reconstruction technique that combines shape-from-silhouette with stereo photo-consistency in a global optimisation that enforces feature constraints across multiple views. Human shape reconstruction is considered where extended regions of uniform appearance, complex self-occlusions and sparse feature cues represent a challenging problem for conventional reconstruction techniques. A unified approach is introduced to first reconstruct the occluding contours and left-right consistent edge contours in a scene and then incorporate these contour constraints in a global surface optimisation using graph-cuts. The proposed technique maximises photo-consistency on the surface, while satisfying silhouette constraints to provide shape in the presence of uniform surface appearance and edge feature constraints to align key image features across views.

1 Introduction

This paper presents a volumetric approach to multiple view shape reconstruction. The target area of human reconstruction is considered. People form a central component of multimedia content and extensive research in computer vision and computer graphics has focused on modelling the shape, appearance and motion of real people [14, 6, 12]. Whole-body images of people present several important challenges for conventional reconstruction techniques with uniform surface appearance, self occlusions, sparse features, non-lambertian surfaces and limited resolution video images with wide baseline cameras often required to achieve 360° coverage.

Image-based reconstruction techniques derive a 2.5D depth representation from two or more cameras using regularised image correspondence. Kanade et al. [6] first reconstructed dynamic scenes of people using a 51 camera dome, fusing multiple 2.5D images into a single 3D surface. Image-based correspondence however fails with uniform appearance where matching is ambiguous, at depth discontinuities where the surface is tangential to the image plane, and surface fusion relies on accurate reconstruction of all surface regions recovered independently in the 2.5D representation.

Volumetric reconstruction instead derives the 3D volume that is consistent with multiple images. A volume representation allows inference of visibility and integration of appearance across multiple camera views. Shape-from-silhouette (SFS) techniques derive

the *visual-hull*, the maximal volume that is consistent with a set of foreground silhouettes. Space-carving [8] techniques provide the *photo-hull*, the maximal volume that is photo-consistent across all visible camera images. 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. The photo-hull suffers from under or over carving according to the photo-consistency criteria. Regularisation has been introduced using a level-set approach [4] although this is susceptible to local-minima according to the initial surface solution and neglects the shape constraint imposed by silhouette images.

Silhouette and photo-consistency cues have been combined in iterative local-surface optimisation techniques. The visual-hull [3, 13] has been used as a robust initial surface model for surface optimisation to match silhouette contours and maximise multiple view photo-consistency. A model-based approach [12] using a prior humanoid model has been used in human shape reconstruction. Surface optimisation is however subject to local minima, the surface is constrained to represent only those structures that are defined in the initial surface and will retain any incorrect structures, a humanoid model with short hair will not reconstruct long hair and the visual-hull will retain all phantom volumes.

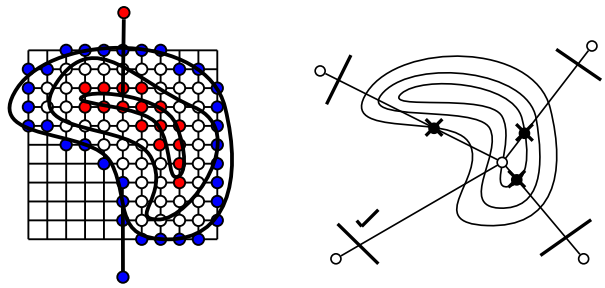
Global optimisation techniques have been presented using graph-cuts for surface reconstruction [10, 15, 11, 5]. Current approaches either neglect silhouette constraints in reconstruction [15], require genus-0 topology [11], or are restricted to a deformation with respect to the visual-hull [5] retaining any phantom structures. This paper presents a unified framework to reconstruct both occluding contour and edge contours constraints and apply the constraints in surface reconstruction, performing global optimisation with graph-cuts. The technique handles non genus-0 surfaces, provides a unified framework to reconstruct and incorporate contour constraints, and reconstructs the surface with a maximum photo-consistency that satisfies SFS without restriction to a predefined deformation with respect to the visual-hull surface.

2 Methodology

Volumetric surface reconstruction is presented using graph-cuts to first derive the occluding contours generating multiple view silhouettes and edge contours generating foreground edge features, surface reconstruction is then constrained to satisfy these contour constraints. The input for reconstruction is a set of camera images together with the camera intrinsic and extrinsic calibration parameters. Foreground segmentation is required to derive a set of foreground images for contour extraction. A volumetric reconstruction of the visual-hull is computed to define a feasible reconstruction space. The output is a triangulated surface mesh contained within the visual-hull that satisfies silhouette and feature contours with a global maximum photo-consistency score integrated across the surface.

2.1 Overview of volumetric stereo with graph-cuts

Surface reconstruction is formulated as the computation of a maximum-flow / minimum-cut solution on a discrete graph. Efficient global optimisation methods have been presented for such a problem and the theoretical relationship to a continuous minimal surface has been established [2]. Reconstruction treats surface recovery as a labelling problem. Each voxel in a discretized volume V forms a node in an undirected graph and voxels are



(a) Graph for a generalized scene (b) Visibility testing in cameras

Figure 1: Volumetric surface reconstruction. (a) The graph is constructed using depth layers from a base surface, first and last layers are connected to source and sink nodes respectively [15]. (b) Visibility is derived by testing for occluders both on the depth layer and base surface.

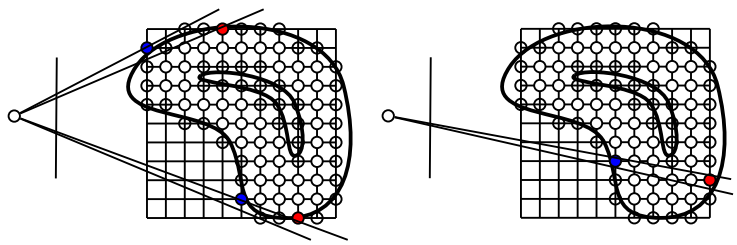
labelled as either inside or outside the true surface. Edges are defined in the graph to link adjacent voxels with a capacity corresponding to the energy function for minimization. A set of external voxels are identified and connected to a source node s and a set of internal voxels connected to a sink t. The minimum $s-t$ cut on the graph labels the remaining voxels and provides the continuous set of edges with a minimum combined energy cost that separate the source s and sink t as illustrated in Figure 1.

The energy function for optimisation is generally defined in terms of a data term E_{data} that imposes photo-consistency on the surface f and a regularization term E_{smooth} introducing spatial smoothness [7]. A discrete graph is constructed by dividing the reconstruction space into a set of depth layers $V = \{l_1, \dots\ l_D\}$ and with a node for every feasible voxel $u \in V$ [10]. The first layer l_1 corresponds to an external set of voxels and the final layer l_D an internal set, each connected to a source s and sink t node respectively with infinite edge capacity. A photo-consistency metric $\rho(u)$ is computed for each node u and edges are constructed between nodes (u,v) using capacities corresponding either to the data-term or smoothness term of the cost function. Data edges are introduced between nodes in adjacent layers (l_i, l_{i+1}) using the photo-consistency score at layer i. A cut introduced on a data-edge then corresponds to a surface passing through the node connected to the source s side of the cut. Smoothing edges are introduced between adjacent nodes within each layer l_i using an average of the photo-consistency. Edge capacities are normalised by edge length and a relative weight $0 < k < 1$ controls the degree of smoothing in reconstruction [2, 9]. A minimum $s-t$ cut on the graph provides the set of minimum capacity data edges with the maximum photo-consistency. The discrete surface solution is extracted as the set of voxels on the source side of the cut.

$$E_{data}(u,v) = 1 - \frac{\rho(u)}{\|\underline{x}(u) - \underline{x}(v)\|} \qquad u \in l_i \quad v \in l_{i+1} \tag{1}$$

$$E_{smooth}(u,v) = 1 - k\frac{\rho(u)+\rho(v)}{2\|\underline{x}(u) - \underline{x}(v)\|} \qquad u,v \in l_i \quad 0 < k < 1 \tag{2}$$

The feasible reconstruction space is defined here using the visual-hull as proposed by Vogiatzis et al. [15]. A layered representation is constructed by successively eroding



(a) Occluding contours (b) Foreground edge features

Figure 2: Graph construction for (a) occluding contours generating an image silhouette and (b) edge contours generating edge features in a foreground image. Nodes form an ordered set with proximal nodes attached to a source s and distal nodes to a sink t in the graph.

and extracting surface layers l_i from the feasible set V. A graph is constructed with the visual-hull surface l_1 connected to the source s and the interior voxels for the final eroded surface layer l_D connected to the sink t as illustrated in Figure 1(a). Throughout this work an 18-connected neighbourhood is used for the graph structure such that data-edges between layers are not restricted to the axes of the discrete volume grid.

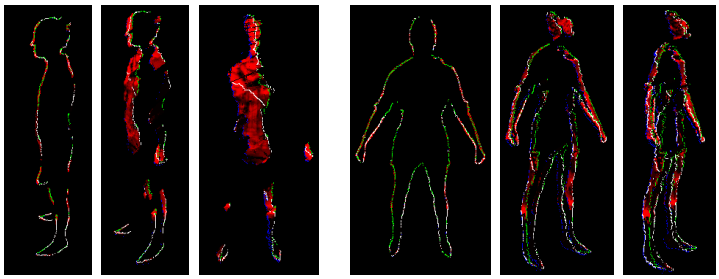
The photo-consistency metric $\rho(u)$ is computed using the Zero-mean Normalised Cross Correlation Score (ZNCC) averaged across all pairs of visible cameras and mapped to the range [0,1]. The ZNCC is invariant to linear intensity variations allowing for a non-Lambertian scene under the assumption of a dichromatic reflection model with a locally planar surface patch and no saturation in the camera images. Photo-consistency $\rho(u)$ is derived for a voxel in a depth layer $u \in l_i$ by testing for occluders against both the layer l_i and the visual-hull surface l_1 as shown in Figure 1(b).

2.2 Surface contour reconstruction

In this section contour reconstruction is formulated as a graph-cut on a subset of the feasible space V. A contour c in an image is defined by a set of image pixels $p \in c$. Each pixel reprojects to a ray in space that intersects the feasible set V as an ordered set of intersections $V_{pi} \subset V, i \in \{1 \ldots I\}$. Where only one intersection occurs $I = 1$ the true contour generator u_p is guaranteed to lie on the intersection segment between the first and last voxel in the ordered set. We have the following constraint for the contour generators $u_p \in V$ that will generate a subset of a contour c.

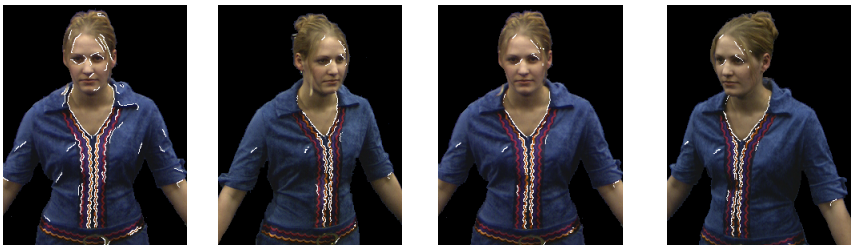
$$(u_p \in V_{pi}) \ \forall \ (p \in c) \Leftrightarrow (I = 1) \tag{3}$$

The projection for each voxel $u \in V$ is precomputed in a camera image and the feasible set for contour reconstruction is defined as all voxels that project onto all image contours $c \in C$ with only one intersection segment $I = 1$. A graph is constructed with a node for each voxel in the feasible set. The first node in each intersection V_{pi} is connected to the source s and the last to the sink t as shown in Figure 2. Data edges are defined between neighbouring voxels in each intersection segment and smoothing edges are defined between neighbouring voxels in adjacent intersection segments. The minimum $s-t$ cut on the graph derives the set of surface points u_p generating the contours in an image.



(a) Silhouette contour 1 (b) Silhouette contour 2

Figure 3: Graphs constructed on the visual-hull surface to derive the occluding contour corresponding to two silhouette images, showing three viewpoints for each graph.



(a) Edge features (b) Left stereo pair (c) Consistent edges (d) Right stereo pair

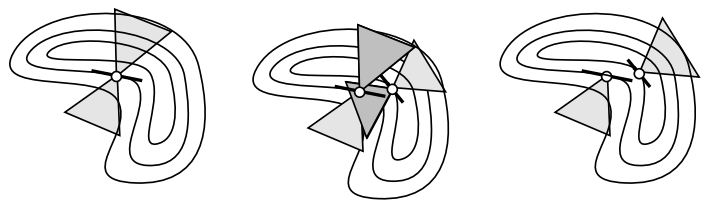
Figure 4: Extracted edge features in a selected camera are reconstructed as the set of *left-right* consistent edge contours in the stereo camera pairs.

2.2.1 Surface rim reconstruction

The silhouette of an object in an image is generated by a set of occluding contours on the surface of the object termed *surface rims*. The visual-hull of the object is reconstructed as the intersection of the visual cones generated by reprojecting multiple image silhouettes into space. The surface rims therefore have the property that they lie on the surface of the visual-hull. For each camera image the set of image contours C for the silhouette are derived and a graph is constructed for the corresponding feasible reconstruction space $V_{pi}(I=1)$. Edge capacities are calculated using photo-consistency $\rho(u)$ computed across all cameras for which each voxel is visible. The minimum $s-t$ cut on the graph derives the surface rims on the visual-hull that generate the silhouette contour. Figure 3 shows the graphs constructed for two silhouette images.

2.2.2 Edge contour reconstruction

Image edges have a dominant local intensity variation that provide a strong cue for stereo matching. Reliable matches can also be derived by enforcing *left-right* consistency between stereo camera pairs. Edges in an image are derived using a conventional edge-filter such as the Canny-Deriche edge detector. Edges orientated along epipolar lines will have ambiguous matches and can be neglected by only considering the component of the image



(a) Constraint point (b) Inconsistent regions (c) Consistent regions

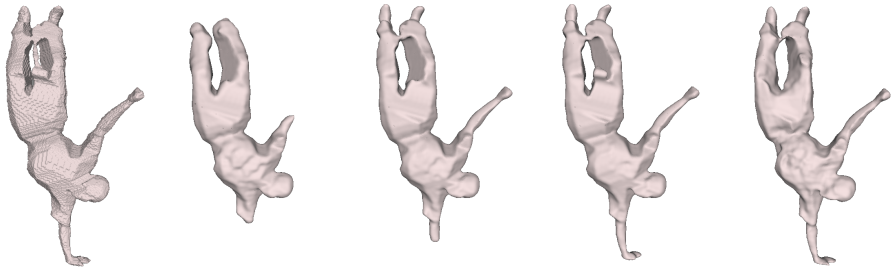
Figure 5: (a) Contour constraints generate an infeasible surface region within the reconstruction space in a directed graph. (b) Inconsistent constraints are detected and then (c) removed.

gradient directed along the baseline between cameras in the filter. For a selected camera a subset of images are considered for all cameras forming stereo pairs as shown in Figure 4. Edges are derived in the foreground images and a graph constructed for the feasible reconstruction space $V_{pi}(I = 1)$. The surface contours are then derived from the minimum $s - t$ cut on the graph as the set of surface voxels that correspond to an edge feature in a stereo pair, rejecting occluded contours that appear in one camera only and ambiguous contours where reconstruction does not match features across views. Once the contours for all images are reconstructed, the set of *left-right* consistent surface contours are identified where a consistent set of surface voxels is derived for matched edge features. Figure 4 shows the *left-right* consistent edge features reconstructed for a camera image.

2.3 Constrained surface reconstruction

In this section surface reconstruction in the feasible set V is formulated to satisfy a set of contour constraints. The assumption is made that the underlying scene surface is locally injective to the depth layer representation illustrated in Figure 1(a) such that no over-folding occurs. Surface reconstruction can then be performed using a directed graph in which data edges are defined in the direction of increasing depth within the graph with infinite capacity edges in the reverse direction. Over-folding becomes unfeasible as the corresponding cut would then include an infinite capacity edge. For each constraint point, the corresponding node is attached to the source s with infinite capacity and connected nodes on the directed data edges are connected to the sink t with infinite capacity. In a directed graph, nodes on increasing depth layers become constrained to the sink t and on decreasing depth layers to the source s such that the constraint is enforced in the minimum $s - t$ cut.

With inexact contour reconstruction and discretization on the volumetric grid it is feasible to introduce conflicting constraints as illustrated in Figure 5(b). The connectivity of the graph structure defines the region in which no-overfolding can occur. Overlapping regions corresponding to conflicting source side and sink side constraints can therefore be detected by propagating the constrained regions from each contour point. Conflicting s and t regions are detected and removed as illustrated in Figure 5(c). The minimum s-t cut on the volumetric graph provides the minimal surface that maximises photo-consistency on the surface and satisfies non-conflicting contour constraints. The underlying surface is finally extracted by a local polynomial approximation using the Moving Least Squares (MLS) framework [1].



(a) Visual-hull (b) $\lambda = 0.5$ [15] (c) $\lambda = 1.0$ [15] (d) $\lambda = 1.5$ [15] (e) Proposed

Figure 6: Volumetric stereo reconstruction in the feasible space defined by (a) the visual-hull, (b)-(d) using a ballooning term [15] and (e) using contour constraints.

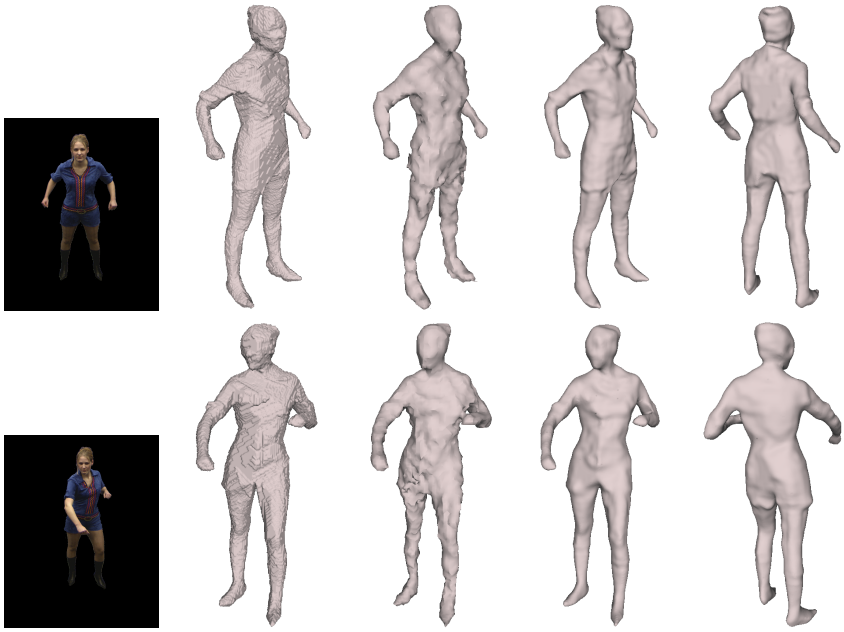
3 Results

Two sets of experiments were performed to evaluate the proposed technique, first with limited resolution whole-body images and then with wide-baseline camera positions for subjects wearing clothing with both uniform appearance and edge features. In the first test 13 camera views are used with a baseline of 30° for 360° degree coverage with standard definition (SD) video images at 0.4MPixel resolution. In the second test 8 cameras are used with a baseline of 45° for 360° degree coverage with high definition (HD) video images at 2MPixel resolution. Figures 7(a), 8(a) show the camera images for the two experiments.

The advantage of incorporating explicit contour constraints in reconstruction is demonstrated in Figure 6. Volumetric stereo [15] provides a feasible surface reconstruction incorporating appearance across multiple views allowing for limited wide-baseline camera positions. However, without explicit contour constraints a ballooning term is required to ensure that the reconstruction does not simply provide a minimum area solution. The proposed technique removes the requirement for a variable ballooning term using surface contours to constrain the reconstructed surface. It is important to note that using occluding contours where only a single intersection occurs ($I = 1$) removes the phantom volume in the visual-hull that would remain in iterative surface optimisation techniques.

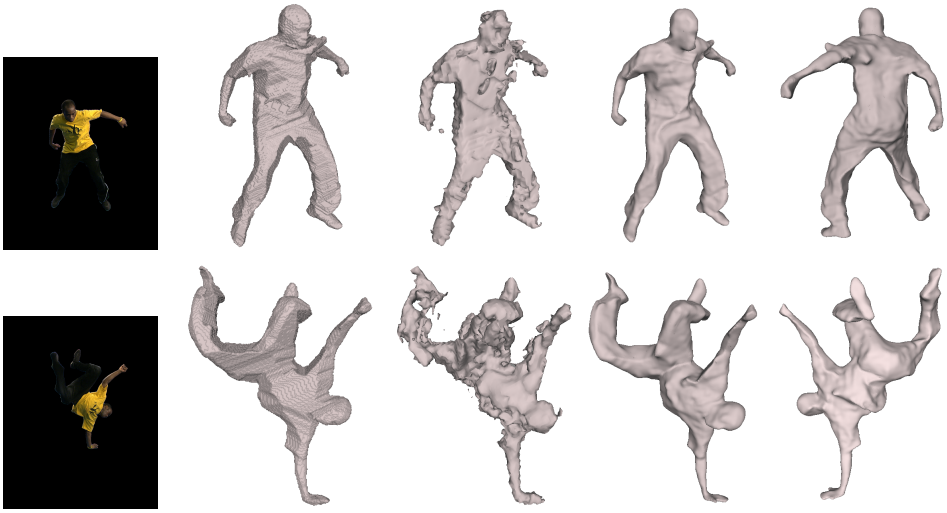
Reconstruction is compared against a baseline using the visual-hull and merged 2.5D stereo. The reconstructed surface is shown in Figures 7,8 for different body poses in the two experiments. Several points are illustrated. SFS provides only an upper-bound on the true surface and does not provide concavities that are occluded in silhouette images. Image-based stereo correspondence fails in the presence of uniform appearance. The proposed technique combines complementary cues from SFS and stereo correlation to reconstruct shape with both uniform and varying surface appearance. Shape reconstruction at the clothing boundaries using the edge feature constraints can be seen.

Reconstruction is formulated to incorporate explicit edge feature constraints to ensure that consistent intensity discontinuities in the images are matched. A limited quantitative evaluation is performed using the photo-consistency metric $\rho(u)$ accumulated across every surface point corresponding to a feature pixel in a camera image. Table 1 shows the scores for 5 different body poses in the two different experiments. As would be expected



(a) Camera (b) Visual-hull (c) 2.5D Stereo (d) Proposed (e) Proposed

Figure 7: Experiment 1: Reconstruction compared to shape-from-silhouette and multi-view stereo for a 13 camera setup with 30° camera baseline and 0.4MPixel resolution.



(a) Camera (b) Visual-hull (c) 2.5D Stereo (d) Proposed (e) Proposed

Figure 8: Experiment 2: Reconstruction compared to shape-from-silhouette and multi-view stereo for an 8 camera setup with 45° camera baseline and 2MPixel resolution.

the visual-hull provides the lowest consistency with no appearance matching, multiple-view stereo increases the photo-consistency at the edge features where the local intensity structure in the images enables matching, and the proposed technique provides the highest photo-consistency with explicit constraints on the feature matches where *left-right* consistent features are found in the images.

Experiment 1	Visual-hull	2.5D Stereo	Proposed Technique
pose1	0.70	0.77	0.81
pose2	0.75	0.79	0.82
pose3	0.74	0.80	0.82
pose4	0.71	0.78	0.81
pose5	0.66	0.76	0.79
Experiment 2	Visual-hull	2.5D Stereo	Proposed Technique
pose1	0.58	0.67	0.71
pose2	0.53	0.56	0.74
pose3	0.56	0.69	0.76
pose4	0.51	0.46	0.71
pose5	0.58	0.64	0.79

Table 1: ZNCC accumulated across all surface points corresponding to extracted image features.

Several limitations in the proposed technique should be noted. Discretisation in volumetric reconstruction should encompass the reprojection error in the camera system and a 0.5cm voxel size is used in this work. At this level discretization is of the same order as the geometric variations of the surface. Removing quantization in surface extraction using MLS then leads to over smoothing of surface detail as seen at the nose in Figures 7, 8. The global optimisation technique requires precomputation of visibility to define photo-consistency within the reconstruction space whereas iterative surface refinement approaches allow incremental estimation of visibility [4]. Finally, explicit reconstruction of occluding contours in the volume can be ambiguous in regions of uniform appearance and unique constraints may also not exist in the presence of self-occlusions. Enforcing occluding contour constraints [5] can then either place an invalid constraint or provide insufficient constraints in reconstruction.

4 Conclusion

A novel framework is presented to reconstruct contour constraints and enforce the constraints in surface reconstruction using graph-cuts for efficient global optimisation. A volumetric approach is adopted allowing integration of appearance across multiple views and inference of visibility. The proposed technique leverages shape-from-silhouette to provide shape in the presence of uniform surface appearance, feature constraints to enforce alignment of edge features between images and stereo photo-consistency to recover the surface that optimally aligns appearance between images in reconstruction. Reconstruction is applied to the problem of whole-body human shape reconstruction with limited resolution images and wide baseline camera positions. The technique demonstrates improved shape reconstruction compared to baseline implementations of shape-from-silhouette, multiple view stereo and unconstrained volumetric stereo with graph-cuts.

References

[1] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C.T. Silva. Computing and rendering point set surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 9(1):3—15, 2003.

[2] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. *IEEE International Conference on ComputerVision (ICCV)*, pages 26–33, 2003.

[3] C. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, December 2004.

[4] O. Faugeras and R. Keriven. Variational principles, surface evolution, pde's, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, 1998.

[5] Y. Furukawa and J. Ponce. Carved visual hulls for image-based modeling. *European Conference on Computer Vision*, 2006.

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

[7] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 82–96, 2002.

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

[9] C. Proulx and S. Roy. A maximum flow approach to the volumetric reconstruction problem. *Proceedings of the British Machine Vision Conference (BMVC)*, 2005.

[10] S. Roy and I. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. *IEEE International Conference on Computer Vision (ICCV)*, pages 492–502, 1998.

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

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

[13] J. Starck and A. Hilton. Virtual view synthesis of people from multiple view video sequences. *Graphical Models*, 67(6):600–620, November 2005.

[14] S. Vedula, S. Baker, and T. Kanade. Image-based spatio-temporal modeling and view interpolation of dynamic events. *ACM Transactions on Graphics*, 24(2):240–261, 2005.

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