University of Alberta

Tracking human joint motion for turntable-based static model reconstruction

by

Neil Birkbeck, Martin Jagersand, and Dana Cobzas

Technical Report TR 09-03
Jan 2009

DEPARTMENT OF COMPUTING SCIENCE
University of Alberta
Edmonton, Alberta, Canada

Abstract

We propose a method that makes standard turntable-based vision acquisition a practical method for recovering models of human geometry. A human subject typically exhibits some unintended joint motion while rotating on a turntable. Ignoring such motion causes shape-from-silhouette to excessively carve the model, resulting in loss of geometry (especially on limbs). We utilize silhouette cues and appearance consistency with an initial automatically recovered skinned-model to recover this joint motion, or wobbling. The recovered joint motion gives the calibration of each rigid body of the subject, allowing for temporal fusion of image cues (silhouettes and texture) used to refine the geometry. Our method gives improved results on real datasets when considering both silhouette overlap and texture consistency. The recovered geometry is useful in vision tasks such as multi-view image-based tracking of humans, where the recent trend of using a priori laser-scanned geometry could be replaced with a more cost effective vision-based geometry.

1 Introduction

For some time now the benefits of turntable-based vision acquisition systems for low cost 3D modeling have been recognized and exploited [6, 14]. Turntables boast the ability to quickly acquire an image stream about an object that can quickly be calibrated and easily be foreground segmented for use in both silhouette and stereo reconstruction. Light variation due to rotation of the object, useful for appearance or reflection model estimation, is easy to introduce [33]. Of course, the paradigm becomes less practical for large scale objects, but for such cases structure and motion has matured enough to be a good alternative for the calibration (e.g., the ARC 3D Web-service). In this work we argue that turntable acquisition is still feasible for human scale geometry, something that has only been exploited in few works [8, 10] and, in the case of some, it was only used for the recovery of appearance [2].

There is no doubt that convenient vision-based acquisition of static human geometry is useful, with example applications ranging from gaming to anthropometric studies. There exist full body laser range finders built exclusively for the task of recovering dense static human geometry, but this hardware comes at a premium (e.g., Cyberware ^{TM}'s Whole Body 3D Scanner \$200K+). In terms of applications in vision, a recent trend has seen many of the multi-view human tracking and deformation recovery methods being formulated around an initial laser scanned geometry [13, 12, 4]. In fact many methods in this category go on to recover deformations over time from vision, but have skipped the application of vision in the first step by relying on the scanned geometry [12].

One solution that is commonly used in capturing human geometry from vision uses a large set of fixed, pre-calibrated cameras that are observing a moving person [31, 21, 29]. Individually geometry for each time frame is reconstructed either using visual hull [30] or multi-view stereo [21] and then related to each other either using differential constrains like scene flow [31], through feature point correspondences [29], or registered with marker-based motion capture data in the coordinate system of the joint [22]. We take a different approach and propose a method that acquires human geometry using traditional turntable approach that requires only two cameras and reconstructs a model of the rotating human unified in time. Full geometry at each time frame cannot be recovered due to the low number of cameras (2) in our setup.

One limitation in simply extending the turntable-based approaches to a human scale geometry is the fact that a rotating human is not rigid and will undoubtedly move over time while rotating. Such motion causes methods like shape-from-silhouette to excessively carve the object (Fig. 1) and causes misalignment of any recovered appearance. Since the human is a kinematic chain containing a hier-

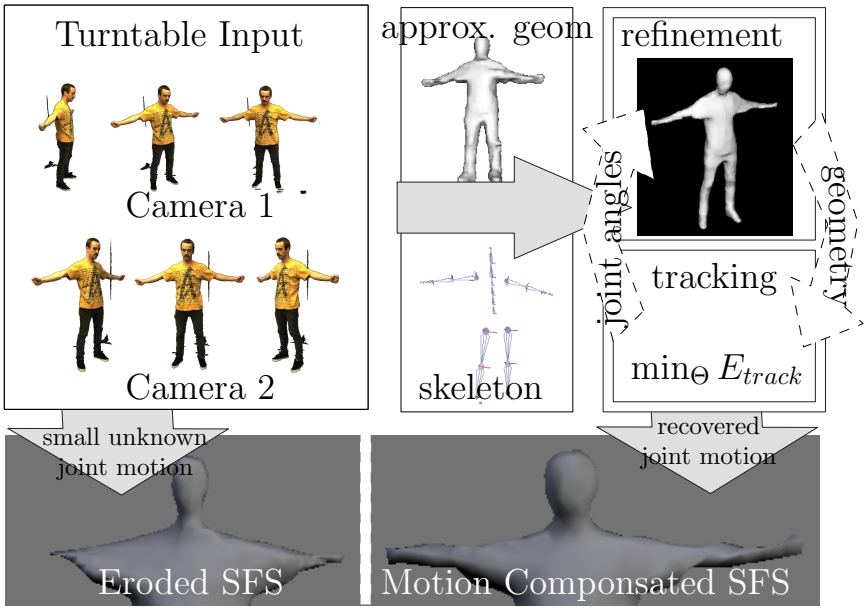

Figure 1: Overview of our solution. SFS with no motion compensation illustrates eroded body. Interleaving motion estimation with SFS gives more accurate result.

archy of coordinate systems, this problem of registration can not be solved by a simple application of single rigid body calibration. As silhouettes have always been a strong cue in turntable acquisition and human motion recovery, we propose to solve the joint motion calibration problem through an interleaved tracking and model recovery step. Our contributions are two-fold:

- using as few as two cameras, the small kinematic human motion relative to a rotating turntable is tracked by utilizing silhouette and appearance consistency while enforcing kinematic constraints.
- recovered joint angles for a kinematic structure are used to re-compute a unified shape-from-silhouette model that is the union of the visual hull for each of the kinematic links.

2 Related Work

In the context of recovering dense static geometric models of humans from vision-based methods, many of the general multi-view stereo methods for static scene reconstruction are relevant(e.g., [23]). For humans specifically, some attention

has been directed to using as few as two or three images to quickly instantiate a deformable human template model [25] . In our case, we are more concerned with convenient capture of human geometry under limited hardware assumptions; therefore, we focus on recovering the joint motions of a rotating human so as to utilize all silhouette observations in the geometry reconstruction.

Classical feature-based correspondences or feature tracks, such as those used in standard structure from motion (SFM), offer one route to recover these joint motions. Articulated structure from motion factorization techniques decompose such feature tracks into rigid parts and the corresponding points, but are often based on restricted camera models [28, 34]. On the other hand, given that feature tracks are segmented into corresponding parts, the more recent applications of SFM that refine Euclidean camera parameters based on dense matches could also be used to recover the rigid deformation of individual joints [15]. We feel that such feature-based methods may still be prone to failure in regions where few features are available, such as the arms which tend to be one of the more problematic regions.

As the geometry of these problematic regions is well classified by silhouettes, it is useful to consider the use of silhouettes for the purpose of calibration. Calibrating the relative position of cameras in a multi-view environment using dynamic silhouettes has been considered [24, 7], but in our case we assume the relative pose of cameras is known. Alternatively, similar cues such as epipolar tangents, frontier points, or silhouette consistency have also been used to calibrate the position of cameras viewing a scene under restricted turntable motion [18, 16]. Again, it is not the turntable motion given a rigid geometry we approximately recover, as we assume that the turntable motion is known; we are instead trying to recover the arbitrary, possibly small, motion of each joint relative to the turntable.

One of the most relevant methods for combining silhouettes over time utilizes both silhouette and image appearance cues. The shape-from-silhouette over time work of Cheung *et al*. [9] recovers the motion of a rigidly moving object observed by multiple image sequences by the use of frontier points and a silhouette constraint. The rigid transformation from one time frame to the next is found through a constraint that colored surface points (e.g., frontier points) are transformed onto similar image colors in the following frame while projecting inside of the silhouette. This method is used also to fuse images for recovery of human geometry under turntable motion and perform multi-view tracking [9, 10]. Unfortunately, the method relies on the colored surface points which could be hard to extract in the case of a two or three camera setup.

Some integration of silhouettes between time steps is accomplished by the spatio-temporal SFS method of Aganj *et al*., but the approach seems to be more useful for interpolating between SFS geometries at independent time steps [1]. In terms of a joint parameter estimation, the vast assortment of multi-view human

tracking methods can be though of as solving this problem [3, 17, 20, 26]. Many of these approaches also combine multiple cues, such as stereo, flow and silhouettes, for the purpose of tracking a known geometry. A practical use of the silhouette is to minimize the exclusive-or between input silhouette and model silhouette [26]; this cost function is closely related to silhouette-based camera calibration [7, 18].

Many of the multi-view tracking methods also try to refine geometries over time [19], deform temporal geometries between time-steps [29], or ensure that the silhouette of the tracked model is consistent with input silhouettes (e.g., [32]). These dynamic geometries are often reconstructed per time instant (e.g., often 6-8 or more views are available), meaning they rely mostly on the inter-camera correspondence between numerous fixed cameras for reconstructing geometry. In our case we have two widely separated views that cannot be used to reconstruct an independent geometry per frame. Instead, we exploit the intra-camera relationship for geometry reconstruction by recovering and compensating for the restricted human motion that occurs on the turntable.

3 Tracking & Refinement

We assume that the motion of the human rotating on the turntable is governed completely by the joint angles of its kinematic skeleton. The problem is then to recover both the geometry, G, and these joint angles, $\mathbf{\Theta}$, such that the geometry deformed by the joint angles is consistent with the two input image streams.

As input we have two image streams $I_{L,t}, I_{R,t}$ and silhouette images $S_{L,t}, S_{R,t}$ at time $t \in \{1, T\}$. The projection matrices $\mathbf{P}_{L,t} = [\mathbf{K}_L|0]\mathbf{E}_t$ and $\mathbf{P}_{R,t} = [\mathbf{K}_R|0]\mathbf{E}_{R,L}\mathbf{E}_t$ are also available. The relative pose between the cameras, $\mathbf{E}_{R,L}$, is fixed, and the motion of the cameras relative to the turntable is characterized only by the known transformation $\mathbf{E}_t$ (recovered using a pattern placed on the turntable - see Section 4).

Based on the observation that multi-view silhouette-driven human tracking is often successful with an approximate geometric model, we propose to solve this problem by interleaving tracking and refinement. In summary, the entire procedure involves:

1. *initialize* geometry, G, and align a kinematic structure
    - this initial geometry is based on a traditional SFS where we grow the silhouettes slightly to ensure initial model has all appendages
2. *track*

Figure 2: Capture setup illustrating the typical position of L and R cameras.

- utilize geometry G to recover joint angles, $\mathbf{\Theta}$, ensuring that motion is small, agrees with images, and keeps feet stationary

3. *refine*
   - use $\mathbf{\Theta}$ to register image observations in coordinates of each joint
   - compute SFS geometry in space of each joint
   - take union of all SFS geometries, and attach to kinematic structure
4. *iterate* tracking and refinement

The *tracking* and *refinement* components are essential and most relevant to our contribution so we discuss them first in the context of a generic model followed by details of our model. The only constraints on the generic model are that its motion can be parameterized by a set of angles and that a new model can be attached to a posed skeleton.

3.1 Tracking

Assuming some geometric model parametrized only with joint angles, e.g., $G(\boldsymbol{\theta}_t)$, we treat the recovery of all the joint angles $\mathbf{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_T\}$ as the optimization of a cost function that contains a linear combination of several terms:

$$\min_{\mathbf{\Theta}} E = E_{data} + \alpha_{kin} E_{kin} + \alpha_{smooth} E_{smooth} \tag{1}$$

The data cost, E_{data}, incorporates the agreement of the model with the image data, and is further broken down into a silhouette cost, E_{sil}, and a texture cost, E_{tex}:

$$E_{data} = E_{sil} + \alpha_{tex} E_{tex} \tag{2}$$

The silhouette energy, based on an energy used in motion tracking [27], measures agreement of the model with the input silhouettes and is computed as a sum of XOR's over all input images:

$$E_{sil} = \sum_{t=1}^{T} \sum_{i \in \{L,R\}} \sum_{\mathbf{x}} S_{t,i}(\mathbf{x}) \otimes P_{t,i}(G(\boldsymbol{\theta}_t), \mathbf{x}) \tag{3}$$

where the shorthand $P_{t,i}(G(\boldsymbol{\theta}_t))$ denotes the projected silhouette of the geometry by $P_{t,i}$. The texture energy is used to ensure that the motion recovery respects the appearance information leading to recovered joint angles that have some appearance coherence, which is useful when estimating an appearance model.

$$E_{tex} = \sum_{t=2}^{T} \sum_{i \in \{L,R\}} \sum_{\mathbf{x} \in S_{t,i}} \| I_{t,i}(\mathbf{x}) - T_{t-1,i}(\mathbf{x}, G, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) \|^2 \tag{4}$$

The texture cost is a sum of squared distance cost, computed by rendering the model in the current time step while texturing with the image and joint parameters in the previous step. Such a mapping transforms the previous input image to the current time step by warping through the model (requires the joint parameters at time t and $t-1$). With this formulation, texture coherence is only considered in the intra-camera sense. This was done as we use two cameras with with a wide baseline that have little overlap in observed regions. Furthermore, this texture energy is like a flow between images and does not need color calibration between cameras.

The smoothness term prefers no joint motion from one frame to the next:

$$E_{smooth} = \sum_{t=2}^{T} \| \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} \|^2 \tag{5}$$

Finally, due to the assumption of our input being a human rotating on a platform, the kinematic term, E_{kin}, enforces the constraint that the feet stay on the ground. This energy term measures deviations from the feet position, $\mathbf{X}_{foot}$, in frames $t > 1$ from their position at time $t = 1$

$$E_{kin} = \sum_{t=2}^{T} \sum_{foot} \| (\mathbf{X}_{foot}(\boldsymbol{\theta}_t) - \mathbf{X}_{foot}(\boldsymbol{\theta}_1)) \|^2 \tag{6}$$

Due to the discrete nature of the silhouette XOR term, we use Powell's method to optimize the cost function [27]. As the motions are small we can assume the parameters are as they were in the initial frame of the sequence and simultaneously optimize all the parameters for all frames.

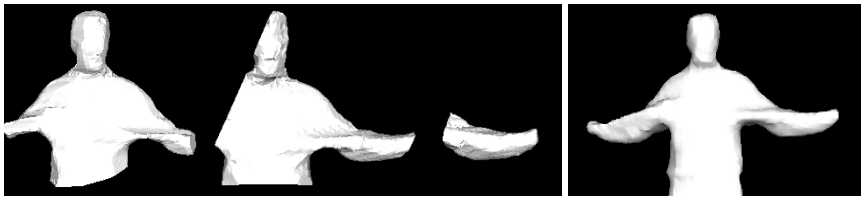

Figure 3: Pieces on the left from overlapping regions computing using SFS are merged into a single manifold geometry using the MI data structure.

3.2 Refinement

Tracking gives an updated estimate of coordinate transforms of each link at each time of the image sequence which we use to integrate all the silhouette observations. This is a straightforward process that involves interpreting each link as a rigid body and concatenating the joint to world coordinate transform with the world to camera coordinate system. With this transformation, SFS, can then be applied in a straightforward manner.

For each link we use this procedure to recover a link geometry by only considering a bounding box around each link. The bounding box is obtained from the current geometry as the bounding box of the vertices whose skinning weights to that link are above a threshold. These geometries will overlap somewhat, but this is not a limitation as each part of the geometry will lie within all the silhouettes for the sequence considered. The geometries computed for each part are originally disconnected. A manifold geometry is obtained by taking the volumetric union these disjoint geometries using the Marching Intersections (MI) data structure (see Fig. 3), where the union occurs in the pose of the first frame. Subsequently, this geometry is attached to the skeleton (see Section 3.3 for details).

3.3 Model

Our particular model consists of two parts: a mesh geometry and a kinematic structure. The geometry is used to *skin* the skeleton; the motion of it is determined solely by the kinematic model–an assumption we used during the tracking.

3.3.1 Kinematic Model

The kinematic hierarchy is represented as a tree of transformations. Each node is positioned in the coordinate system of its parent node, $P(b)$ with a Euclidean transformation $\mathbf{T}_b$ and has a set of rotational freedoms (the root also has translation),

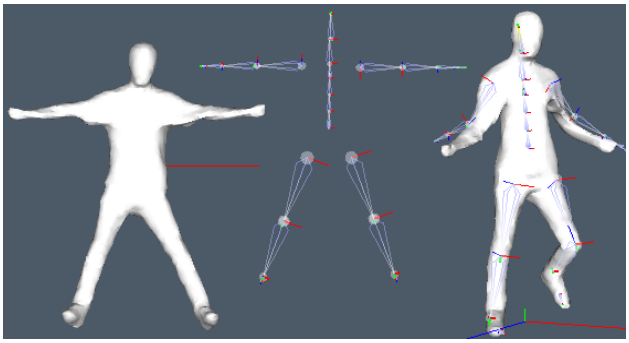

Figure 4: An illustration of an unposed or *rest* geometry (e.g., $\mathbf{v}_k(\mathbf{0})$), the corresponding unposed skeleton, and a posed geometry.

$\mathbf{R}_b(\boldsymbol{\theta}_b)$. The transformation from a joint to world coordinates is then

$$\mathbf{M}_b([\boldsymbol{\theta}_b, \boldsymbol{\theta}_{anc}]) = \mathbf{M}_{P(b)}(\boldsymbol{\theta}_{anc})\mathbf{T}_b\mathbf{R}_b(\boldsymbol{\theta}_b) \tag{7}$$

where the parent transformation is influenced by a set of ancestor joint angles, $\boldsymbol{\theta}_{anc}$. The root is an exception to this structure as it has no parent and its freedoms are a full Euclidean transformation. For notational convenience we will treat $\mathbf{M}_b$ as a function of all joint angles, $\boldsymbol{\theta}$, although freedoms of children have no affect on the parent transformation. Each joint (other than the root) is affected by at most 3 parameters.

We extract a default kinematic structure (e.g., the $\mathbf{T}_b$) complete with joint angle limits from a subject in the CMU motion capture database [11]. Redundant parameters, such as those for wrists or fingers are removed from this model before optimization (see Table 1 for a listing of the degrees of freedom and kinematic structure). We optimize the lengths of the kinematic links to align the structure to the human subject. The registration is done by locating approximate joint positions in the initial geometry (detected through assumptions on body size) and optimizing the kinematic parameters and scales such that these joint position constraints are met using inverse kinematics.

3.3.2 Kinematic & Geometry Coupling

The geometric model is attached to the skeletal model using linear blend skinning (LBS). In LBS a vertex deforms through a linear combination of a set of joints it has been associated with

$$\mathbf{v}_k(\boldsymbol{\theta}) = \sum_{b \in B(k)} w_{k,b}\mathbf{T}_b(\boldsymbol{\theta})\hat{\mathbf{v}}_k \tag{8}$$

Bone	Parent	Freedoms
Root	nil	R_x,R_y,R_z,Tx,Ty,Tz
Back	Root	$R_x \in [-20, 45]$ $R_y, R_z \in [-30, 30]$
Thorax	Back	$R_x \in [-20, 45]$ $R_y, R_z \in [-30, 30]$
Clavicle	Thorax	$R_y \in [-10, 20]$ $R_z \in [-20, 0]$
Humerus	Clavicle	$R_x \in [-60, 90]$ $R_z \in [-90, 90]$
Radius	Humerus	$R_x \in [0.01, 170]$
Femur	Root	$R_x \in [-160, 20] R_z \in [-70, 60]$
Tibia	Femur	$R_x \in [0.01, 170]$
Foot	Tibia	–

Table 1: A breakdown of the bone names, their freedoms, and their parents for a total of 34 freedoms.

where $\hat{\mathbf{v}}_k$ is the vertex in rest position, $B(k)$ is the set of links to which vertex k is attached and $w_{k,b}$ is the weight of association of vertex k with bone b. The transformation matrix $\mathbf{T}_b(\boldsymbol{\theta}) = \mathbf{M}_b(\boldsymbol{\theta})\hat{\mathbf{M}}_b^{-1}$, where $\hat{\mathbf{M}}_b = \mathbf{M}_b(\mathbf{0})$ is the rest transformation matrix for bone b and $\mathbf{M}(\boldsymbol{\theta})$ is the animated pose of bone b. Given a posed kinematic skeleton (e.g., as a result of tracking or manual initialization in the first frame) we extract the vertex skinning weights automatically using the heat diffusion process of Baran and Popovic [5].

In our case the geometry is computed in context of a posed kinematic structure, e.g., the vertices $\mathbf{v}_k(\boldsymbol{\theta}_{pose})$) are already deformed with joint parameters $\boldsymbol{\theta}_{pose}$) The heat weights are assigned to the geometry in this posed frame, so the rest geometry must be obtained through the inverse of the transformation in Eq. 8, i.e., $(\sum_{b \in B(k)} w_{k,b}\mathbf{T}_b(\boldsymbol{\theta}_{pose}))^{-1}$.

For the purpose of evaluating the model we also generate a single texture map. The texture coordinates of the models are automatically determined by identifying key points on the feet, crotch, armpits, hands and head of the model, computing a vertex-edge-path through these vertices and fixing these key coordinates; the remaining coordinates are found using a conformal mapping. These salient vertices map to predefined locations in the texture map, giving a semi-consistent parametrization of the different meshes (Fig. 5). .

4 Experiments

For the experiments we have captured three data-sets of human subjects rotating on a turntable (Fig. 6). All of the data-sets contain three video streams; two of the streams were used for reconstruction and the third was used for comparison. The video sequences in the *Yellow Shirt* and the *Red Sweater* datasets each contain 30

Figure 5: Average texture illustrating texture space.

Figure 6: Sample input images for two of the views for the *Yellow Shirt*, *Red Sweater*, and *Green Sweater* data-sets.

images, and the *Green Sweater* dataset sequences contain 22 images. All of the images are 800x600 color images captured from Point Grey grasshopper cameras. The external positions of the cameras were calibrated in advance and kept fixed. A calibration pattern positioned on the turntable was used to calibrate the relative position of the turntable with respect to these cameras over the image streams.

In each case we bootstrapped our algorithm with a geometry that was obtained from all of the images in the data-sets using SFS; the silhouette boundaries were extended (by roughly 5-6 pixels) to ensure that the extremities were present in the initial geometry. Figure 8 illustrates the final geometry (with no weight on the texture term), and the SFS geometry that would result if no motion compensation were used. In all cases we can see that the original SFS is eroded, with parts of the arms missing and the bodies shaved too far in general. The motion compensation successfully recovers these parts of the geometry. Figure 7 illustrates the silhouette agreement after alignment for the *Red Sweater* data-set, where even SFS geometry

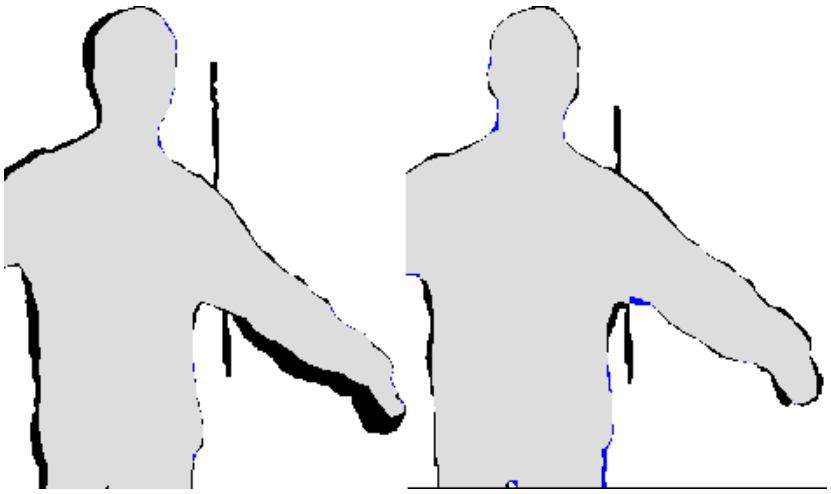

Figure 7: The *Red Sweater* data-set had the least motion, but even still the SFS geometry (left) disagrees with the input silhouette (black indicates regions of input not covered by geometry). Motion recovered silhouette matches better (right).

looks reasonable.

For numerical comparison we compare the silhouette energy for the model (averaged over the frames), and the texture energy for the model, E_{tex}. The numerical results corresponding to the *Yellow Shirt* and *Green Sweater* data-sets illustrate that the refinement with iterative SFS does reduce the XOR score. This is illustrated in Table 2 where several quality metrics are shown for the *SFS* geometry, refinement with no texture weight (i.e., $\alpha_{tex} = 0$) *Ref*, and refinement with texture weight in latter parts of optimization *+Tex*. From this data we make the following observations

- the XOR score goes down with refinement (even when using a texture term, Fig. 9)
- and by using the texture term we get more consistent texture scores (in the refinement).

We now consider the XOR results on the video stream that was not used during the reconstruction. the XOR scores also went down in that image stream when compensating for motion. For the *Red Sweater* data set the numbers through the refinement were 20185, 17010, and 16410, illustrating that the refinement is actually moving toward the true visual hull. This is in comparison to the score of 21735 obtained on the standard SFS model from the other two sequences with no

	Yellow			Green		
Stage	SFS	Ref	+Tex	SFS	Ref	+Tex
Xor	11651	8792	8739	13521	9884	9876
Tex	883	923	788	1437	1468	1294

	Red		
Stage	SFS	Ref	+Tex
Xor	11418	8580	8891
Tex	604	1426	635

Table 2: Consistency measures for data-sets with no motion recovery (SFS), the refined geometry with no texture weight (Ref) and with texture weight (+Tex)

motion compensation. Similar observations were made on the *Yellow Shirt* data-set where values were 8846, 7253, 6230 (SFS was 8325), and the *Green Sweater* data set where refinement scores decreased from 74027, 71362, to 71069 (SFS score was 73306). The overall higher scores in this image stream are due to poor background segmentation resulting in background pixels being labeled as foreground and raising the score.

Refinement of the geometry only affects the E_{sil} and E_{tex} terms because it does not change joint angles. One may question the validity of using only SFS in the refinement, as SFS does not directly minimize either of these terms, meaning that on successive tracking/refinement iterations the energy could in fact go up. Although this is possible, in practice we have found that the E_{sil} term often does go down. For example, on the *Red Sweater* data-set the score for the initial geometry was 45968, which reduced to 10496, 8855, and 8580 after interleaved geometry estimates (SFS score was 11418). Similar observations were made for the other data-sets.

5 Conclusion

We have presented an iterative method that uses as few as two camera streams to recover small human motion primarily by using silhouette cues. The recovered motion allows the registration of silhouettes to improve the geometry using SFS.

One limitation of our model is that we do not optimize the texture cost in the mesh refinement. A more accurate geometry could be obtained using stereo consistency in the intra-camera sequence, but we argue that the registration we recover is a necessary pre-requisite for this stage. Another limitation is that our method needs to be bootstrapped with an initial geometry. We currently based this geometry on an enveloping geometry that is obtained by growing the silhouette boundaries. We would like to further explore the sensitivity of our solution to this initial position.

In future work we would like to explore using this model in the context of tracking. Another possible future direction is to see if refining the model in this manner can be done in an online manner with general motion.

References

[1] E. Aganj, J.-P. Pons, F. Segonne, and R. Keriven. Spatio-temporal shape from silhouette using four-dimensional delaunay meshing. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[2] N. Ahmed, H. Lensch, and H.-P. Seidel. Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):663–674, 2007. Member-Christian Theobalt and Member-Marcus Magnor.

[3] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 349–356, Washington, DC, USA, 2005. IEEE Computer Society.

[4] L. Ballan and G. M. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *3DPVT*, Atlanta, GA, USA, June 2008.

[5] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. *ACM Trans. Graph.*, 26(3):72, 2007.

[6] A. Baumberg, A. Lyons, and R. Taylor. 3D S.O.M. - a commercial software solution to 3d scanning. In *Proceedings of Vision, Video, and Graphics (VVG'03)*, pages 41–48, July 2003.

[7] E. Boyer. On using silhouettes for camera calibration. In *ACCV*, 2006.

[8] K. M. Cheung, S. Baker, J. K. Hodgins, and T. Kanade. Markerless human motion transfer. In *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission*, September 2004.

[9] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *International Journal of Computer Vision*, 62(3):221 – 247, May 2005.

[10] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking. *International Journal of Computer Vision*, 63(3):225 – 245, August 2005.

[11] CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[12] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, pages 1–10, New York, NY, USA, 2008. ACM.

[13] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less 3d feature tracking for mesh-based human motion capture. In A. M. Elgammal, B. Rosenhahn, and R. Klette, editors, *Workshop on Human Motion*, volume 4814 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2007.

[14] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. In *4th International Conference on 3D Digital Imaging and Modeling (3DIM'03)*, pages 46–53, October 2003.

[15] Y. Furukawa and J. Ponce. Accurate camera calibration from multi-view stereo and bundle adjustment. pages 1–8, 2008.

[16] Y. Furukawa, A. Sethi, J. Ponce, and D. Kriegman. Structure and motion from images of smooth textureless objects. In *ECCV*, 2004.

[17] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *CVPR '96*.

[18] C. Hernandez, F. Schmitt, and R. Cipolla. Silhouette coherence for camera calibration under circular motion. 29(2):343–349, February 2007.

[19] A. Hilton and J. Starck. Multiple view reconstruction of people. In *3DPVT '04: Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium on (3DPVT'04)*, pages 357–364, Washington, DC, USA, 2004. IEEE Computer Society.

[20] R. Kehl, M. Bray, and L. V. Gool. Full body tracking from multiple views using stochastic sampling. In *CVPR '05*.

[21] J.-P. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *CVPR '05*.

[22] P. Sand, L. McMillan, and J. Popović. Continuous capture of skin deformation. *ACM Trans. Graph.*, 22(3):578–586, 2003.

[23] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR '06*.

[24] S. Sinha and M. Pollefeys. Camera network calibration from dynamic silhouettes. In *CVPR*, 2004.

[25] J. Starck, A. Hilton, and J. Illingworth. Human shape estimation in a multi-camera studio. *British Machine Vision Conference (BMVC)*, pages 573–582, 2001.

[26] C. Theobalt, J. Carranza, M. A. Magnor, and H.-P. Seidel. Combining 3d flow fields with silhouette-based human motion capture for immersive video. *Graphical Models*, 66(6):333–351, 2004.

[27] C. Theobalt, E. de Aguiar, M. Magnor, and H.-P. Seidel. *Reconstructing Human Shape, Motion and Appearance from Multi-view Video*, chapter Reconstructing Human Shape, Motion and Appearance from Multi-view Video, pages 29–58. Springer, Heidelberg, Germany, November 2007.

[28] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *cvpr*, 2005.

[29] K. Varanasi, A. Zaharescu, E. Boyer, and R. P. Horaud. Temporal surface tracking using mesh evolution. In *Proceedings of the Tenth European Conference on Computer Vision*, volume Part II of *LNCS*, pages 30–43, Marseille, France, October 2008. Springer-Verlag.

[30] 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, April 2005.

[31] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):475 – 480, March 2005.

[32] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3):1–9, 2008.

[33] M. Weber, A. Blake, and R. Cipolla. Towards a complete dense geometric and photometric reconstruction under varying pose and illumination. In *BMVC*, 2002.

[34] J. Yan and M. Pollefeys. Articulated motion segmentation using ransac with priors. In *ICCV Workshop on Dynamical Vision*, 2005.

Figure 8: Reconstruction without motion compensation (e.g., SFS) on the left, followed by the refined model (middle), and a textured model (single average texture).

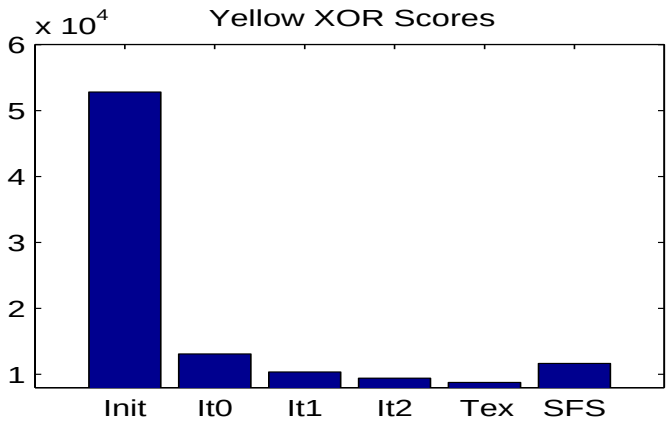

Figure 9: XOR decrease through refinement compared to SFS