Reconstructing Non-stationary Articulated Objects in Monocular Video using Silhouette Information

Saad M. Khan and Mubarak Shah
University of Central Florida, Orlando, FL, USA.

Abstract

This paper presents an approach to reconstruct non-stationary, articulated objects from silhouettes obtained with a monocular video sequence. We introduce the concept of motion blurred scene occupancies, a direct analogy of motion blurred images but in a 3D object scene occupancy space resulting from the motion/deformation of the object. Our approach starts with an image based fusion step that combines color and silhouette information from multiple views. To this end we propose to use a novel construct: the temporal occupancy point (TOP), which is the estimated 3D scene location of a silhouette pixel and contains information about duration of time it is occupied. Instead of explicitly computing the TOP in 3D space we directly obtain it's imaged(projected) locations in each view. This enables us to handle monocular video and arbitrary camera motion in scenarios where complete camera calibration information may not be available. The result is a set of blurred scene occupancy images in the corresponding views, where the values at each pixel correspond to the fraction of total time duration that the pixel observed an occupied scene location. We then use a motion de-blurring approach to de-blur the occupancy images. The de-blurred occupancy images correspond to a silhouettes of the mean/motion compensated object shape and are used to obtain a visual hull reconstruction of the object. We show promising results on challenging monocular datasets of deforming objects where traditional visual hull intersection approaches fail to reconstruct the object correctly.

1. Introduction

In this paper we present a novel approach to reconstruct the 3D shape of an object from silhouettes obtained in a monocular video sequence with the object undergoing rigid or non-rigid motion. Traditionally visual hull based approaches rely on object silhouettes obtained from multiple time-synchronized cameras or if a single camera is used for a fly-by (or a turn table setup) the scene is assumed to be static. These constraints greatly limit the applicability of visual hull based approaches to controlled laboratory conditions. In real-life applications, a sophisticated multiple-camera setup may not be available. If a single camera is used to capture multiple views by going around the object, it is not reasonable to assume that the object will remain static over the course of time it takes to obtain views of the object, especially if it is a person, animal or vehicle on the move.

Though in the past there has been some work on using visual hull reconstruction in monocular video sequences of rigidly moving objects to recover shape and motion [13][9][2] [8], these methods involve the estimation of 6 DOF rigid motion of the object between successive frames. To handle non-rigid motion the use of multiple cameras becomes indispensable [3]. Unlike these approaches we do not require the detection of surface feature points in 3D (frontier points, colored surface points) for the estimation and eventual compensation of the motion of the scene object. Rather we introduce the concept of motion blurred scene occupancies, a direct analogy of the motion blurred image but in a 3D object scene occupancy space. Similar to a motion blurred picture caused by the movement of a scene object (or the camera) and the camera sensor accumulating scene information over the exposure time, 3D scene occupancies will be mixed with non-occupancies where there is motion resulting in a *motion blurred occupancy space*. By de-blurring this data with appropriate point spread functions (PSF), we are able to obtain the motion compensated 3D shape of the object. Note that our approach is different from the traditional structure from defocus/deblur approaches [6] [10]. There the objective is to obtain a depth map/surface of the scene from one or more blurred (out of focus) appearance images by adjusting camera focal length, or recovering motion of the object that caused the motion blurred appearance.

Our approach takes a different route to recover structure from multiple views obtained from a monocular video sequence of a non-stationary object. Instead of using motion blurred appearance image/s we fuse silhouette information from multiple views to create motion blurred scene occu-

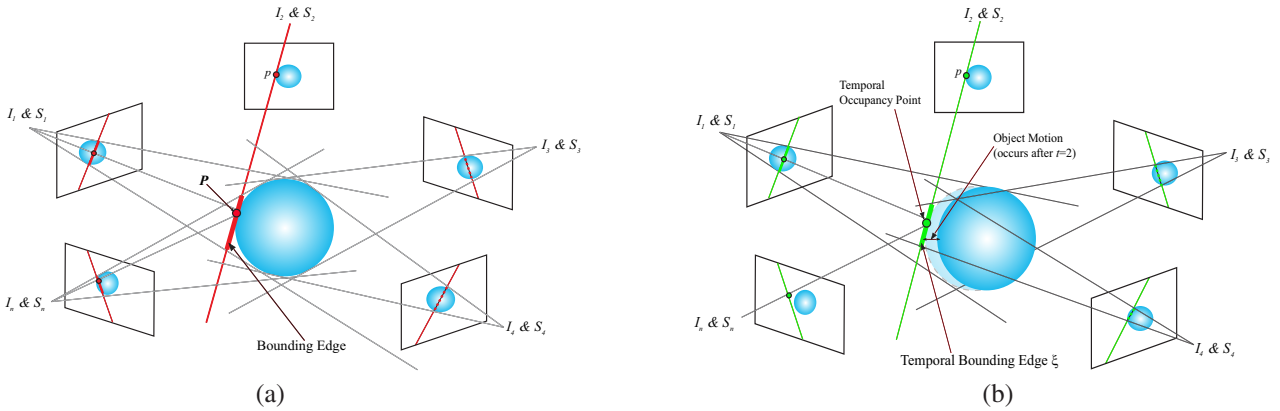


Figure 1. In case of a stationary object we can obtain the bounding edge for a pixel on the foreground silhouette by extending a ray through the pixel and selecting the section of the ray that projects to within the bounds of silhouettes in all views. This process is shown in (a) where the bounding edge corresponding to pixel p in view I_2 is highlighted with a bold red segment of the red ray. When the object is undergoing motion the ray through a silhouette pixel is not guaranteed to project to within the bounds of silhouettes of other views. In this case for pixel p we have a temporal bounding edge which is the section of the ray through p that projects to the highest number of silhouettes as shown in (b). The temporal occupancy point corresponding to p is also shown. This is the point on the temporal bounding edge that when projected in the visible images has minimum color variance and is good estimate of the 3D scene point that is imaged at p.

pancy information, where greater blur (lesser occupancy value) is interpreted as greater mixing of occupancy with non-occupancy in the total time duration. We then use a motion deblurring approach to obtain the mean/motion compensated 3D shape of the scene object over the duration of time.

The rest of the paper is organized as follows. In section 2 we discuss related work. In section 3 we describe details of our approach. In section 4 we report results from our experiments on a variety of challenging scenarios. We conclude this paper in section 5.

2. Related Work

Visual hull methods [1] can yield surprisingly accurate shape models. Over the years a number of variants have evolved including surface representation [14], voxel representation [16], or image-based representation [11], elegantly summarized and evaluated in a recent survey of the literature [12]. A limiting constraint on these approaches is the requirement of multiple time synchronized cameras, or if only a monocular video sequence is available then the scene object must remain static (in effect simulating multiple views obtained at the same time). If both the object has motion and we have a monocular video sequence then this assumption is violated and the silhouettes obtained from the different views no longer carve out the object's visual hull.

In the relatively recent past approaches that combine visual hull and stereo reconstruction [4] have been proposed to handle rigid motion of the object in monocular video. In Wong et al. [13] the motion was assumed to be circular. Frontier points were extracted from the silhouette boundary and used to estimate the axis of rotation. In [9] a local parabolic structure was defined on the surface of a smooth curved object and epipolar geometry was used to localize frontier points. In [2] the 6 DOF motion is estimated by combining both color and silhouette information (CSPs). A parallel work by Cheung et al. [3] can handle non-rigid motion of the object but requires the use of multiple cameras.

Our approach takes a different route to recover structure from a monocular video sequence of a non-stationary object. We hypothesis that due to motion space occupancies are mixed or blurred with non-occupancies over time. Given a occupancy grid with occupancies accumulated over time we propose to use a motion deblurring approach to obtain the mean 3D shape of the scene object.

3. Approach

Silhouette information has been used in the past to estimate occupancy grids for the purpose of detection and reconstruction. Due to the inherent nature of visual hull based approaches if the silhouettes correspond to a non-stationary object obtained at different time steps (monocular video), grid locations that are not occupied consistently will be carved out. As a result the reconstructed object will only have an internal body core (consistently occupied scene locations) survive the visual hull intersection. Our first task is therefore to identify occupancy grid locations that are occupied by the scene object and for the durations that they are occupied. In essence scene locations giving rise to the silhouettes in each view need to be estimated.

3.1. Obtaining Scene Occupancies

Let $\{ I_t, S_t \}$ be the set of color and corresponding foreground silhouette information generated by a stationary ob-

ject $\mathbf{O}$ in T views obtained at times $t = 1..., T$ in a monocular video sequence (e.g. the camera flying around the object). Let p_i^j be a pixel in the foreground silhouette image S_i. With the camera center of view i, p_i^j defines a ray r_i^j in 3D space. If the scene object is stationary, then a portion of r_i^j is guaranteed to project inside the bounds of the silhouettes in all the views, and in past literature it has been referred to as the *bounding edge* [2], see figure 1(a). Assuming the object to be Lambertian and the views to be color balanced, the 3D scene point P_i^j corresponding to p_i^j can be estimated by searching along the bounding edge for the point with minimum color variance when projected to the $visible$ color images.

Now, if object $\mathbf{O}$ is non-stationary and P_i^j is not consistently occupied over the time period $t = 1 : T$ then r_i^j is no longer guaranteed to have a bounding edge. There may be no point on r_i^j that projects to within object silhouettes in every view, in fact there may be views where r_i^j projects completely outside the bounds of the silhouettes as shown in figure 1(b). Since the views are obtained sequentially in time, the number of views in which r_i^j projects to within silhouette boundaries would in turn put an upper bound on the amount of time (w.r.t. to total duration of video) P_i^j is guaranteed to be occupied by $\mathbf{O}$. Let us define as *temporal occupancy* τ_j^i, the fraction of total time instances T (views) where r_i^j projects to within silhouette boundaries and temporal bounding edge ξ_i^j as the section of r_i^j that this corresponds to as shown in figure 1(b). We can formally state aforementioned ideas in the following proposition:

Proposition For a silhouette point p_i that is the image of scene point P_i, τ_i provides an upper bound on the duration of time it is guaranteed to be occupied and determines the temporal bounding edge ξ_i on which P_i must lie.

In the availability of scene calibration information, ξ_i^j and τ_i^j can be obtained by successively projecting r_i^j in the image planes and retaining the section that projects to within the maximum number of silhouette images. To refine our localization of the 3D scene point P_i^j (corresponding to the silhouette pixel p_i^j) along ξ_i^j, we develop another construct called the *temporal occupancy point* obtained by enforcing the appearance/color constancy constraint as described in the next section.

3.1.1 Temporal Occupancy Points

If the views of the object are captured at a rate faster than it's motion, then without loss of generality a non-stationary object $\mathbf{O}$ can be considered piece-wise stationary: $\mathbf{O} = \{\, \mathbf{O}_{1:s_1}, \mathbf{O}_{s_1+1:s_2}, ..., \mathbf{O}_{s_k:T} \}$, where each s_i marks a time where there is motion in the object. This assumption is easily satisfied in high capture rate videos where for small batches

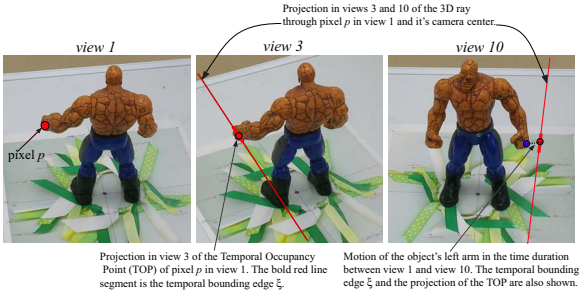


Figure 2. Three frames from a monocular sequence of a non-rigidly deforming object (motion in the left arm after view 3). For the pixel marked with a red circle in view 1 the projection of it's temporal bounding edges and TOPs are shown in views 3 and 10.

of frames non-stationary objects tend to be rigid. With the previous assumptions of Lambertian surfaces and color balanced views, having piece-wise stationarity would justify a photo-consistency check along the temporal bounding edge for scene point localization. We can then proceed with a linear search along the temporal bounding edge ξ_i^j for a point that touched the surface of the object. Such a point will have the property that it's projection in the $visible$ images has minimum color variance (color constancy constraint). We refer to this point as the *Temporal Occupancy Point (TOP)* as shown in figure 1(b), and use it as the estimated localization of the 3D scene point P_i^j that gave rise to the silhouette pixel p_i^j.

In figure 2 we demonstrate this process on some real data used in our experiments. The figure shows three views from a monocular camera sequence (flyby) as the object moves it's left arm. Pixel p marked with a red circle corresponding to the left hand in view 1 is selected for demonstration. The 3D ray back projected through this pixel is imaged in views 3 and 10, shown by the red lines. Notice that due to the motion of the object (left arm moving down) in the time duration between views 1 and 10, the ray does not pass through the corresponding left hand pixel in view 10 (marked with a blue circle). In fact the projection of the ray is completely outside the bounds of the object silhouette in view 10. The temporal bounding edges and the TOPs corresponding to pixel p are computed and their projections in view 3 and 10 are also shown.

Since we are using monocular video sequences, it may not be the case that we have complete camera calibration at each time instant, particularly if the camera motion is arbitrary. Our strategy is therefore to use a purely image-based approach. For each silhouette pixel, instead of determining it's corresponding TOP explicitly in 3D space, we directly obtain the projections (images) of the TOP in each view. If the object was stationary and the scene point visible in every view, then a simple stereo based search algorithm could be used. Given the fundamental matrices between

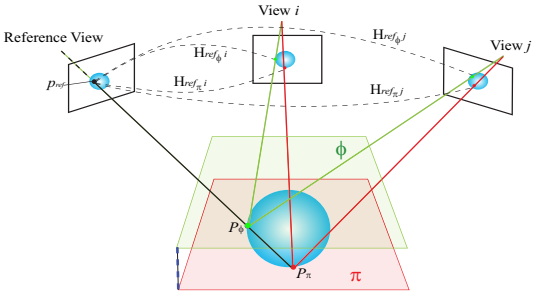


Figure 3. In the absence of complete camera calibration, 3D scene points on a ray passing through a pixel can be directly imaged in other views by warping the pixel with homographies induced between views by a pencil of parallel planes intersecting the ray.

views, the ray through a pixel in one view can be directly imaged in other views using the epipolar constraint [17]. The images of the TOP can then be obtained by searching along the epipolar lines (in the object silhouette regions) for a correspondence across views that has minimum color variance. Since neither the object is stationary nor the scene point guaranteed to be visible from every view, a stereo based approach as described above is not viable. As can be seen in figure 2, pixel p in view 1 has an epipolar line in view 10 (image of the projected ray through p) that is outside the bounds of the object silhouette making it impossible to search along the epipolar line for a correct correspondence. We, therefore, propose to use homographies induced between views by a pencil of planes for a point to point transformation.

Consider figure 3. The image of the 3D scene point P_ϕ (corresponding to the image point p_{ref} in the reference view) can be directly obtained in other views by warping p_{ref} with the homography induced by a plane ϕ that passes through P_ϕ as shown in figure 3. To obtain this homography, we can use a ground plane reference system. Given the homography induced by a ground scene plane and the vanishing point of the normal direction, homographies of planes parallel to the ground plane in the normal direction can be obtained using the following relationship [7]:

$$H_{i_\phi j} = (H_{i_\pi j} + [0|\gamma \mathbf{v}_{ref}])(I_{3\text{x}3} - \frac{1}{1+\gamma}[0|\gamma \mathbf{v}_{ref}]). \quad (1)$$

The parameter γ determines how far up from the reference (ground in our case) plane, the new plane is. The projection of the temporal bounding edge ξ_i^j in the image planes can be obtained by warping p_i^j with homographies of successively higher planes (by incrementing the value of γ) and selecting the range of γ for which p_i^j warps to within the largest number of silhouette images. The image of p_i^j's TOP in all the other views is then obtained by finding the value of γ in the previously determined range, for which p_i^j and its homographically warped locations have minimum color variance in the visible images. The upper bound on occupancy duration τ_i^j is evaluated as the ratio of the number of views where ξ_i^j projects to within silhouette boundaries and the total number of views. This value is stored at the imaged locations of p_i^j's TOP in all other views.

3.1.2 Building Blurred Occupancy Images

As described above, for a silhouette pixel we can obtain the image location of its TOP in every other view. We uniformly sample the boundary of the object silhouette in each view and project their TOPs in all the views. The accumulation of these projected TOPs delivers a corresponding set of images that we call the blurred occupancy images: $B_t; t = 1, ..., T$. The pixel values in each image are the occupancy durations τ of the TOPs. Examples are shown in figure 4 and the analogy with motion blurred images is apparent. Due to the motion of the object, regions in space are not consistently occupied resulting in some occupancies blurred out with non-occupancies which is reflected in the blurred occupancy images. The algorithmic procedure is described in the following steps:

Generate blurred occupancy images $B_t; t = 1, ..., T$.

- **for** each silhouette image
  - Uniformly sample silhouette boundary
  - **for** each sampled silhouette pixel p
    1. Obtain temporal bounding edge ξ and occupancy duration τ
       - As described in 3.1.1 transform p to other views using multiple plane homographies.
       - Select range of γ (planes) for which p warps to within the silhouette boundaries of the largest number of views.
    2. Find projected location of TOP in all other views
       - Search along ξ (values of plane γ)
       - Project point to $visible$ views
       - Return if minimum variance in appearance amongst the views.
    3. Store value τ at projected locations of TOP in each B_t.
  - End **for**.
- End **for**.

3.2. Motion Deblurring

The motion blur in the blurred occupancy images can be modelled as the convolution of a blur kernel with the latent

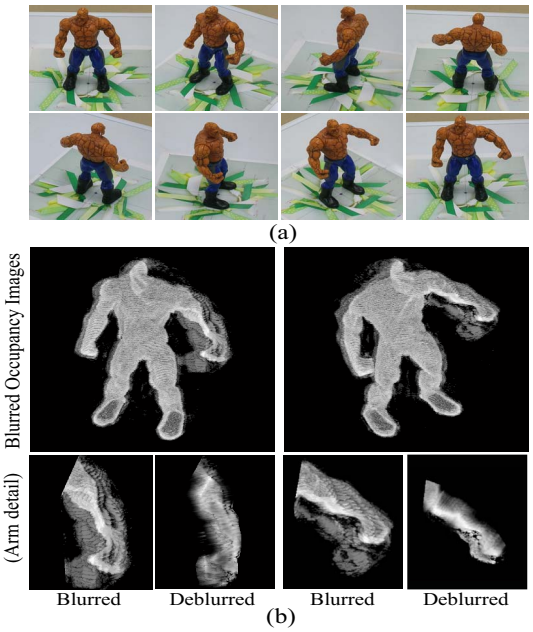


Figure 4. (a) Eight of the 20 views from a monocular dataset. Notice the left arm of the model is moving (compare first and last images). (b) Two of the 20 *blurred occupancy images*. Due to the motion of the arm some sections of the scene (where the moving arm passes through) are not consistently occupied resulting in the blurred silhouette/occupancy images.

occupancy image plus noise:

$$B = L \otimes K + n, \tag{2}$$

where B is the blurred occupancy image, L is the latent or unblurred occupancy image, K is the blur kernel also known as the point spread function (PSF) and n is additive noise. Conventional blind deconvolution approaches focus on the estimate of K to deconvolve B using image intensities or gradients. In traditional images, there is the additional complexity that may be induced by the background that may not undergo the same motion as the object. The PSF has a uniform definition only on the moving object. This however is not a factor in our case since the information in blurred occupancy images corresponds only to the motion of the object. Here we do not propose a new method to compute the blur PSF since that is not the focus of this work and there have been several successful blind deconvolution algorithms developed in the recent past [19][18][20]. We use an approach similar to the recent work by Jia [21]. They first segment the foreground object as a blurred transparency layer and use the transparency information in a MAP framework to obtain the blur kernel. By avoiding taking all pixel colors and complex image structures into computation their approach has the advantage of simplicity and robustness but requires the estimation of the object transparency or alpha matte. The object occupancy information in our blurred occupancy maps once normalized in the [0-1] range and can be directly interpreted as the transparency information or an alpha matte of the foreground object.

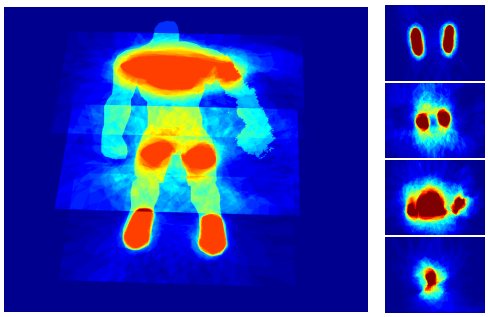
The blur filter estimation maximizes the likelihood that the resulting image, when convolved with the resulting PSF, is an instance of the blurred image, assuming Poisson noise statistics. The process de-blurs the image and refines the PSF simultaneously, using an iterative process similar to the accelerated, damped Lucy-Richardson algorithm. We start with an initial guess of the PSF as simple translational motion. This is fed into the blind deconvolution approach that iteratively restores the blurred image and refines the PSF to deliver the de-blurred occupancy maps $L_t; t = 1, ..., T$, which are used in the final reconstruction.

It should be noted that our deblurring approach assumes uniform motion blur but that may not be the case in natural scenes. For instance due to the difference in motion between the arms and the legs or a walking person the blur patterns in occupancies may be different and hence different blur kernels will need to be estimate for each section. This is a very challenging problem and though there is some very recent work on estimating blur kernels in the case of non-uniform blurring [22], this problem is beyond the scope of this work and will be addressed in future studies. In our method for PSF estimation the user specifies different crop regions of the blurred occupancy images each with uniform motion, which are then restored separately.

3.3. Final Reconstruction

Once motion deblurred occupancy images have been generated, the final step is to perform a probabilistic visual hull intersection. Conventional approaches can be used [15] and for our purposes the recent image-based approach of Khan et al. [7] is suitable as it handles arbitrary camera motion without requiring full calibration. We here provide a brief description of this approach, interested readers are directed to [7] for details.

The 3D structure of objects is modelled as being composed of an infinite number of cross-sectional slices, with the frequency of slice sampling being a variable determining the granularity of the reconstruction. Using planar homographies induced between views by a reference plane in the scene (ground) occupancy maps L_is' (foreground silhouette information) from all the available views are fused into a reference view (arbitrarily chosen) performing visual hull intersection in the image plane. This process delivers a 2D grid of object occupancy likelihoods representing a cross-sectional slice of the object. Consider a reference plane π in the scene inducing homographies $H_{i_\pi j}$, from view i to view j. Warping L_is to a occupancy map in a reference view L_{ref} , we have the warped occupancy maps: $\hat{I}_i = [H_{i_\pi j} L_i]$. Visual hull intersection on π is achieved by



(a) (b)

Figure 5. After deblurring these are used to perform a slice based reconstruction of the object (b) Three of the 100 slices are overlaid onto a reference view (deblurred occupancy map). (c) The slices are shown separately.

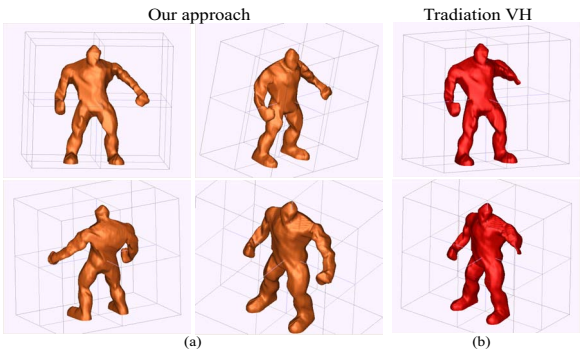


Figure 6. (a) Different views of the final reconstruction of the dataset shown in figure 4. Notice how the left arm of the model that undergoing non-rigid motion is accurately reconstructed. (b) Shows the reconstruction if traditional visual hull intersection is used on the same data. The arm is carved out due to the motion.

fusing the warped occupancy maps:

$$\theta_{ref} = \prod_{i=1}^{n} \hat{L}_i, \tag{3}$$

where θ_{ref} is the projectively transformed grid of object occupancy likelihoods or an object slice. Notice how using this homographic framework visual hull is being performed in the image plane without requiring to go in 3D space.

Subsequent slices or θs of the object are obtained by extending the process to planes parallel to the reference plane in the normal direction. Homographies of these new planes can be obtained using the relationship in equation 3. Occupancy grids/slices are stacked on top of each other creating a three dimensional data structure: $\Theta = [\theta_1; \theta_2; \ldots \theta n]$ that encapsulates the object shape. Θ is not an entity in the 3D world or a collection of voxels. It is, simply put, a logical arrangement of planar slices, representing discrete samplings of the continuous occupancy space. Object structure is then segmented out from Θ i.e., simultaneously from all the slices by evolving a smooth surface $\mathcal{S} : [0, 1] \rightarrow \mathbb{R}^3$ using level sets that divides Θ between the object and the background similar to the approach in [5].

4. Results and Experiments

We have tested our approach on several challenging monocular datasets. Figure 4 shows our 'The Thing' dataset. It consists of 20 views of a humanoid model captured with a camera moving around the object while the object deforms non-rigidly. (The left arm of the model is moving). In figure 4(b) two of the twenty blurred occupancy images (one in each view) produced using our approach are shown. Notice the occupancies in the region surrounding the left arm are blurred due to the motion of the arm. This region is selected and shown with more detail in the images in the second row of 4(b) together with the deblurred results. The motion deblurred occupancy images are then used to reconstruct the object using the image-based approach as described in section 3.3. A hundred slices were used to reconstruct the object. In figure 5(a) the reference view (from the deblurred occupancy images) used in reconstruction is shown with three of the hundred slices overlaid. The slices are also shown separately in log scale in 5(b) (redder is higher likelihood).

Figure 6(a) shows our reconstruction results. Notice that the left arm of the model that is undergoing motion is accurately reconstructed. There is some loss of detail in the reconstruction primarily due to limited number of views (twenty) and since we did not use 3D calibration (monocular un-calibrated camera sequence), performing visual hull intersection on cross-sectional slices using planar homographies. Yet to qualitatively assess the accuracy of our results we show in figure 6 (b) the reconstruction if the object is assumed to be rigid during the sequence and our occupancy deblurring approach is not used. It can be clearly seen that the left arm of the model is carved out by the visual hull intersection due to it's motion.

4.1. Quantitative Analysis

To quantitatively analyze our algorithm we conducted an experiment in which we obtained several monocular sequences of an object. In each flyby of the camera the object was kept stationary but after each cycle the posture of the object changed a little. We call this the 'Superman' dataset and is shown in figure 7(a). It consists of seven flybys of the camera around the object as the object deforms (both arms moving) after each sequence but is rigid within each. We call these the *rigid sequences* and each consists of 14 views of the object at a resolution of 480x720 with the object occupying a region of approximately 150x150 pixels. Figure 7(a) shows three of the seven rigid sequences. Notice the changing postures of the object between sequences. This data was used to obtain seven rigid reconstructions of the

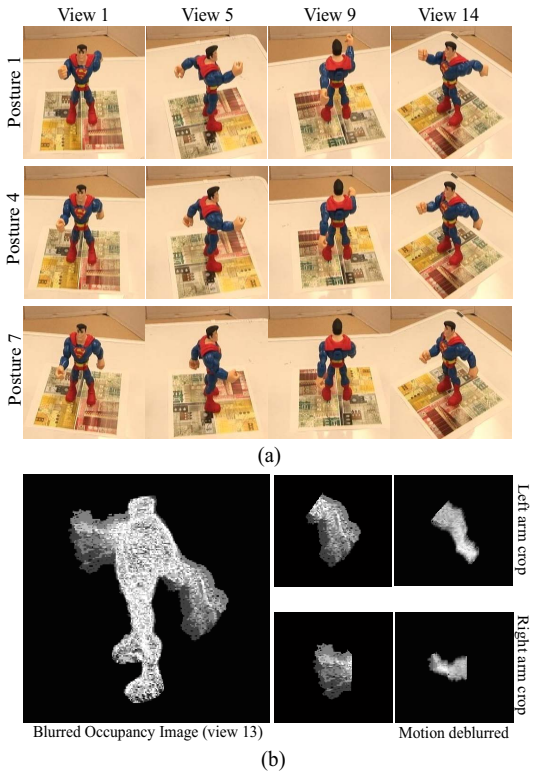


Figure 7. (a) Each row shows four views (of fourteen) from one of the seven monocular sequences in the dataset. The object is rigid within each sequence but changes posture between sequences by moving the arms (notice both arms moving progressively inwards). A monocular sequence of a non-rigidly deforming object is assembled by selecting two views in order from each rigid sequence. (b) The blurred occupancy image (one of fourteen) produced using our approach. The cropped, detail sections on the arms are shown on the right together with the deblurred results.

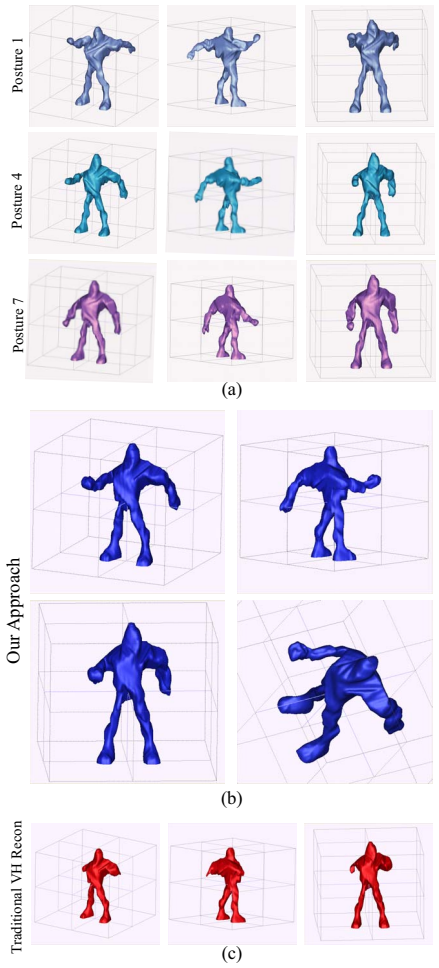


Figure 8. (a) Three of the seven visual hull reconstructions from the seven rigid sequences shown in figure 7(a). (b) Visualization of the reconstruction using our occupancy deblurring approach on the assembled non-rigid monocular sequence. Notice that the moving arms are accurately reconstructed using our approach but are carved out if we use conventional visual hull intersection that assumes the object is rigid as can be seen in the visualization in(c).

object, three of which are shown in figure 8(a).

A monocular sequence of a non-rigidly deforming object was assembled by selecting two views from each rigid sequence in order, thereby creating a set of fourteen views of the object as it changes posture (deforms non-rigidily). Reconstruction on this assembled non-rigid, monocular sequence was performed using our occupancy de-blurring approach and the visualization of the results are shown in figure 8(b). Notice using our approach the arms of the object are accurately reconstructed which are carved out when traditional visual hull intersection is used as shown in figure 8(c). For a quantitative analysis we compared our reconstruction results with each of the seven reconstructions from the rigid sequences. All the reconstructions were aligned in 3D (w.r.t the ground plane coordinate system) and the similarity was evaluated using a measure of the ratio of overlapping and non-overlapping voxels in the 3D shapes. The similarity measure is described as:

$$S_i = \left(\frac{\sum_{\forall v \in \mathbb{R}^3} \left((v \in O_{test}) \oplus (v \in O^i_{rig}) \right)}{\sum_{\forall v \in \mathbb{R}^3} \left((v \in O_{test}) \wedge (v \in O^i_{rig}) \right)} \right)^2, \quad (4)$$

where v is a voxel in the voxel space $\mathbb{R}^3$, O_{test} is the 3D reconstruction that needs to be compared with, O^i_{rig} the visual hull reconstruction from ith rigid sequence. S_i is the similarity score i.e. the square of the fraction of non-overlapping to overlapping voxels that are a part of the reconstructions, closer S_i is to zero greater the similarity.

In figure 9 we show a plots of the similarity measure. For the red plot (traditional visual hull reconstruction) the similarity is consistently quite low (measure high). This is expected since the moving parts of the object (arms) are

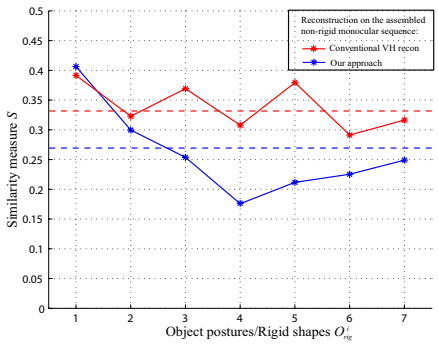


Figure 9. Plot of the similarity measure between reconstructions from the assembled monocular sequence and the rigid sequences in the 'Superman' dataset.

carved out by the visual hull intersection as can be seen in the visualization of this reconstruction in figure 8(c). For the blue plot (our approach) notice a clear dip in the similarity measure value at rigid shape 4 demonstrating quantitatively that the result of using our approach is most similar to this shape. This also corroborates what can be visually observed by comparing our reconstruction results shown in figure 8(b) with the reconstruction of the fourth rigid sequence shown in the second row of figure 8(a).

5. Conclusions

We have proposed an image-based approach to reconstruct non-stationary, articulated objects from silhouettes obtained with a monocular video sequence. Our approach starts with an silhouette fusion step that combines color and silhouette images to produce the blurred occupancy images, where the values at each pixel correspond to the fraction of the total time duration that the pixel observed an occupied scene location. We then use a motion de-blurring approach to de-blur the occupancy images. The de-blurred occupancy images correspond to a silhouettes of the mean/motion-compensated object shape and are used to obtain a visual hull reconstruction of the object. We have shown compelling results on challenging monocular datasets of non-stationary, articulated motion where traditional visual hull intersection approaches fail to reconstruct the object correctly.

Acknowledgments

This research was funded in part by the U.S. Government VACE program.

References

[1] A. Laurentini. The Visual Hull Concept for Silhouette- Based Image Understanding. IEEE TPAMI, 1994.

[2] G. K. M. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time:a 3D reconstruction algorithm combining shape-frame-silhouette with stereo. CVPR03.

[3] G. K. M. Cheung, S. Baker, and T. Kanade. Shape-From-Silhouette of Articulated Objects and its Use for Human Body Kinematics Estimation and Motion Capture. CVPR03.

[4] G. P. Stein, A. Shashua. Model-Based Brightness Constraints: On Direct Estimation of Structure and Motion. TPAMI 2000.

[5] A. Yezzi and S. Soatto, Stereoscopic segmentation, IJCV, 2003.

[6] S. Nayar and Y. Nakagawa. Shape from focus. IEEE TPAMI, 1994.

[7] S. M. Khan, P. Yan, M. Shah. A Homographic Framework for the Fusion of Multi-view Silhouettes. IEEE ICCV 2007.

[8] L. Torresani, A. Hertzmann, and C. Bregler. Non-Rigid Structure-From-Motion: Estimating Shape and Motion with Hierarchical Priors. IEEE TPAMI 2008.

[9] B. Vijayakumar, D. Kriegman, and J. Ponce. Structure and motion of curved 3D objects from monocular silhouettes. IEEE CVPR 1996.

[10] P. Favaro, S. Soatto. A variational approach to scene reconstruction and image segmentation from motion-blur cues. IEEE CVPR 2004.

[11] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image Based Visual Hulls. In SIGGRAPH, 2000.

[12] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. IEEE CVPR 2006.

[13] K. Wong and R. Cipolla. Structure and motion from silhouettes. IEEE ICCV 2001.

[14] S. Sullivan and J. Ponce. Automatic Model Construction, Pose Estimation, and Object Recognition from Photographs using Triangular Splines. IEEE TPAMI, 1998.

[15] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm, ICCV 2001.

[16] K. Kutulakos and S. Seitz. A Theory of Shape by Space Carving. IJCV, 2000.

[17] O. Faugeras and B. Mourrain, On the Geometry and Algebra of the Point and Line Correspondences Between N Images, IEEE ICCV 1995.

[18] R. Raskar, A. Agrawal, and J. Tumblin. Coded exposure photography: Motion deblurring via fluttered shutter. SIGGRAPH, 2006.

[19] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. SIGGRAPH, 2006.

[20] M. Ben-Ezra and S. K. Nayar. Motion-based motion deblurring. TPAMI, 2004.

[21] J. Jia. Single Image Motion Deblurring Using Transparency. In proceedings of CVPR 2007.

[22] S. Cho, Y. Matsushita and S. Lee. Removing Non-Uniform Motion Blur from Images.IEEE ICCV 2007.