Vision, Modeling, and Visualization (2017)
M. Hullin, R. Klein, T. Schultz, A. Yao (Eds.)

Template-Based 3D Non-Rigid Shape Estimation from Monocular Image Sequences

L. Kausch[1] , A. Hilsmann[1,2] and P. Eisert[1,2]

[1]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Germany
[2]Humboldt University, Berlin, Germany

Abstract

This paper addresses the problem of reconstructing non-rigid 3D geometries from temporal image sequences captured with only a single camera under full perspective projection. Without the knowledge of a shape deformation model, this task is severly under-constrained, because multiple shape configurations can produce the same image projections. The challenge remains even if a template 3D model of the static, un-deformed state is available, because the depth along the line of sight is unkown. Often, this is handled by assuming an orthographic camera model. In contrast, we address a full perspective camera model. Also, our reconstruction is not limited to the model parts that are visible in the current image, but deformation is estimated for the entire template across the temporal sequence. In a first step, we compute a template of the geometry in un-deformed pose, assuming that the object was captured while being static. Next, the object starts to deform while being captured by a single camera, and the non-rigid shape is reconstructed sequentially by estimating the camera position and the deformations with respect to the template model. Our objective minimization function combines image data and temporal consistency information, and constrains the deformation space by a rotation-invariant volumetric graph Laplacian and as-rigid-as-possible constraints defined on the tesselation of the template model. The method is evaluated on synthetic and real data, including different object classes, thereby concentrating on the class of articulated deformations.

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Curve, surface, solid, and object representation I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Motion, Shape, Tracking

1. Introduction

3*D* shape from 2*D* image observation is a subset of inverse problems encountered in human and computer vision, known as 3*D* reconstruction problems. For a human it is usually an easy task to infer the 3*D* structure shown in an image. Similarly, in the computer vision field, 3*D* reconstruction from multi-view image sets, displaying a static object, is a well explored topic [TMHF99]. When the object does not move, the available redundancy from multiple views can be exploited. However, many objects exist that can take different 3*D* configurations. As soon as the object deforms over time, rigid methods will fail. If the deforming object is captured by only a single camera, resulting in monocular image information per object configuration, the reconstruction turns into an underdetermined problem. In order to solve the problem, prior information about the unknown object is needed.

The approaches for non-rigid monocular reconstruction presented in the literature can be divided into three main classes. Traditional non-rigid structure from motion approaches (NRSfM) reconstruct deformable objects from points being tracked throughout a video sequence [BHB00, RRA13, ZTH13, AmMAC14]. A recent review can be found in [WYJZ15]. Incorporating a statistical shape prior, the 3*D* deforming shapes are represented by a linear combination of basis shapes. The basis shapes and the shape coefficients are recovered from the image projection matrix together with the camera motion in a batch approach. This limits the applicabitily of these methods to models with small deformations where the whole surface is visible during the acquisition time. Those methods that do consider feature occlusion model occlusions as outliers [THB08]. Hence, these methods cannot handle severe occlusions that are likely to occur, especially for volumetric deformable objects. The second class involves machine learning approaches that make use of 3*D* training data to learn a deformation model [GWBB09, MSS*17]. Thus, these methods are only applicable to a restricted number of object classes, namely to those where an appropriate training set in terms of 3*D* scans is available. Thirdly, template-based methods [SF10, BGCC12, KKBJ16, HE09] assume the availability of a 3*D* shape model prior to reconstruction and attempt to estimate the surface deformation in a frame-to-frame approach for consecutive images.

Since we are interesed in reconstructing the entire deformable volumetric shapes, our method has to deal with occlusions and disocclu-

DOI: 10.2312/vmv.20171257

sions in the monocular image sequence. Moreover, no prior information about the deformation model is available. Our work builds on template-based approaches. As shape prior, we combine two constraints that supplement each other. On the one hand side, we constrain the volumetric deformable reconstruction problem using the volumetric graph Laplacian (VGL) introduced by [ZHS*13]. The Laplacian encodes surface details as the difference between each mesh vertex and the average of its neighbors. By applying the Laplacian to a volumetric graph defined inside the mesh, VGL imposes volumetric contraints that penalize unnatural volume changes. This is combined with an as-rigid-as-possible (ARAP) constraint [SA07] with the objective of preserving the object surface properties during deformation.

Our method can be divided in two components: The first step deals with the template computation of the rest pose. For this purpose, the object is captured in its initial state with a multi-view camera set-up, such that traditional rigid reconstruction techniques can be employed for template generation [Wu13, WACS11, BBE14]. The template serves as a geometric and topological prior for the next step, where the template model is modified in order to satisfy the image data fitting constraints imposed by the new input frame depicting the object in a deformed state. The fitting constraints consist of point correspondences and color-dependent silhouette constraints. Unlike other template-based approaches [KKBJ16], our method does not require any user-input to establish 2*D*-3*D* correspondences. The approach is evaluated on a variety of generic volumetric objects

The remainder of this paper is structured as follows: The next section gives an overview on related works, followed by a section that describes our method to reconstruct a deformable volumetric object from monocular images under full perspective projection. Finally, experiments and results on synthetic and real sequences are presented in Section 4.

2. Related Work

There are very few non-rigid reconstruction methods presented in the literature that attempt to reconstruct an entire deformable volume from monocular image observations. The presented approach can be categorized into template-based reconstruction methods. These approaches have in common that they assume the 3*D* shape in one reference image to be known prior to reconstruction. Point correspondences between this reference image and a current image are established and the goal is to recover the deformations of the 3*D* template vertices such that the shape best conforms to the image observation. Still, the depth of the vertices along the line of sight is ill-constrained and different penalty functions have been proposed to overcome these ambiguities, including temporal consistency across consecutive images and geometric constraints on the template shape, cp. [SF10] for an extensive review.

Early template-based reconstruction methods focused on the reconstruction of developable surfaces, that are fully observed during acquisition [SHF07, SMNLF08, BGCC12, OVF12].

Volumetric Non-Rigid Reconstruction is even more challenging than the reconstruction of planar-like surfaces, because only the front part of the object surface is visible in the image, while the back surface and the interior have to be inferred without direct image information. Recently, a few methods have been presented that make template-based approaches applicable to volumetric objects:

[VA13] combine a template- and silhouette-based reconstruction approach under orthographic projection. The deformation is constrained by volumetric inextensibility constraints defined on virtual nodes in the mesh interior. This method requires considerably less point correspondences, but is limited to objects that have a plane of symmetry parallel to the image plane and does not infer concavities. Inspired by this, we go even further and address generic objects without topological restrictions under full perspective projection.

[KKBJ16] deform a 3*D* template to fit user-clicked 2*D*-3*D* correspondences under weak perspective projection. This method uses the ARAP [SA07] as shape constraint but allows for non-uniform, local deformations by imposing a sparsity constraint on local stiffness. The method simultaneously estimates an object specific stiffness model and the deformation of the mesh with respect to several different object instances in a global optimization. It is not optimized for large pose difference between the template and the pose depicted in the target images where it can result in erroneous camera estimation. Moreover, reconstructed parts may be bent in an unnatural direction. We, too, employ an ARAP deformation constraint and add a temporal consistency constraint. This enables feature tracking such that user input is not necessary anymore.

[YRCA15] compute a dense template using a short rigid sequence. Sequentially, a photometric cost is minimized that simultaneously estimates dense image correspondences and 3*D* deformations. The deformation is regularized spatially by the ARAP surface functional without any additional volumetric constraints.

[ZNI*14] present a template-based non-rigid reconstruction framework that achieves real-time performance on diverse scenes. In contrast to our preconditions, their methods work on temporal depth image sequences captured with a stereo camera setup. In a frame-to-frame approach they first align the template model to the current depth data and subsequently, perform non-rigid surface fitting by minizing geometric and photometric constraints, where the deformation is penalized by an ARAP shape regularizer.

Our contribution is a non-rigid shape estimation method that is general applicable to a broad range of object classes. The only prerequisites are that the object is captured in rest pose and that the surface contains enough details for feature extraction. We do not require any user-input to establish correspondences between the template model and the current image. In contrast to many other approaches, our method uses a full perspective camera model and handles fully volumetric objects.

3. Method for Non-Rigid Reconstruction

We formulate the problem of monocular deformable reconstruction in the following way: Given a deformable monocular RGB image sequence $\left\{ I^f : \Omega^f \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3 \right\}_{f=1}^{F}$ together with a 3*D* template model $\mathbf{X}^0$ as input, the aim is to estimate a deforming surface $\mathbf{X}^f$ across a temporal sequence $f \in \{1, \cdots, F\}$. We use a perspective camera model and assume that the RGB camera capturing the deforming object is calibrated such that the internal camera parameters are known. The perspective projection of a point $\mathbf{x}^f \in \mathbb{R}^3$ at frame f with external camera parameters $\mathbf{R}^f$ and $\mathbf{t}^f$ and calibration

matrix $\mathbf{K}_f$ is defined by a function $g^f : \mathbb{R}^3 \times \mathbf{R}^f \times \mathbf{t}^f \times \mathbf{K}_f \rightarrow \Omega^f \subset \mathbb{R}^2$.

3.1. Template Computation

The 3*D* template model consists of a triangular mesh with N surface vertices $\mathbf{X}^0 = \left\{ \mathbf{x}_i^0 \in \mathbb{R}^3 | \, i = \{1, \cdots, N\} \right\}$. Connection between these vertices are defined via a neighborhood set $\mathcal{N}_i$, that includes all the vertex indices that are connected to vertex $\mathbf{x}_i^0$. The mesh topology is constant for the entire sequence. The template mesh is computed from a rigid multi-view image sequence using structure-from-motion [WACS11], subsequent point cloud densification [FP10], and Poisson surface meshing [KH13]. The pipeline for template mesh creation is shown in Figure 1a-1c.
To further enable volume deformation constraints without making strong object assumptions like kinematic skeletons or parametric shape models, volume vertices $\mathbf{Y}^0 = \left\{ \mathbf{y}_i^0 \in \mathbb{R}^3 | \, i = \{1, \cdots, M\} \right\}$ are added in the mesh interior and the volumetric template is tesselated with tetrahedra constrained by the volume vertices [Si15]. The inner volume edges are inferred by a set $\mathcal{M}_i$, similar to the surface mesh topology. It is assured that volume vertices are evenly distributed including thin regions and their distance is similar to the average surface edge distance, guaranteeing equally shaped tetrahedra. The template volume graph is depicted in Figure 1d for an exemplary object. The template serves as a geometrical and topological prior for non-rigid reconstruction.

3.2. Energy function for non-rigid reconstruction

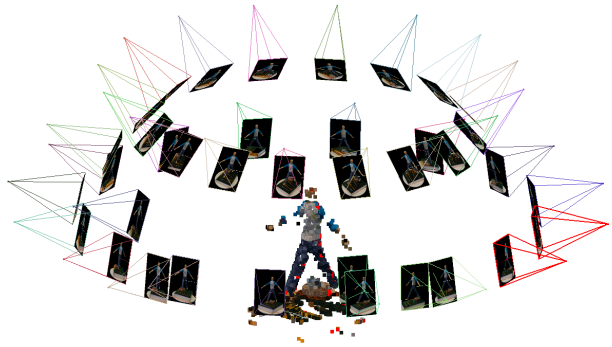
The goal is to determine the locations of the template vertices $\mathbf{X}^t$ at any time t. We assume a known internal calibration matrix $\mathbf{K}$. In addition, we expect the sourrounding scene of the deforming object to contain sufficiently rigid parts for global rotation and translation estimation $(\mathbf{R}^f, \mathbf{t}^f)$, similar to [YWSHSH15]. Then, the unknown deformation of the template mesh is estimated by minimization of a non-linear energy function in a frame-to-frame approach using the Levenberg-Marquardt-Algorithm. Consequently, the running time complexity grows linear with the number of frames. The optimization value is initialized for each frame with the estimated shape from the previous time instance. The energy function (1) comprises two main terms, one accounts for the image data fitting (E_{fit}), the other controls the smoothness of the deformation (E_{reg}). For an arbitrary time instant t, the optimization problem can be formulated as

$$\min_{\mathbf{X}^f, \mathbf{Y}^f} E(\mathbf{X}^f, \mathbf{Y}^f) = \min_{\mathbf{X}^f, \mathbf{Y}^f} E_{fit}(\mathbf{X}^f, \mathbf{Y}^f) + E_{reg}(\mathbf{X}^f, \mathbf{Y}^f). \quad (1)$$

The fitting term E_{fit}, given by

$$E_{fit}(\mathbf{X}^f, \mathbf{Y}^f) = \lambda_p E_{point}(\mathbf{X}^f) + \lambda_s E_{sil}(\mathbf{X}^f, \mathbf{Y}^f) \text{ with } \lambda_p, \lambda_s \in \mathbb{R},$$

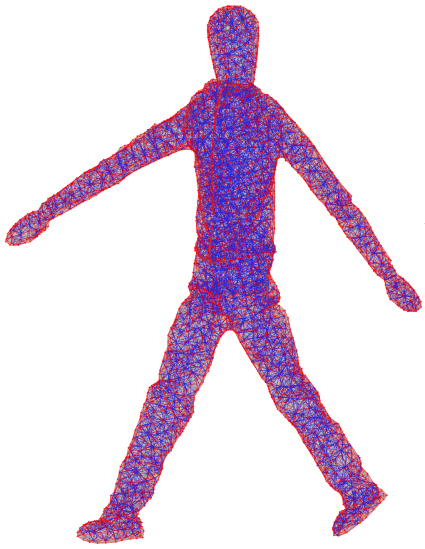
consists of a weighted sum of point correspondences (2) and silhouette constraints, as in [VA13], but extended with color conditions (3). Both terms require image data information.
The point correspondences assure that specific 3*D* surface points project to the correct image location. For this purpose, inter-frame 2*D*-2*D* SIFT matches are computed [FE13], that can be related to points on the surface of the previously computed 3*D*



(a) *Camera pose estimation with SFM [WACS11]*

(b) *Dense point cloud [FP10]*

(c) *Poisson surface mesh [KH13]*



(d) *Volumetric graph structure*

Figure 1: *Pipeline for template construction from a rigid image sequence using structure-from-motion, followed by interior volume tesselation. Surface vertices and edges are colored in red, while volume vertices and edges are highlighted in blue.*

shape model in terms of barycentric coordinates., assuming that the mesh facets are sufficiently small such that they remain flat as the surface deforms and perspective effects are negligible. This results in a set of $2D$-$3D$ correspondence pairs $C^f = \left\{ (\mathbf{u}_i^f, \mathbf{v}_i^f) \,|\, \mathbf{u}_i^f \in \Omega^f, \mathbf{v}_i^f = \sum_{j=1}^3 b_{i,j}^f \mathbf{x}_{k_{i,j}^f}^f \in \mathbb{R}^3 \right\}_{i=1}^{P^f}$, where $b_i^f \in \mathbb{R}^3$ is the barycentric coordinate of a surface point $\mathbf{v}_i^f$ that is contained in a mesh triangle with vertex indices $k_i^f \in \mathbb{R}^3$ and $\mathbf{u}_i^f$ is the corresponding $2D$ image location. P_f specifies the number of correspondences. Hence, the point correspondence constraint can be formulated as

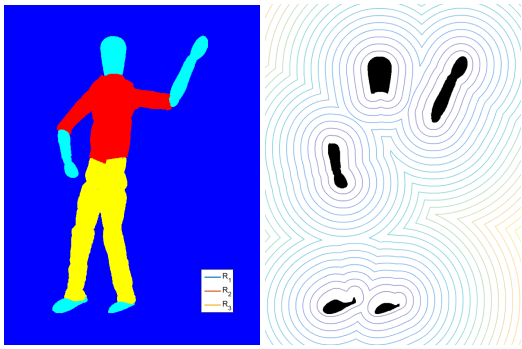
$$E_{point}(\mathbf{X}^f) = \frac{1}{P^f} \sum_{i=1}^{P^f} \left\| \left(g^f(\Sigma_{j=1}^3 b_{i,j}^f \mathbf{x}_{k_{i,j}^f}^f) - \mathbf{u}_i^f \right) \right\|^2. \quad (2)$$

The silhouette constraint penalizes volume configurations that project outside the image silhouette. This requires the input images to be segmented [LSS09]. Similar to [VA13], we compute for each image location the Euclidean distance to the closest silhouette point. For complicated deformations, vertices may project into a silhouette part that is related to a different surface region. To alleviate these apparent minima, we compute color-based silhouettes for specified color ranges. If a color is associated to each template surface vertex, one can constrain the visible vertices to project inside the related color-based silhouette while the non-visible and volume vertices are constrained to project inside the union of all these silhouettes. A visibility flag $\psi_i \in \{0,1\}$ for each surface vertex is computed prior to each Levenberg-Marquardt update step by rendering the current $3D$ volume with the estimated camera parameters. Let $D_{\mathcal{R}}^f : \Omega^f \to \mathbb{R}$ define the Euclidean distance for each pixel to the closest pixel inside the silhouette and let $D_{\mathcal{R}_j}^f$ define the distance map for specified color ranges such that $\cap_j \mathcal{R}_j = \emptyset$ and $\cup_j \mathcal{R}_j = \mathcal{R}$, where $\mathcal{R}$ is the region of the entire silhouette. In addition, a region flag η_i indicates for each vertex $\mathbf{x}_i$ the index j of the corresponding color region $\mathcal{R}_j$. Figure 2 visualizes this for one exemplary input image. Then, the color-based silhouette constraint is given by

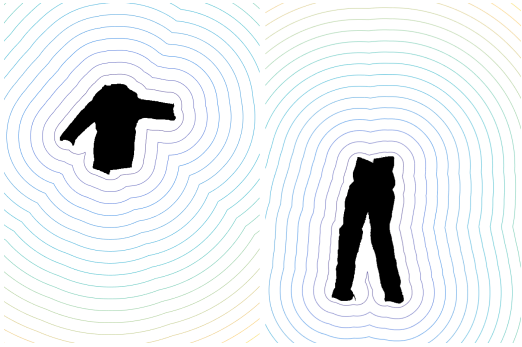
$$E_{sil}(\mathbf{X}^f, \mathbf{Y}^f) = \frac{1}{N} \sum_{i=1}^{N} \left(\psi_i \sum_j \mathbb{1}_{\eta_i}(j) \left\| D_{\mathcal{R}_j}^f(g^f(\mathbf{x}_i^f)) \right\|^2 + (1-\psi_i) \left\| D_{\mathcal{R}}^f(g^f(\mathbf{x}_i^f)) \right\|^2 \right) + \frac{1}{M} \sum_{i=1}^{M} \left\| D_{\mathcal{R}}^f(g^f(\mathbf{y}_i^f)) \right\|^2. \quad (3)$$

The monocular image information alone leaves reconstruction ambiguities. Thus, a regularization term that comprises a weighted combination of three terms is added: Firstly, a temporal smoothness term that penalizes strong frame-to-frame deformations (4), secondly, spatial smoothness is imposed by an as-rigid-as possible functional on the mesh surface (5), and thirdly, volume preservation is realized by a rotation-invariant volumetric graph Laplacian (6). This can be formulated as

$$E_{reg}(\mathbf{X}^f, \mathbf{Y}^f) = \gamma_t E_{temp}(\mathbf{X}^f) + \gamma_s E_{surface}(\mathbf{X}^f) + \gamma_v E_{volume}(\mathbf{Y}^f)$$



(a) *Definition of color-based regions.* **(b)** *Distance map corresponding to* $\mathcal{R}_1$.



(c) *Distance map corresponding to* $\mathcal{R}_2$. **(d)** *Distance map corresponding to* $\mathcal{R}_3$.

Figure 2: *Definition of color regions* $\mathcal{R}_i$ *with related color-based distance maps that define for each pixel the distance to the closest pixel inside the specified region. The Euclidean distance is visualized with contour lines. The union of all color-based regions is equal to the silhouette.*

with weighting coefficients $\gamma_t, \gamma_s, \gamma_v \in \mathbb{R}$.
The temporal smoothness term can be formulated as

$$E_{temp}(\mathbf{X}^f) = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}_i^f - \mathbf{x}_i^{f-1} \right\|^2, \quad (4)$$

where $\mathbf{X}^{f-1}$ defines the $3D$ surface of the previous time frame. This constraint encourages temporally smooth deformations.
The second term allows local surface deformations that do not alter the relative locations between vertices and each of their neighbors, thereby preserving surface details. This deviation is measured by the as-rigid-as possible criterion [SA07] defined by

$$E_{surface}(\mathbf{X}^f) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left\| (\mathbf{x}_i^f - \mathbf{x}_j^f) - R_i(\mathbf{x}_i^{f-1} - \mathbf{x}_j^{f-1}) \right\|^2, \quad (5)$$

where R_i defines the local rotations for one vertex between consecutive frames, taking the surrounding neighborhood into account. The local rotation is re-computed in each optimization step.
These two terms retain temporal and surface smoothness during deformation, but unnatural volume changes can still occure due to strong deformations. To preserve the volume, we impose the graph

Laplacian on the interior template graph structure. The volumetric graph Laplacian was introduced by [ZHS*13] and applied to transfer curve-based deformations of 2*D* cartoon characters to 3*D* meshes. It is applied to the interior volume graph $\left(\mathbf{Y}^f, \{\mathcal{M}_i\}_i\right)$ and can be formulated as

$$E_{volume}(\mathbf{Y}^f) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}\left\|\mathcal{L}(\mathbf{y}_i^f) - R_i\mathcal{L}(\mathbf{y}_j^0)\right\|^2, \quad (6)$$

where $\mathcal{L}(\mathbf{y}_i) = \mathbf{y}_i - \frac{1}{|\mathcal{M}_i|}\sum_{j\in\mathcal{M}_i}\mathbf{y}_j$ defines the Laplacian coordinates of the volume vertices, paticularly $\mathcal{L}(\mathbf{y}^0)$ specify the Laplacian coordinates of the template model in undeformed state. These are transformed by a local rotation, computed just as in the ARAP case, thus, allowing for locally rigid changes in the Laplacian difference vectors.
Each criterion term is normalized such that they are not influenced by the resolution of the surface, the density of the volume sampling or the number of feature correspondences.

4. Experimental Results

In this section, we present results obtained on three different datasets, covering different object classes. Two datasets, the pixar lamp (Figure 6) and the sackboy sequence (Figure 7), are synthetically generated. Hence, they allow for quantitative evaluation. The third data sequence of a jointed doll (Figure 5), obtained from real image observations, is evaluated qualitatively. The datasets used in the experiments are publically available at https://cvg.hhi.fraunhofer.de/.

4.1. Real data

The jointed doll sequence was acquired with a single RGB camera. Between each time instance the 3*D* object configuration is modified as in a stop motion film. This results in a smooth articulated movement of the limbs across the sequence. The sequence includes particular challenges due to intra-object occlusion, that can be noted in the last three frames of Figure 5, where the right arm is occluded by the body, and the left hand moves in front of the face. For template creation, the object was captured in its rest pose with a structure from motion method. To guide our monocular non-rigid reconstruction procedure, feature correspondences between consecutive frames are established as shown in Figure 3 [FE13]. Final results of our volumetric non-rigid reconstruction can be seen in Figure 5 for some exemplary frames of the jointed doll sequence, as well as in the supplemental video. The presented method is able to reconstruct the present 3*D* deformation while preserving the interior volume properties.

4.2. Synthetic data

Two further image sequences were created synthetically to enable quantitative performance evaluation of the proposed method. The rigged pixar lamp is publically available at http://uploaded.net/file/3vl8g79c. The sackboy model http://voila3d.com/model.php?view=LittleBigPlanet_Sackboy_3d_model__QU2UNIBX9UCSM0IS0UEW2F0H8 was rigged in blender and

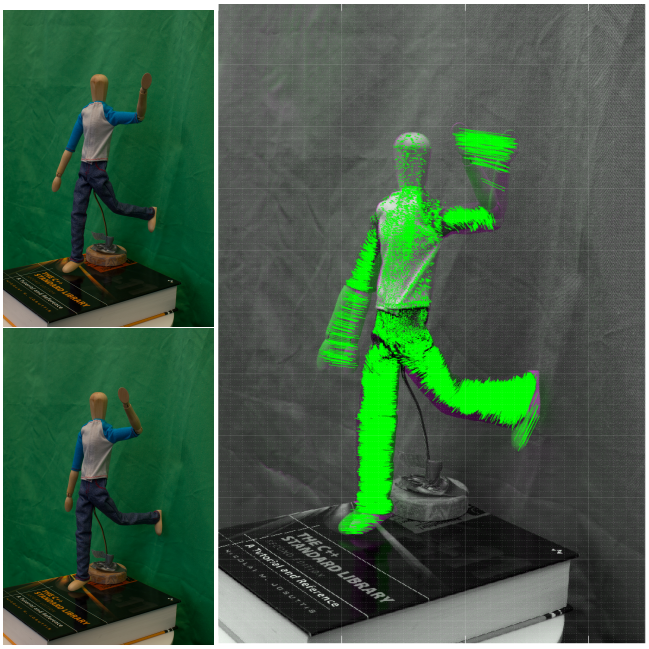


Figure 3: *2D SIFT correspondences highlighted in green between two consecutive image frames shown in the left column.*

both models were animated and rendered, in order to generate the 2*D* monocular image sequence. The articulated joints of the two models are shown in Figure 4. The 3*D* mesh of the first frame is

Figure 4: *Definition of articulated joints for the two synthetically generated deformation sequences. Joints are labeled in red.*

used as the template model in rest pose. Qualitative results for the 3*D* deformation estimation are shown in Figure 6 and 7 and in the supplemental video.
In order to access the quality of the estimated reconstruction with respect to the original 3*D* sequence quantitatively, we evaluated the Hausdorff distance [CRS98]. This metric computes the largest distance that occurs between one point on one mesh and its closest surface point on the other mesh. The distance metric is not symmetric. We compute the one-sided Hausdorff distance from the original mesh to the reconstructed mesh, sampled at

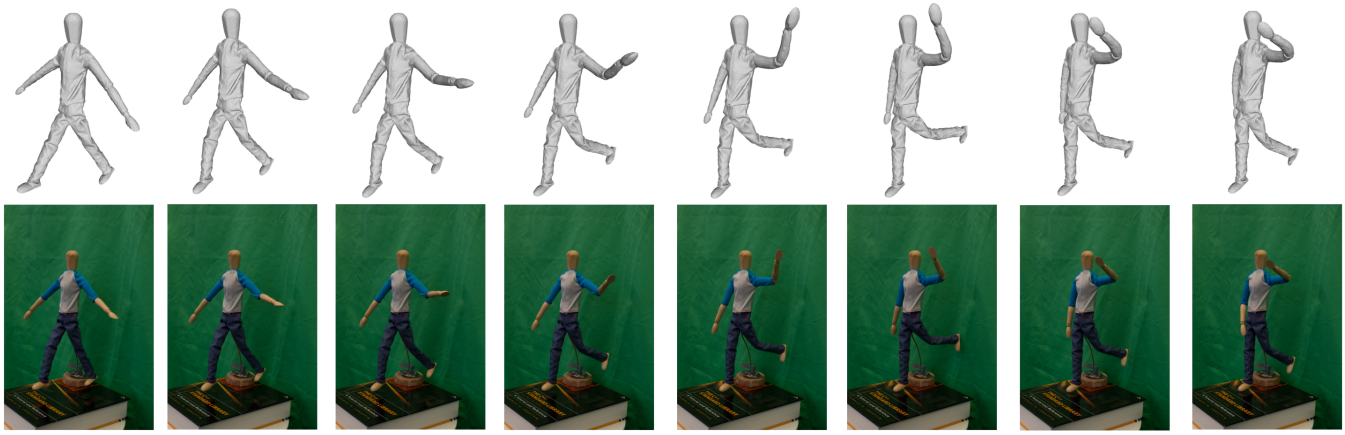


Figure 5: *A template model for the jointed doll (first column), generated from a rigid multi-view sequence, is modified to be consistent with a monocular image sequence (bottom). At each time instance the camera parameters are esimated from rigid correspondences in the background (here: book).*

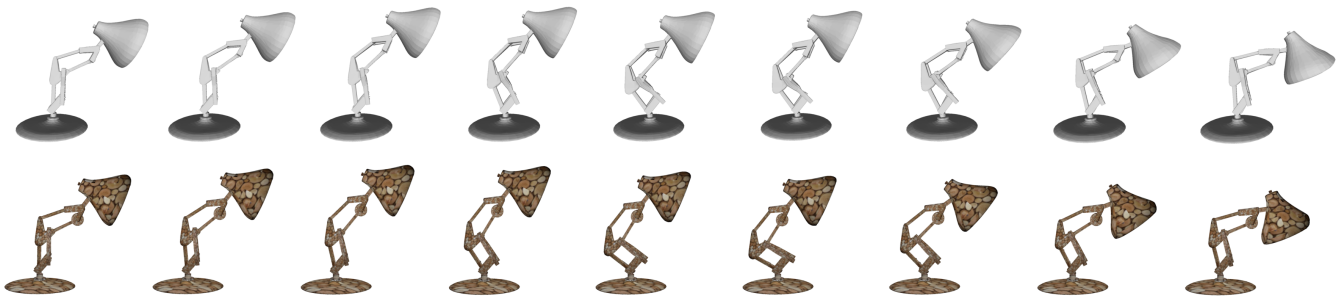


Figure 6: *A dynamic 3D volume model is generated (top) that captures different object configurations shown in a monocular image sequence (bottom) for the pixar lamp.*

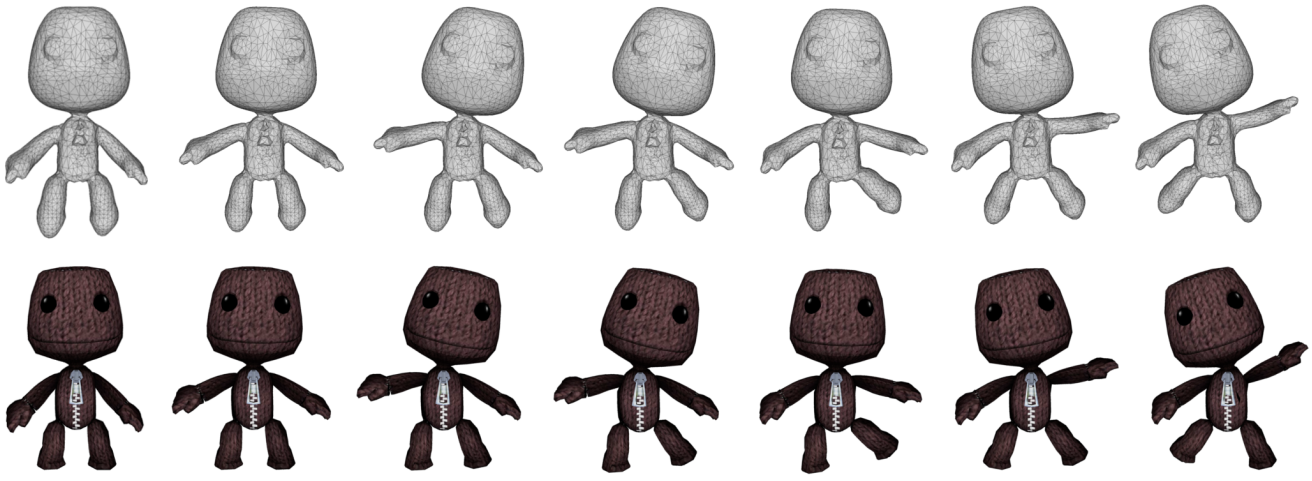


Figure 7: *Non-rigid shape estimation from our method on the sackboy sequence.*

f	2	3	4	5	6	7	8	9	10	11
pixar lamp	1.91	1.95	2.05	2.18	2.40	2.66	2.91	3.10	3.21	3.23
sackboy	3.43	3.37	3.31	3.28	3.23	3.21	3.17	3.13	3.12	3.12

f	12	13	14	15	16	17	18	19	20
pixar lamp	3.21	3.16	3.05	3.00	2.98	2.99	2.97	2.95	2.90
sackboy	3.11	3.08	3.10	3.11	3.10	3.12	3.18	3.21	3.25

Table 1: *One-sided Hausdorff distance between the true mesh deformation and the estimated deformed mesh, generated from the knowledge of monocular image information per time instance. The unity is in percent with respect to the diagonal of the object bounding box.*

all vertex locations. Let the true vertex locations be given by $\mathbf{V}^f = \left\{ \mathbf{v}_i^f \in \mathbb{R}^3 | \, i = \{1, \cdots, N\} \right\}$ and let $M^f = (\mathbf{X}^f, \mathcal{N}_i)$ define the triangulated surface of the estimated mesh.
Then, the one-sided Hausdorff metric can be formulated as

$$d_H^f = \sup_{\mathbf{v}_i^f \in \mathbf{V}^f} \inf_{\mathbf{x}^f \in M^f} \left\| \mathbf{v}_i^f - \mathbf{x}^f \right\|$$

for all frames f.
Table 1 lists the values for the two 20-frame synthetic datasets, where frame 1 corresponds to the rest pose. The error distance is measured with respect to the diagonal of the bounding box of the mesh and is expressed in percent. For both sequences the maximal error is below 3.32% in relation to the diagonal length across all frames.

5. Conclusion

We presented a non-rigid reconstruction approach from monocular images under full perspective projection. The ill-conditioned problem is regularized by utilizing the knowledge about a 3*D* template model in rest pose and imposing surface and volume constraints on this geometry. The data term is a pairwise term that encourages correct projection of corresponding points and at the same time guides the deformation by silhouette information where color consistency can be incorporated. The method fulfils our objective of generality, because it is independent of any user input and capable to cope with volumetric deforming objects. We have shown results on a novel real world camera sequence, as well as a qualitaive evaluation on two new synthetic sequences.
As we are interested in learning deformation parameters from 2*D* images, a next step will be to apply motion separation algorithms to the 3*D* sequences, and use the obtained results for joint estimation. The information about partially rigid parts could then be used to improve the deformation estimation for articulated objects.

Acknowledgments

This research has received funding from the EUs Horizon 2020 research and innovation programme under grant agreement number 687757 (REPLICATE).

References

[AmMAC14] AGUDO A., M. MONTIEL J. M., AGAPITO L., CALVO B.: Online dense non-rigid 3d shape and camera motion recovery. In *British Machine Vision Conference (BMVC)* (2014). 1

[BBE14] BLUMENTHAL-BARBY D., EISERT P.: High-resolution depth for binocular image-based modelling. *Computers & Graphics 39* (2014), 89–100. 2

[BGCC12] BARTOLI A., GÉRARD Y., CHARDEBECQ F., COLLINS T.: On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), pp. 2026–2033. 1, 2

[BHB00] BREGLER C., HERTZMANN A., BIERMANN H.: Recovering non-rigid 3d shape from image streams. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2000), vol. 2, pp. 690–696. 1

[CRS98] CIGNONI P., ROCCHINI C., SCOPIGNO R.: Metro: Measuring error on simplified surfaces. *Computer Graphics Forum 17* (1998), 167–174. 5

[FE13] FURCH J., EISERT P.: An iterative method for improving feature matches. In *IEEE International Conference on 3DTV* (2013), pp. 406–413. 3, 5

[FP10] FURUKAWA Y., PONCE J.: Accurate, dense, and robust multiview sterepsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32*, 8 (2010), 1362–1376. 3

[GWBB09] GUAN P., WEISS A., BALAN A. O., BLACK M. J.: Estimamting human shape and pose from a single image. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 1381–1388. 1

[HE09] HILSMANN A., EISERT P.: Joint estimation of deformable motion and photometric parameters in single view video. In *IEEE International Conference on Computer Vision (ICCV)* (2009), pp. 390–397. 1

[KH13] KAZHDAN M., HOPPE H.: Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG) 32*, 3 (2013), 29. 3

[KKBJ16] KANAZAWA A., KOVALSKY S., BASRI R., JACOBS D.: Learning 3d deformation of animals from 2d images. *Computer Graphics Forum 35*, 2 (2016), 365–374. 1, 2

[LSS09] LIU J., SUN J., SHUM H.-Y.: Paint selection. *ACM Transactions on Graphics (ToG) 28*, 3 (2009), 69. 4

[MSS*17] MEHTA D., SRIDHAR S., SOTNYCHENKO O., RHODIN H., SHAFIEI M., SEIDEL H.-P., XU W., CASAS D., THEOBALT C.: VNect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)* (2017). 1

[OVF12] ÖSTLUND J., VAROL A., FUA P.: Laplacian meshes for monocular 3d shape recovery. In *European Conference on Computer Vision (ECCV)* (2012), pp. 412–425. 2

[RRA13] R.GARG, ROUSSOS A., AGAPITO L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 1272–1279. 1

[SA07] SORKINE O., ALEXA M.: As-rigid-as-possible surface modelling. In *Symposium on Geometry Processing* (2007), vol. 4. 2, 4

[SF10] SALZMANN M., FUA P.: Deformable surface 3d reconstruction

from monocular images. *Synthesis Lectures on Computer Vision 2*, 1 (2010), 1–113. 1, 2

[SHF07] SALZMANN M., HARTLEY R., FUA P.: Convex optimization for deformable surface 3-d tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2007). 2

[Si15] SI H.: Tetgen, a delaunay-based quality tetrhedral mesh generator. *ACM Transactions on Mathematical Software (TOMS) 41*, 2 (2015), 11. 3

[SMNLF08] SALZMANN M., MORENO-NOGUER F., LEPETIT V., FUA P.: Closed-form solution to non-rigid 3d surface registration. In *European Conference on Computer Vision (ECCV)* (2008), pp. 581–594. 2

[THB08] TORRESANI L., HERTZMANN A., BREGLER C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 30*, 5 (2008), 878–892. 1

[TMHF99] TRIGGS B., MCLAUCHLAN P., HARTLEY R., FITZGIBBON A.: Bundle adjustment: A modern synthesis. In *Vision Algorithms Workshop: Theory and Practice* (1999), pp. 298–372. 1

[VA13] VICENTE S., AGAPITO L.: Balloon shapes: reconstructing and deforming objects with volume from images. In *IEEE International Conference on 3DTV* (2013), pp. 223–230. 2, 3, 4

[WACS11] WU C., AGARWAL S., CURLESS B., SEITZ S. M.: Multicore bundle adjustment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2011). 2, 3

[Wu13] WU C.: Towards linear-time incremental structure from motion. In *IEEE International Conference on 3DTV-Conference* (2013), pp. 127–134. 2

[WYJZ15] WANG Y., YAN X., JIANG M., ZHEN J.: Research on non-rigid structure from motion: A literature review. *Journal of Fiber Bioengineering and Informatics 8* (2015), 751–760. 1

[YRCA15] YU R., RUSSEL C., CAMPBELL N. D. F., AGAPITO L.: Direct, dense and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 918–926. 2

[YWSHSH15] YUCER K., WANG O., SORKINE-HORNUNG A., SORKINE-HORNUNG O.: Reconstruction of articulated objects from a moving camera. In *IEEE International Conference on Computer Vision Workshops* (2015), pp. 28–36. 3

[ZHS*13] ZHOU K., HUANG J., SNYDER J., LIU X., BAO H., GUO B., SHUM H.-Y.: Large mesh deformation using the volumetric graph laplacian. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35* (2013). 2, 5

[ZNI*14] ZOLLHÖFER M., NIESSNER M., IZADI S., REHMANN C., ZACH C., FISCHER M., WU C., FITZGIBBON A., LOOP C., THEOBALT C., STAMMINGER M.: Real-time non-rigid reconstruction using rgb-d camera. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 156. 2

[ZTH13] ZHANG X., TANG A., HUNG Y.: A decomposition method for non-rigid structure from motion with orthographic cameras. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)* (2013), p. 1. 1