ANIMATABLE 3D MODEL GENERATION FROM 2D MONOCULAR VISUAL DATA

Philipp Fechteler[1,2], *Lisa Kausch*[1], *Anna Hilsmann*[1,2] *and Peter Eisert*[1,2]

[1]Fraunhofer HHI, [2]Humboldt Universität zu Berlin, Germany

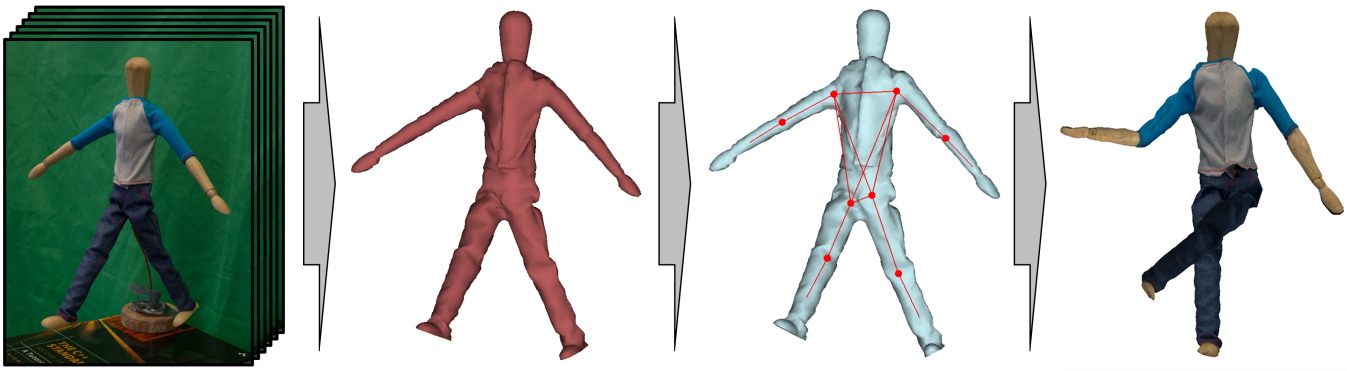

Fig. 1. Visualization of pipeline (from left to right): monocular RGB input sequence, reconstructed 3D geometry, reconstructed animatable model based on kinematic skinning, textured and animated model.

ABSTRACT

In this paper, we present an approach for creating animatable 3D models from temporal monocular image acquisitions of non-rigid objects. During deformation, the object of interest is captured with only a single camera under full perspective projection. The aim of the presented framework is to obtain a shape deformation model in terms of joints and skinning weights that can finally be used for animating the model vertices. First, the monocular rigid shape estimation problem is solved by computing a template model of the object in rest pose from an image sequence. Next, the unknown external camera parameters and the deformation for each vertex are estimated alternately in a sequential approach. The resulting consistent non-rigid shape geometries are used to compute a kinematic skeleton control structure including skinning weights and optimized shape. For that, a completely data-driven optimization scheme is used, which iterates over three steps: (a) optimization of pose for each frame as well as joint parameters consistent over the entire sequence, (b) optimization of rest pose vertices to enhance the shape and (c) optimization of skinning weights for improved deformation characteristics. With experimental results on publicly available synthetic as well as real-world datasets, we demonstrate the quality of the proposed approach. The resulting models with fixed topology and rigged with skeleton and skinning weights can be animated in existing render engines.

***Index Terms*—** deformation model, animation, monocular non-rigid reconstruction, articulation estimation

1. INTRODUCTION

Tracking of articulated objects in monocular image sequences is a challenging problem due to non-rigidity, self-occlusion, high degree of freedom, and inherent ambiguity. Many existing methods [1, 2] incorporate prior knowledge by employing a 3D pose database. These approaches have in common that the deformation models need to be pre-trained and in consequence are specific to the underlying database. Additionally, the models' shape need to be adapted to the observed targets.

In contrast, we build articulated models directly from 2D visual data without including any task specific model assumptions and without requiring a 3D pose database. We combine template-based monocular non-rigid reconstruction with a subsequent articulated model extraction approach. The presented method is generally applicable to a wide range of different non-rigid objects, where the observed deformations can be broadly explained through kinematic deformations. Besides, the approach only requires a simple single camera setup and is easy to use because of its automatic work flow.

The overall task of articulated model estimation is targeted by combining two different estimation methods: monocular, deformable structure from motion [3] and automated model rigging [4].

For the first step, a template-based volumetric non-rigid 3D reconstruction approach from monocular images under full perspective projection is exploited [3]. With monocular image information, this task is highly ambiguous, because multiple shape configurations can produce identical image projections. The ill-conditioned problem is regularized by utilizing knowledge about a 3D template model in rest pose and by imposing surface and volumetric constraints on the geometry. The method is independent of any user-input and capable to cope with fully volumetric objects, where the back surface and the interior have to be inferred without direct image information.

The second step is based on a data-driven approach [4] to calculate an animatable model from the topologically consistent meshes from the previous step. For this mesh sequence, an initial kinematic

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
doi 10.1109/ICIP.2018.8451630

skeleton structure is computed using k-means. In a data-driven optimization loop, the vertices, skinning weights as well as joints are optimized in order to resemble as closely as possible the training set. The output of the approach is a fully animatable, articulated 3D mesh model that has automatically been created from a sequence of 2D images of the object in different sample poses.

2. DEFORMABLE MODEL ESTIMATION

The pipeline for our deformable model estimation from monocular visual data is depicted in Fig. 1 and 2. Initially, a short image sequence of the object of interest is captured in a static period and used for template generation. The template resembles the object shape in rest pose. Additionally, the object articulation is captured with a single camera system providing monocular input for the non-rigid shape estimation module. This component estimates temporal, volumetric deformation of the initial template from single images having a partial view on the object. The output of this component is an animated mesh sequence with vertex consistency across time that serves as pose database for the articulated model estimation module. This module calculates an optimized set of rest pose vertices to resemble as good as possible the given pose database with an articulated mesh model, as well as the required skinning weights and skeletal joints. The resulting model is animatable through a kinematic skeleton control structure and mimics closely the captured object. Finally, to further increase the visual quality of resulting animations, a texture can be computed from the multi-view sequence of the object, using standard texturing techniques.

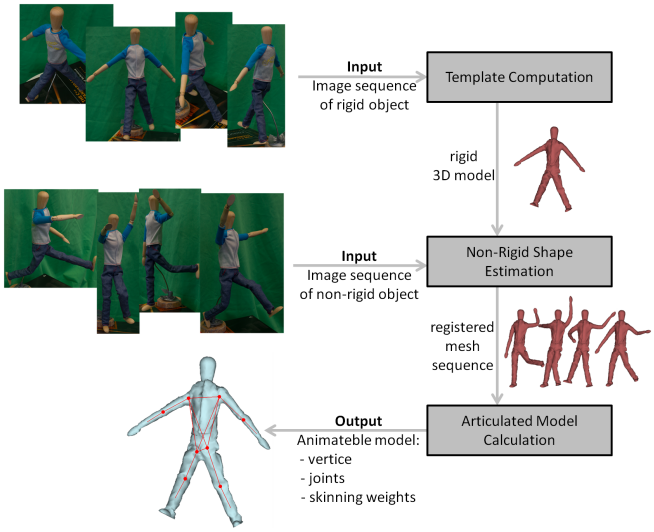

Fig. 2. Overview of processing pipeline.

2.1. Template Computation

The first step to calculate an animatable model is to reconstruct a 3D template of the captured object, with the pipeline shown in Fig. 3. This 3D template model is computed from the rigid image sequence using structure-from-motion [5], subsequent point cloud densification [6], and Poisson surface meshing [7]. To further enable volume deformation constraints without making strong object assumptions, volume vertices are added in the mesh interior and the volumetric template is tessellated with tetrahedra constrained by the volume vertices [8]. The inner volume edges are inferred 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. This template model serves as a geometrical and topological prior for the non-rigid reconstruction.

Fig. 3. Initial template generation (from left to right): camera pose estimation with SfM, dense point cloud, Poisson surface mesh, volumetric graph structure.

2.2. Non-Rigid Shape Estimation

The volumetric shape estimation from a non-rigid monocular image sequence is solved sequentially with a global reference template [3]. The internal calibration matrix is assumed to be known. After template computation, the external camera parameters and the unknown deformation for each vertex are estimated alternately in a sequential approach. While the camera parameters are estimated from rigid background correspondences, the unknown deformation is estimated at each time step by solving a non-linear optimization problem. The objective function used here combines fitting constraints with a regularization that controls the smoothness of the deformation. Let $\mathbf{X}^0$ denote the 3D template model together with the monocular non-rigid image sequence $\{\mathbf{I}^f, f = 1, \ldots, F\}$ for different time instants f. Then, we optimize for the unknown shape $\mathbf{X}^f$ by solving

$$\min_{\mathbf{X}^f} E(\mathbf{X}^f|\mathbf{I}^f, \mathbf{X}^0) = \min_{\mathbf{X}^f} E_{\text{fit}}(\mathbf{X}^f|\mathbf{I}^f) + E_{\text{reg}}(\mathbf{X}^f|\mathbf{X}^0).$$

The pairwise fitting constraint E_{fit} consists of a weighted sum of 3D-2D point correspondence constraints assuring that dedicated surface points project to related image locations and a silhouette constraint with color consistency extension that constrains the entire volume to project into the image silhouette. The regularization constraint

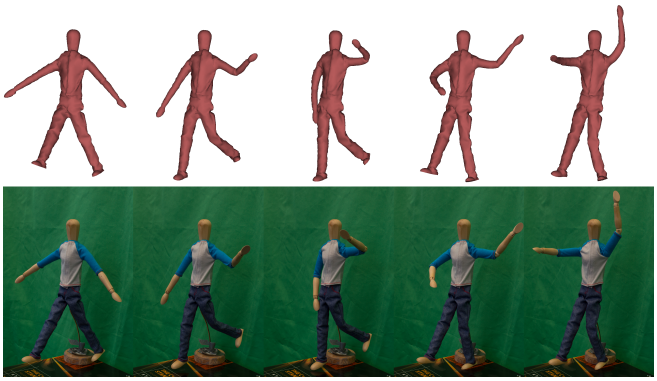

Fig. 4. Reconstructed poses obtained with non-rigid reconstruction approach from captured images.

includes temporal, surface and volume smoothness assumptions defined on the internal graph structure of the template model. Thus, the reconstruction is secured to result in a smooth deforming sequence that preserves surface details as well as avoids volume shrinkage. The output is a deforming mesh sequence with fixed triangle topology according to the defined template model, as shown in Fig 4.

2.3. Articulated Model Calculation

In order to calculate an animatable model from the input sequence consisting of registered example meshes of the captured object in different poses, we employ the data-driven optimization framework presented in [4]. This method uses as skinning function a combination of Linear Blend Skinning (LBS) and Dual quaternion Linear Blending (DLB) in order to reduce typical skinning artifacts.

Initially, the mesh of the animatable model is randomly taken from the set of registered example meshes. Using k-means, approximately rigid body parts are clustered, forming an initial skeleton structure and binary skinning weights (rigid kinematics without any smooth interpolation) [9]. Further, the optimization framework needs as input the initial configuration for each joint as well as the global alignment (rotation and translation) for each example mesh. This is addressed by treating each joint/limb of each mesh independently as an Orthogonal Procrustes alignment problem [10] which is solved efficiently via Singular Value Decomposition.

In order to solve for optimal rest pose vertices, skinning weights and joints, to resemble closely the example meshes, we setup an objective function E_{kin} consisting of a typical data term as well as a shape prior. The data term enforces that the vertices of the animated model are close to the corresponding vertices in the example meshes. Using a mesh Laplacian [11], the shape prior term (weighted with $\alpha \geq 0$) enforces that the vertices of the animated kinematic shape model have similar neighborhood properties as their corresponding vertices in the example meshes.

With $\mathcal{L}()$ being the uniform Laplace operator and $\mathcal{S}_f()$ being the LBS/DLB-based skinning function parameterized to transform the models vertices from rest pose to the pose of frame f, the objective function of the kinematic shape model is written for F example meshes, each having $|\mathbf{X}^0|$ vertices, as:

$$E_{\text{kin}} = \sum_{\substack{f=1\ldots F \\ j=1\ldots|\mathbf{X}^0|}} \left|v_j^{\text{scan}_f} - \mathcal{S}_f\left(v_j\right)\right|^2 + \alpha \left|\mathcal{L}\left[v_j^{\text{scan}_f}\right] - \mathcal{L}\left[\mathcal{S}_f\left(v_j\right)\right]\right|^2$$

with $v_j^{\text{scan}_f}$ being the jth vertex of example mesh f and v_j being the corresponding j^{th} vertex of the kinematic shape model in rest pose.

The optimization to attain the desired kinematic shape model adapted to the training set iterates successively over three optimization steps:

Pose and Joints Optimization refines the pose parameters for each frame as well as the kinematic joints for the whole sequence via Gauss-Newton by using the analytic differentiation of the objective function E_{kin}.

Shape Optimization enhances the vertex locations of the model in rest pose by rearranging the equation system into one huge sparse matrix and solving it directly for the least squares solution.

Skinning Weight Optimization improves the deformation characteristics of the model using the sampling based coordinate descent algorithm [12].

The resulting kinematic shape model is optimized to resemble the input training set as close as possible.

Finally, the animatable model can be equipped with photo-realistic appearance properties by computing a texture from the rigid input sequence using standard texturing techniques, as shown in Fig. 1 and 6. In order to import the animatable model into standard modeling/animation software, straightforward linearization techniques like [13] can be employed for conversion.

3. EXPERIMENTAL RESULTS

Results for our animatable 3D model generation framework from monocular 2D image input are presented on two publicly available datasets [14]: the synthetic sackboy sequence (consisting of 20 non-rigid input images + rigid template mesh) and the real-world puppet sequence (consisting of 20 rigid and 18 non-rigid input images). While the sackboy dataset is synthetically generated, the puppet dataset was acquired with a single RGB camera. During acquisition, the object configuration was modified.

In the synthesized sackboy image sequence with 980×1280 resolution, the head, arms and legs have been moved. Consequently, the resulting animatable 3D model (2.5k vertices and 5k triangles) contains a skeletal control structure consisting of 5 joints, as shown in Fig. 5.

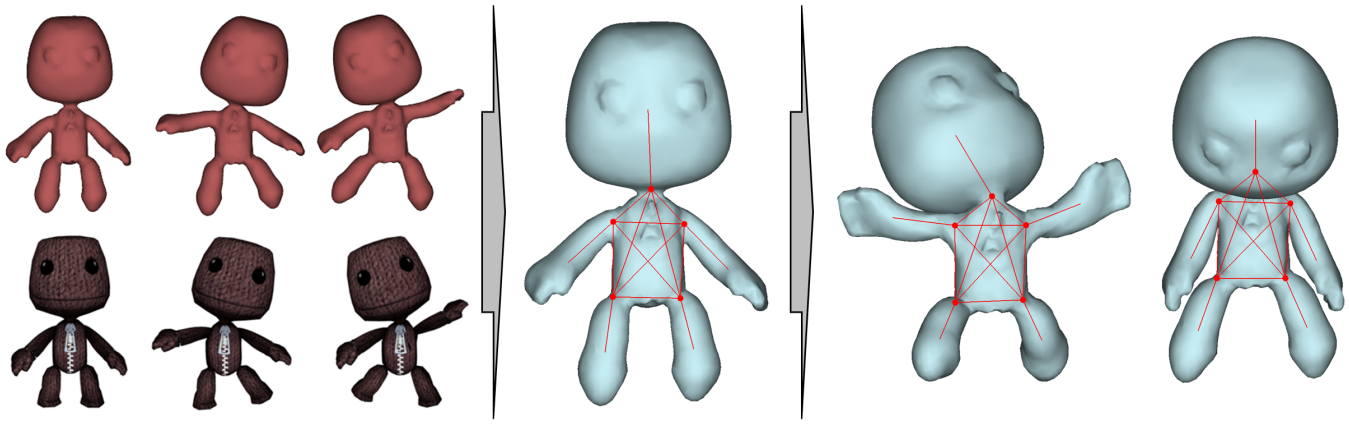

Fig. 5. Reconstructed poses (top left) obtained from synthesized images (bottom left), generated animatable model (middle) and example animations of new poses, which have been synthesized and not captured (right).

Fig. 6. Textured animation examples of animatable model calculated from puppet dataset, with rest pose in the middle.

During capturing of the real-world puppet dataset with a 3456×5184 resolution Canon EOS 550D DSLR camera, the shoulders, elbows, thighs and knees where moved. As a result the corresponding animatable 3D model (3.4k vertices and 6.8k triangles) contains a skeletal control structure consisting of 8 joints, as shown in Fig. 1. To demonstrate the achieved visual quality, example animations with a 1024×1024 resolution texture extracted from the rigid input sequence are shown in Fig. 6.

4. CONCLUSION

In this paper, we presented a fully automatic approach for creating animatable 3D models from temporal monocular 2D image acquisitions of non-rigid objects. During deformation, the object of interest is captured with only a single camera under full perspective projection. With the presented framework, an animatable 3D shape deformation model based on kinematic skinning is calculated from captured image sequences, consisting of vertices, skinning weights and joints. Initially, the monocular rigid shape is estimated by computing a template model of the object in rest pose from a rigid image sequence. Next, the external camera parameters and the unknown deformation for each vertex are estimated alternately in a sequential approach. The resulting consistent non-rigid shape geometries are used to compute a kinematic skeleton control structure including skinning weights and optimized shape. This completely data-driven optimization uses 3 steps to approximate the registered example meshes as closely as possible: pose and joints optimization, shape optimization and skinning weight optimization. With experiments on publicly available datasets the achieved quality of this approach is demonstrated.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] P. Guan, A. Weiss, A. Balan, and M. Black, "Estimating Human Shape and Pose from Single Images," in *IEEE 12th International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009, pp. 1381–1388.

[2] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei amd H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera," in *ACM Transaction on Graphics (TOG)*, 2017.

[3] L. Kausch, A. Hilsmann, and P. Eisert, "Template-Based 3D Non-Rigid Shape Estimation from Monocular Image Sequences," in *Proc. Vision, Modeling and Visualization (VMV)*, Bonn, Germany, September 2017.

[4] P. Fechteler, A. Hilsmann, and P. Eisert, "Example-based Body Model Optimization and Skinning," in *Proc. EUROGRAPHICS*, Lisbon, Portugal, May 2016.

[5] C. Wu, S. Agarwal, B. Curless, and S. Seitz, "Multicore Bundle Adjustment," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 3057–3064.

[6] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[7] M. Kazhdan and H. Hoppe, "Screened Poisson Surface Reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 29:1–29:13, July 2013.

[8] H. Si, "TetGen, a Delaunay-Based Quality Tetrahedral Mesh Generator," *ACM Trans. Math. Softw.*, vol. 41, no. 2, pp. 11:1–11:36, Feb. 2015.

[9] B. Le and Z. Deng, "Smooth Skinning Decomposition with Rigid Bones," *ACM Transactions on Graphics (SIGGRAPH Asia2012)*, vol. 31, no. 6, December 2012.

[10] D. Akca, "Generalized Procrustes Analysis and its applications in Photogrammetry," Tech. Rep., ETH Zurich, 2003.

[11] O. Sorkine, "Differential Representations for Mesh Processing," *Computer Graphics Forum*, 2006.

[12] A. Jacobson, Z. Deng, L. Kavan, and JP Lewis, "Skinning: Real-time Shape Deformation," in *ACM SIGGRAPH Courses*, 2014.

[13] L. Kavan, S. Collins, and C. O'Sullivan, "Automatic Linearization of Nonlinear Skinning," in *Proc. Symposium on Interactive 3D Graphics and Games*, Boston, MA, USA, February 2009.

[14] Lisa Kausch, "Volumetric Deformable Structure From Motion Free Dataset," https://cvg.hhi.fraunhofer.de/VoldSfM.html.