Simultaneous Pose Estimation and Augmentation of Elastic Surfaces from a Moving Monocular Camera

Nazim Haouchine, Marie-Odile Berger, Stephane Cotin

Simultaneous Pose Estimation and Augmentation of Elastic Surfaces from a Moving Monocular Camera

Nazim Haouchine* Inria, Mimesis Group
Marie-Odile Berger† Inria, Magrit Group
Stephane Cotin‡ Inria, Mimesis Group

ABSTRACT

We present in this paper an original method to estimate the pose of a monocular camera while simultaneously modeling and capturing the elastic deformation of the object to be augmented. Our method tackles a challenging problem where ambiguities between rigid motion and non-rigid deformation are present. This issue represents a major lock for the establishment of an efficient surgical augmented reality where endoscopic camera moves and organs deform. Using an underlying physical model to estimate the low stressed regions our algorithm separates the rigid body motion from the elastic deformations using polar decomposition of the strain tensor. Following this decomposition, a constrained minimization, that encodes both the optical and the physical constraints, is resolved at each frame. Results on real and simulated data are exposed to show the effectiveness of our approach.

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Physically based modeling

1 INTRODUCTION AND RELATED WORKS

The augmentation of non-rigid shapes from a moving monocular camera is considered to be a challenging problem with many applications such as endoscopic surgery guidance [8]. Its complexity arises from two main ambiguities. First, several 3D shape configurations may produce the same 2D projection on an image. Second, the decomposition of the rigid motion, produced by the camera displacement, and the non-rigid motion, corresponding to the inner deformation of the object, is a non-trivial task. These two aspects make the problem under-constrained and have led the community to consider additional constraints like smoothness, shading, isometry or physical priors.

The problem of estimating camera pose while recovering and augmenting a non-rigid 3D object is known as non-rigid Structure-from-Motion (NR-SfM), and can be seen differently depending on the types of deformation that are considered and the inputs that are used. To handle global deformation, statistical and low-rank approaches have been widely used since they can efficiently represent global deformations from a linear combination of basic shapes [5, 7, 6]. They are often used off-line, where all frames are batch processed, however, some studies have investigated the on-line processing for non-rigid motion continuously [17]. These methods, however, fail at capturing local deformations and are most of the time, dedicated to articulated deformations.

Physical priors have been recently introduced for simultaneous and sequential pose and shape estimation. In [2], the authors combined a physical model derived from Navier's equations with an Extended Kalman Filter to efficiently estimate 3D elastic shapes while simultaneously compute the camera pose. Similarly, [3] proposed to incorporate a dynamic particle model into a bundle adjustment framework, where [1] proposed to use modal analysis and reduced models to represent the deformed shape based on vibrations analysis. The problem becomes an eigenvalue problem, where the pose is estimated online using bundle adjustment. These methods show convincing results with small deformations, however they lack experiments with highly elastic deformations and take the assumption of the presence of a known fixed region. Mechanical models have also been used for the estimation of 3D shapes from a monocular camera [13, 8, 12]. These approaches, known as Shape-from-Template (SfT), have first been dedicated to isometric shapes where geometrical models are sufficient to capture the deformation [4] before being extended to elastic shapes using various types of material laws [13, 8, 12]. Although these methods present an adequate use of physical priors, they assume a fixed camera input and have not yet been proven to be reliable with a moving camera.

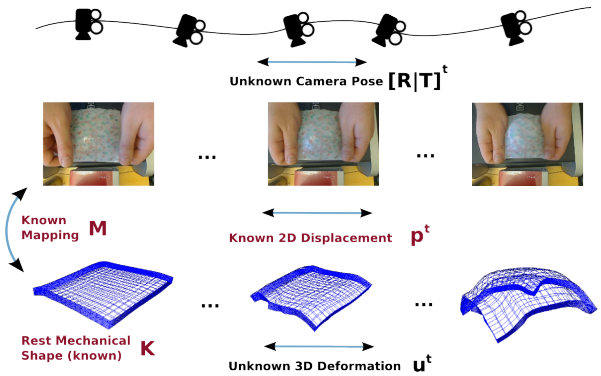


Figure 1: Problem formulation: we aim at sequentially recover the 3D elastic shape corresponding to **u**, from the known reprojected displacement in the image **p** while simultaneously estimating the camera rotation **R** and translation **T**.

Contribution: We introduce in this paper a method capable of estimating online camera pose and 3D elastic shape from a monocular camera using a known template. Because we are interested in elastic deformation, our template consists of a mechanical model with a known geometry and material properties. Our method is based on the decomposition of rigid body motion from elastic deformations using polar decomposition of the strain tensor. The pose is estimated from the rigid body motion by minimizing the reprojection error, while ensuring a temporal camera motion consistancy. 3D shape deformation is computed by solving a constrained minimization that encodes image-points as physical boundary conditions in an elegant manner.

*e-mail: nazim.haouchine@inria.fr
†e-mail: marie-odile.berger@inria.fr
‡e-mail: stephane.cotin@inria.fr

2 Problem formulation

We note $\mathbf{p} = \{p_i \in \mathbb{R}^2\}$ the vector of m non-homogeneous point coordinates where $p_i = (x_i, y_i)^\top$ and $\mathbf{u} = \{u_i \in \mathbb{R}^3\}$ the vector of n nodal positions of the physical model (template) in metric coordinates, where $u_i = (X_i, Y_i, Z_i)^\top$. Assuming an orthographic camera, the projection of the 3D points $\mathbf{u}^t$ onto the frame t can be expressed as:

$$p_i^t = \mathbf{R}^t u_i^t + \mathbf{T}^t \qquad \text{for } i = 1, \dots m \tag{1}$$

where $\mathbf{R}^t$ is a 2×3 truncated rotation matrix and $\mathbf{T}^t$ is a 2×1 translation vector. We introduce the stiffness matrix $\mathbf{K} \in \mathbb{R}^{3n \times 3n}$ that encodes the physical behavior of an object so that the equation of its elastic deformations has the form:

$$\mathbf{K}\mathbf{u} = \mathbf{f} \tag{2}$$

where the vector $\mathbf{f} \in \mathbb{R}^{3n}$ contains the n nodal forces acting on the object.

From the 2D measurements in the image corresponding to $\mathbf{p}^t$, the problem that we address can be formalized as finding for each frame t the components $(\mathbf{R}^t, \mathbf{T}^t, \mathbf{u}^t)$ that satisfy the projection constraints of Eq. (1) and the physical constraints of Eq. (2).

3 Decomposition of Rigid and Non-Rigid Motion

Among state-of-the-art elastic models [15], we choose the Saint Venant-Kirchhoff model [9] since it offers a good trade-off between computation time and accuracy. It handles non-linear deformations, shows rotational invariance and its simplicity allows real time computations. Its material is defined by a stress-strain relationship of the form:

$$\mathbf{S} = \eta(tr\mathbf{E})\mathbf{I}_3 + 2\mu\mathbf{E} \tag{3}$$

where $\mathbf{S}$ is the second Piola stress tensor, $\mathbf{E}$ is the Green-Lagrange strain tensor, $\mathbf{I}_3$ is the 3×3 identity matrix, η and μ are Lame coefficients and can be computed thanks to the elastic parameters of the material E and ν. E is Young's modulus and is a measure of the stiffness of the material while ν is Poisson's ratio and estimates the compressibility of the material.

One of the main properties of the behavior of a deformable object arises from the ability of decomposing a deformation into rigid and non-rigid components [16]. This decomposition can be done using a polar decomposition of the deformation gradient which consists of finding a rotation tensor $\mathbf{R}_b$ that minimizes $\|\mathbf{E}_b - \mathbf{R}_b\|_F^2$, where $\|\cdot\|_F$ is the Forbenius norm and the subscript b denotes the volume body. It gives rise to a decomposition $\mathbf{E}_b = \mathbf{R}_b\mathbf{D}_b$ where $\mathbf{R}_b$ is the measure of the rigid body rotation while $\mathbf{D}_b$ is a measure of the local stretching or contraction of the body. This decomposition is not unique, and the solution is chosen so that $\mathbf{D}_b$ is symmetric and $\mathbf{R}_b$ is orthogonal satisfying $\mathbf{R}_b^\top \mathbf{R}_b = \mathbf{I_3}$.

When considering the whole volume of a deforming object with a large number of degrees of freedom, the extraction of the rigid motion can lead to ambiguities between the camera and the object motion. In order to correctly extract the rigid part caused by the camera motion, we propose to locally decompose each element and to consider only the less stressed element. Assuming the finite element method is used (to discretize partial differential equations of solid continuum mechanics) upon tetrahedral elements, one can use the method describe in [14] to compute the rotation $\mathbf{R}_e$ and translation $\mathbf{T}_e$ of each tetrahedral element.

Let $u_1^0, u_2^0, u_3^0, u_4^0$ be the vertices of a tetrahedral element e in the undeformed configuration, and $u_1^t, u_2^t, u_3^t, u_4^t$ be their positions in the deformed configuration at frame t, and let us define the 4×4 matrices $\mathbf{U}_e^0$ and $\mathbf{U}_e^t$ of the form

$$\mathbf{U}_e^0 = \begin{bmatrix} u_1^0 & u_2^0 & u_3^0 & u_4^0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad , \quad \mathbf{U}_e^t = \begin{bmatrix} u_1^t & u_2^t & u_3^t & u_4^t \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{4}$$

There exists a unique matrix $\mathbf{V}_e = [\mathbf{U}_e^t][\mathbf{U}_e^0]^{-1}$ that describes the transformation of the tetrahedron e so that $u_i^t = \mathbf{V}_e u_i^0$ with $1 \leq i \leq 4$ and $\mathbf{V_e}$ takes the form

$$\mathbf{V}_e = [\mathbf{U}_e^t][\mathbf{U}_e^0]^{-1} = \begin{bmatrix} & \mathbf{E}_e & & \mathbf{T}_e \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} & \mathbf{R}_e\mathbf{D}_e & & \mathbf{T}_e \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5}$$

where $\mathbf{T}_e$ is a 3×1 translation matrix that contains the translational part of the deformation and $\mathbf{E}_e$ is the 3×3 deformation gradient that contains the rotational and elastic parts. Decomposing $\mathbf{E}_e$ using polar decomposition gives $\mathbf{E}_e = \mathbf{R}_e\mathbf{D}_e$ and permits to extract the 3×3 rotation matrix $\mathbf{R}_e$ that now contains only the rotational part of the deformation.

Assuming k elements, we can extract the global rigid motion of the object following:

$$\left[\mathbf{R}_b^t | \mathbf{T}_b^t\right] = \frac{1}{k}\left(\sum_{e=1}^{k} \frac{\|\sigma_{max}^t - \sigma_e^t\|_F}{\|\sigma_{max}^t\|_F}\left[\mathbf{R}_e^t | \mathbf{T}_e^t\right]\right) \tag{6}$$

where σ_e denotes the von Mises stress of the element e computed from the local stretching $\mathbf{D}_e$, and σ_{max} the highest von Mises stress of the object. This formulation permits to consider only the transformation of the rigid part of the deformable body and finally to extract the rigid body motion encoded in the 3×3 rotation matrix $\mathbf{R}_b$ and the 3×1 translation vector $\mathbf{T}_b$ (the subscript b denotes the body).

4 Estimation of camera pose from rigid motion

Once the rigid motion is extracted from Eq. 6, estimating the camera pose is straightforward and follows a classic Perspective-n-Point (PnP) problem [10]. The aim of the PnP problem is to determine the position and orientation of a camera given its intrinsic parameters $\mathbf{A}$ and a set of correspondences between 3D points and their 2D projections.

Recovering the camera pose knowing the rigid motion is reduced at finding $\mathbf{R}^t$ and $\mathbf{T}^t$ that minimize the reprojection error following:

$$\min_{\mathbf{R}^t, \mathbf{T}^t} \sum_{i}^{m} \|\mathscr{I}(p_i^t) - \mathbf{R}^t \mathscr{J}(u_i^t) - \mathbf{T}^t\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ is the Forbenius norm and $\mathscr{J}(\cdot)$ and $\mathscr{I}(\cdot)$ are transform functions that permits to extract the rigid body using $\mathbf{R}_b^t$ and $\mathbf{T}_b^t$. We add to this cost function the temporal consistency terms $\|\mathbf{R}^t - \mathbf{R}^{t-1}\|_F^2$ and $\|\mathbf{T}^t - \mathbf{T}^{t-1}\|_F^2$ to regularize the estimation of the pose between successive frames, so that rotations and translation matrices are coherent altogether. Moreover, the final minimization function can now be written as:

$$\min_{\mathbf{R}^t, \mathbf{T}^t} \sum_{i}^{m} \underbrace{\|\mathscr{I}(p_i^t) - \mathbf{R}^t \mathscr{J}(u_i^t) - \mathbf{T}^t\|_F^2}_{\text{Reprojection error}} + \underbrace{\alpha\|\mathbf{R}^t - \mathbf{R}^{t-1}\|_F^2 + \beta\|\mathbf{T}^t - \mathbf{T}^{t-1}\|_F^2}_{\text{Pose consistancy}} \tag{8}$$

where α and β are weighting parameters set empirically. Note that in contrast to related approaches, no shape regularization is introduced in the cost function since the mechanical model ensures the physical coherence of the shape.

5 Simultaneous pose estimation and augmentation

Let us return to the problem formulated in Section 2 where we seek at simultaneous recover at each frame t the camera pose $(\mathbf{R}^t, \mathbf{T}^t)$ and the 3D shape $\mathbf{u}^t$.

The projection constraints of Eq. 1 can be encoded as boundary conditions for the physical model of Eq. 2 by adjoining m Lagrange multipliers $\lambda_i = (l_x, l_y)^\top$ collected in vector $\lambda \in \mathbb{R}^2$ to form the Lagrangian

$$\mathscr{L}(\mathbf{u}, \lambda) = \frac{1}{2}\mathbf{u}^\top \mathbf{K}\mathbf{u} - \mathbf{u}^\top \mathbf{f} + \lambda^\top([\mathbf{R}]\mathbf{u} + [\mathbf{T}] - \mathbf{p}) \qquad (9)$$

The vector λ can be interpreted as forces required to maintain the boundary conditions $\mathbf{p}$, whereas $\mathbf{u}$ is the vector of solutions of the problem. The notation $[\mathbf{R}]$ denotes a block diagonal matrix of size $2m \times 3n$ built from $\mathbf{R}$ and $[\mathbf{T}]$ a $2m$ vector built from $\mathbf{T}$.

Algorithm 1 Simultaneous pose estimation and shape recovery

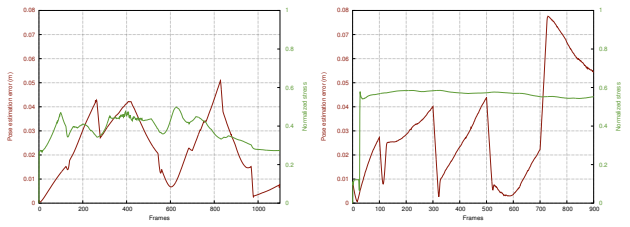
Input: Vector of 2D image positions $\mathbf{p}^t$, mechanical model at rest configuration $\mathbf{u}^0$.
Output: Camera pose $(\mathbf{R}^t, \mathbf{T}^t)$ and 3D shape $\mathbf{u}^t$.

1: **Initialisation:** Build $\mathbf{K}$ from constitutive law, material properties and template model.
2: **while** t **do**
3: Extract the rigid body $(\mathbf{R}_b^t, \mathbf{T}_b^t)$ using polar decomposition.
4: Estimate the camera pose $(\mathbf{R}^t, \mathbf{T}^t)$ from rigid body such as the quantity $\|\mathscr{I}(p_i^t) - \mathbf{R}^t\,\mathscr{J}(u_i^t) - \mathbf{T}^t\|$ is minimized.
5: Estimate the 3D shape of the object $\mathbf{u}^t$ such as it solves $\mathbf{K}\mathbf{u}^t = \mathbf{f}^t$ and $[\mathbf{R}]^t\mathbf{u}^t + [\mathbf{T}]^t = \mathbf{p}^t$ simultaneously.
6: **end while**

Initially, at $t = 0$ the vector of external forces $\mathbf{f}$ is null and both $\mathbf{K}$ and $\mathbf{u}$ are known. Thus, we can decompose the deformation into rigid and non-rigid components which will help at estimating the initial camera pose knowing the correspondence mapping $\mathscr{M}(\cdot)$ between $\mathbf{u}$ and $\mathbf{p}$. At $t > 0$, we sequentially and simultanoulsy solve the system of Eq. 9 following the steps described in Algorithm 1.

6 Results

We present in this section the results obtained using our method on real and synthetic data to demonstrate the ability of our approach to capture and augment 3D large elastic deformations. We first test our approach on computer-generated data where both the object deforms and the camera moves. We then use the dataset of [8] where several video sequences of a silicone-like object undergoing different types of elastic deformation is proposed. We quantify the three-dimensional shape recovery error with respect to a ground truth. In all experiments we use SIFT [11] to detect 2D features and track them over frames and we define the Z-axis as the camera axis. For each set we compute a 3D mean error (in mm) and root-mean square (RMS) error (in mm) as the vertex-to-vertex distance between the reconstructed mesh and the ground-truth mesh.



(a) (b)

Figure 2: Camera pose estimation error with respect to ground truth and model's stress.

We used the framework Sofa [1] to generate elastic deformations of a soft object constrained to produce a 3D deformation. A video sequence of 640×480 images is acquired using a virtual camera with focal length $f_u = f_v = 500$ and principal point at $(u_c, v_c) = (320, 240)$. The camera trajectory and the resulting shapes are reported in Figures 3.

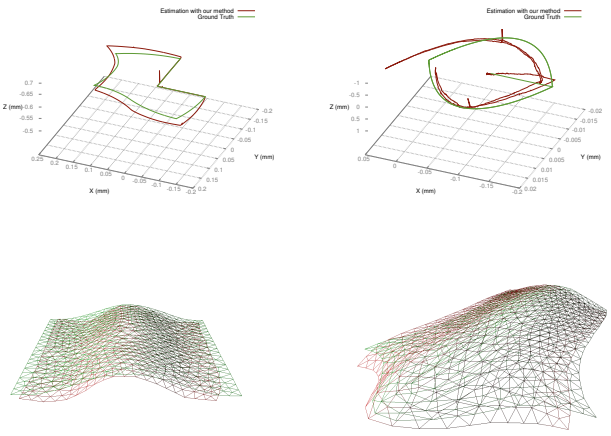


Figure 3: Results obtained on computer-generated data: [Top row] the estimated trajectory and its ground truth and [Bottom row] The recovered 3D mesh and its ground truth. (Left) elasticity = 25%, Mean error = 2.11 mm, RMS error = 2.33 mm — (Right) elasticity = 130%, Mean error = 0.96 mm, RMS error = 1.25 mm

In order to estimate the impact of the deformation on the estimation of the camera pose, we compute the mean value of the von Mises stress σ_{VM} of the object during the deformation and compute the absolute error between the estimated camera trajectory and its ground truth over frames. The von Mises stress is normalized so that 0 represents the rest state and 1 means that the object is highly deforming. This gives an estimation of the state of the object during the decomposition of rigid and non-rigid components. The trajectory error is estimated by a RMS error between the estimated pose and the ground truth at each frame. The results reported in the plots of Figure 2 show that the pose estimation is not linearly dependent on the object state.

We finally test our approach on real data from the silicone-elastic dataset [8]. This dataset consists of a silicone-made object deformed following several configurations with extensibility ranging from 25% to 120%. The silicone strip has a size of $100 \times 100 \times 10$ mm^3 and its stiffness is characterized by Young's Modulus $E = 250000$ Pa and Poisson's ratio $\nu = 0.45$. For each configuration a video sequence with an image resolution of 640×480 is acquired with a monocular camera at 30 fps. We virtually simulate a camera motion by pre-defining a path that includes translations and rotations. We obtain video sequences that are used as input for our algorithm and where we can quantify the pose error w.r.t the predefined path. The 3D shapes at the final state are provided and are considered as ground truth. The resulting 3D shapes, re-texturing and pose estimation are illustrated in Figure 4.

7 Conclusion

We proposed in this paper a method for simultaneously estimating the camera pose and augmenting elastic surfaces. The key idea was to use polar decomposition on an underlying physical model to separate rigid and non-rigid motion. While the rigid body is used to estimate the camera pose, the non-rigid motion is encoded onto a constrained minimization - formulated using Lagrangian Multipliers-

[1] http://www.sofa-framework.org

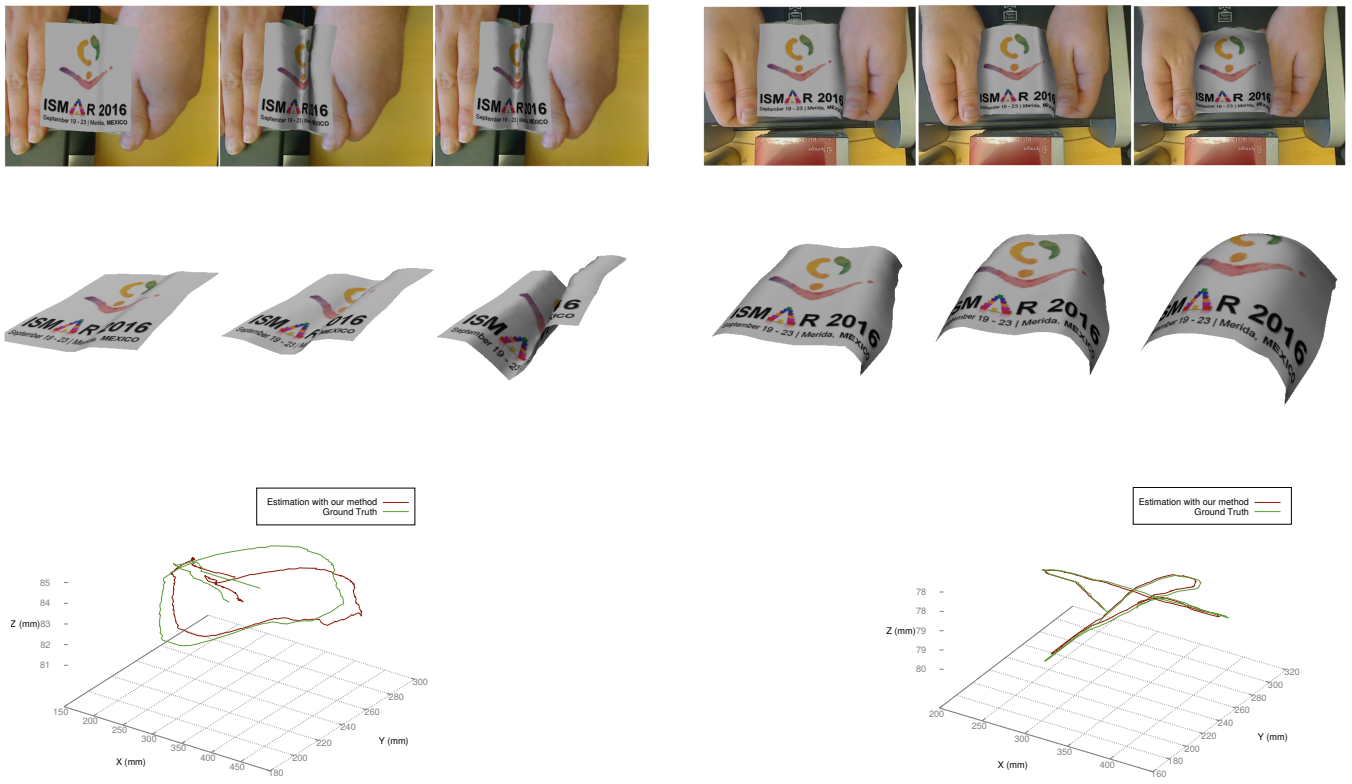


Figure 4: Results on real data: [Top row] input videos with the augmented textures, [Middle row] the resulting 3D shape and [Bottom row] the pose estimation. (Left) elasticity = 30% RMS error = 1.95 mm — (Right) elasticity = 40%, RMS error = 2.1 mm.

to estimate the 3D elastic shape. Promising results were obtained through synthetic and real data, where small absolute camera pose errors were obtained and shape recovery errors close or below 2 mm were reported. Future works will extend the experiments to the comparison with related works and tests on surgical data where such approaches can be of high benefits.

REFERENCES

[1] A. Agudo, L. Agapito, B. Calvo, and J. M. M. Montiel. Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In *Proceedings of 2014 IEEE Conference on CVPR*, pages 1558–1565, June 2014.

[2] A. Agudo, B. Calvo, and J. M. M. Montiel. Finite element based sequential bayesian non-rigid structure from motion. In *Proceedings of 2012 IEEE Conference on CVPR*, pages 1418–1425, 2012.

[3] A. Agudo and F. Moreno-Noguer. Simultaneous pose and non-rigid shape with particle dynamics. In *Proceedings of 2014 IEEE Conference on CVPR*, pages 2179–2187, 2015.

[4] A. Bartoli, Y. Grard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2099–2118, Oct 2015.

[5] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *Proceedings of 2005 IEEE Conference on CVPR*, volume 2, pages 122–128 vol. 2, June 2005.

[6] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of 2013 IEEE Conference on CVPR*, pages 1272–1279, June 2013.

[7] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Proceedings of 2011 IEEE Conference on CVPR*, pages 3065–3072, June 2011.

[8] N. Haouchine, J. Dequidt, M.-O. Berger, and S. Cotin. Single view augmentation of elastic objects. In *Proceedings of 2014 IEEE ISMAR*, pages 199–208.

[9] R. Kikuuwe, H. Tabuchi, and M. Yamamoto. An edge-based computationally efficient formulation of saint venant-kirchhoff tetrahedral finite elements. *ACM Transactions on Graphics (TOG)*, 28(1):8, 2009.

[10] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81(2):155–166, 2008.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[12] A. Malti, A. Bartoli, and R. Hartley. A linear least-squares solution to elastic shape-from-template. In *Proceedings of 2015 IEEE Conference on CVPR*, pages 1629–1637, 2015.

[13] A. Malti, R. Hartley, A. Bartoli, and J.-H. Kim. Monocular template-based 3d reconstruction of extensible surfaces with local linear elasticity. In *Proceedings of 2013 IEEE Conference on CVPR*, pages 1522–1529, 2013.

[14] M. Müller and M. Gross. Interactive virtual materials. In *Proceedings of Graphics Interface 2004*, GI '04, pages 239–246, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

[15] A. Nealen, M. Mller, R. Keiser, E. Boxerman, and M. Carlson. Physically based deformable models in computer graphics. *Computer Graphics Forum*, 25(4):809–836, 2006.

[16] M. Nesme, Y. Payan, and F. Faure. Efficient, physically plausible finite elements. In J. Dingliana and F. Ganovelli, editors, *Eurographics 2005, Short papers*, Trinity College, Dublin, Irlande, 2005.

[17] M. Paladini, A. Bartoli, and L. Agapito. *Sequential Non-Rigid Structure-from-Motion with the 3D-Implicit Low-Rank Shape Model*, pages 15–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.