Direct, Dense, and Deformable: Template-Based Non-Rigid 3D Reconstruction from RGB Video

Rui Yu Chris Russell
University College London

Neill D. F. Campbell
University of Bath

Lourdes Agapito
University College London

http://visual.cs.ucl.ac.uk/pubs/ddd/

Abstract

In this paper we tackle the problem of capturing the dense, detailed 3D geometry of generic, complex non-rigid meshes using a single RGB-only commodity video camera and a direct approach. While robust and even real-time solutions exist to this problem if the observed scene is static, for non-rigid dense shape capture current systems are typically restricted to the use of complex multi-camera rigs, take advantage of the additional depth channel available in RGB-D cameras, or deal with specific shapes such as faces or planar surfaces. In contrast, our method makes use of a single RGB video as input; it can capture the deformations of generic shapes; and the depth estimation is dense, per-pixel and direct. We first compute a dense 3D template of the shape of the object, using a short rigid sequence, and subsequently perform online reconstruction of the non-rigid mesh as it evolves over time. Our energy optimization approach minimizes a robust photometric cost that simultaneously estimates the temporal correspondences and 3D deformations with respect to the template mesh. In our experimental evaluation we show a range of qualitative results on novel datasets; we compare against an existing method that requires multi-frame optical flow; and perform a quantitative evaluation against other template-based approaches on a ground truth dataset.

1. Introduction

The recent emergence of low cost depth sensors, has brought easy and fast acquisition of 3D geometry closer to reality. Systems such as KinectFusion [20] allow users to scan the detailed 3D shape of rigid scenes. The use of RGB-D sensors has also been extended to markerless capture of non-rigid shapes [14, 15] even in real time [18, 37]. At the same time, many multi-camera techniques for marker-less high-end dynamic 3D shape acquisition have been developed over the last decade [8, 33].

In contrast, the acquisition of dense 3D models of generic deformable meshes from a monocular *RGB-only* video stream is significantly harder. The ability to acquire time-varying dense shapes from monocular RGB video would open the door to easy, lightweight non-rigid capture and, perhaps more importantly, from existing video footage or web-based video libraries such as YouTube.

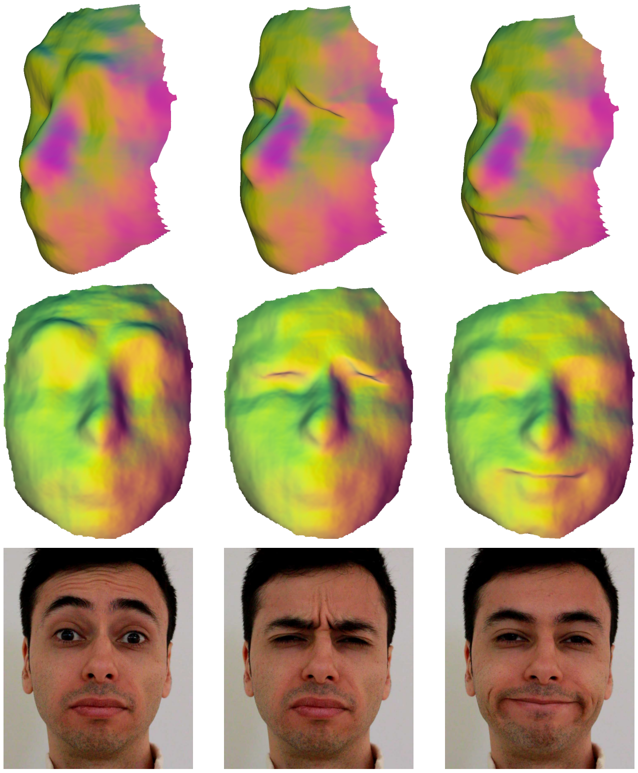

Figure 1: An automatically generated template is warped (top two rows) in a physically plausible manner consistent with a video sequence (bottom) generating rich dynamic 3D meshes, that capture emotive deformations of the mouth and eyes. Each column corresponds to different views of the same frame.

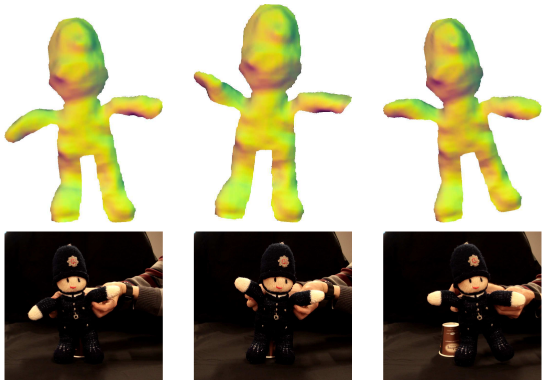

Figure 2: Direct deformable reconstruction from our algorithm on the *bobby* sequence. See section 6.

While spectacular progress has been made in monocular dense 3D reconstruction of static scenes from video [19, 28, 13, 31], direct 3D capture of dense non-rigid shapes from a single video stream lies significantly behind.

Three main successful directions dominate the literature for monocular 3D reconstruction of deformable surfaces. *Model-based* methods [3] use blend-shape models learned from 3D training data in an off-line training step. *Non-rigid structure from motion* (NRS*f*M) approaches offer a model-free formulation for generic shapes but require long term correspondences across a video sequence [4, 7, 10, 22, 23, 32] and are typically batch methods that process the entire sequence at once. Finally, *shape-from-template* approaches [2, 21, 24, 26] offer an attractive sequential frame-to-frame solution but they require a known 3D reference template of the surface and point correspondences between each new frame and the template as input. In addition they have mostly been demonstrated only on simple planar meshes of objects such as paper and cloth.

Two common limitations remain with most NRS*f*M and *shape-from-template* formulations: *(i)* they are typically feature-based which leads to sparse reconstructions or failure with low-textured surfaces and *(ii)* estimation of 2D correspondences and 3D shape inference are decoupled and not solved simultaneously in a direct approach. So far the problem of jointly estimating dense point correspondences and non-rigid 3D geometry from monocular video has received very little attention. Garg *et al*. demonstrated a dense per-pixel NRS*f*M approach but it required dense 2D correspondences to be pre-computed using a multi-frame optical flow method. Pixel-based approaches to template-based reconstruction have been proposed by Malti *et al*. [16] and Suwajanakorn *et al*. [30] but they were only demonstrated on planar surfaces (cloth or paper) [16] or worked exclusively for faces [30].

In this paper we adopt a template-based direct approach to deformable shape reconstruction from monocular sequences. Our contribution is an end-to-end system that builds a dense template from an initial rigid subsequence and subsequently estimates the deformations of the mesh with respect to the 3D template by minimizing a robust photometric cost. Unlike previous template-based direct methods [16, 30] we demonstrate our approach on a variety of generic complex non-planar meshes. While our algorithm is not real-time, it is sequential and relatively fast, typically requiring 3 seconds per frame on a standard desktop machine to optimize a mesh with approximately 25,000 vertices. Ours is the only template-based approach that satisfies all the properties listed in Table 1.

	Zollhofer *et al.* [37]	Malti *et al.* [16]	Suwajanakorn *et al.* [30]	Garg *et al.* [10]	Ours
Template-free	✗	✗	✗	✓	✗
Direct	✓	✓	✓	✗	✓
RGB-only	✗	✓	✓	✓	✓
Monocular	✗	✓	✓	✓	✓
Perspective camera	✓	✓	✗	✗	✓
Frame-to-frame	✓	✓	✓	✗	✓
Generic shapes	✓	✗	✗	✓	✓
Closed mesh with self-occlusion handling	✓	✗	✗	✗	✓

Table 1: Comparison of our approach with other dense competitors for reconstructing deformable shapes. Ours is the only template-based dense approach that only uses monocular RGB data; is frame-to-frame; direct; and suitable for reconstructing generic shapes.

2. Related Work

Very few methods attempt dense and direct reconstruction of non-rigid shapes from monocular sequences. There are three areas of research that have inspired and influenced our work: *non-rigid structure from motion*, *shape-from-template* and *RGB-D based non-rigid capture*. We now describe the most related approaches from each of these fields.

The most related NRS*f*M method to ours is the dense monocular non-rigid reconstruction algorithm by Garg *et al*. [10]. Although their algorithm reconstructs dense per-pixel models, noticeably, it is a batch process that requires multi-frame optic flow over the entire sequence as an input. No attempt was made to solve the dense correspondence and reconstruction problems simultaneously. As such, if the

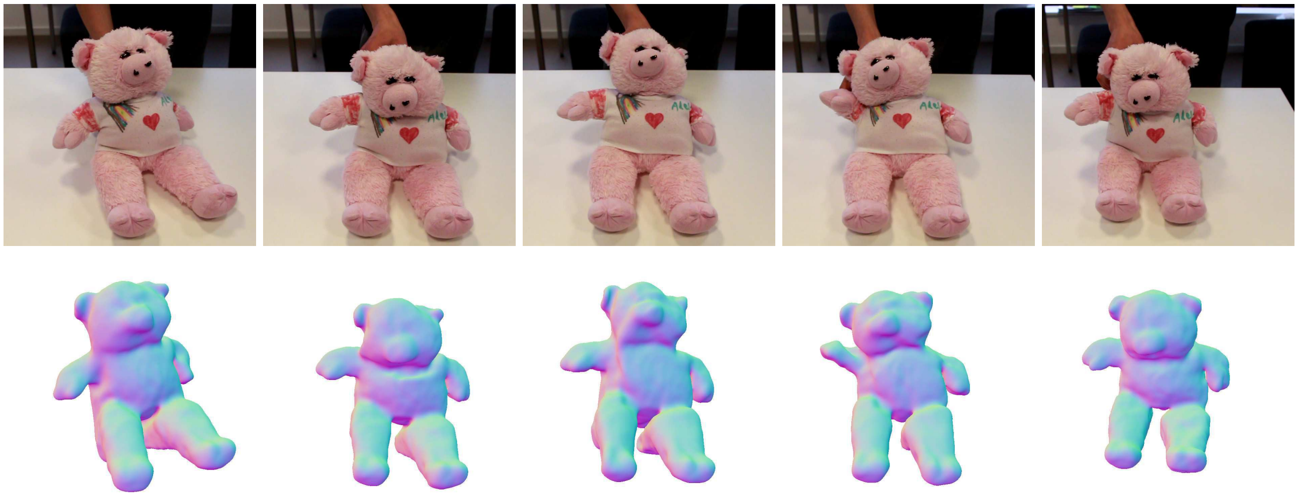

Figure 3: Direct deformable reconstruction from our algorithm on the pig sequence.

flow generation fails, a good reconstruction is not possible.

Our method also shares strong similarities with work in the area of **shape from template** [2, 24, 21, 26]. Many approaches have been proposed mostly taking advantage of the constraints imposed by isometric or conformal deformations [2, 17, 25]. While most template approaches are feature-based and only reconstruct based on a small number of points, Malti *et al.* [16] departs by proposing a direct pixel-based variational framework that exploits visibility constraints. However, their method was only demonstrated on flat isometric surfaces. The recent work of Suwajanakorn *et al.* [30] reconstructs RGB-only videos of faces of celebrities. Similarly to our method, they formulate template-based non-rigid reconstruction as a frame-to-frame energy minimization that optimizes a direct photometric cost. However, their method is limited to reconstructing human faces as their template reconstruction approach is specifically tailored to them. Also related is the monocular face capture system of Garrido *et al.* [11]. While their work also minimizes a photometric cost and the deformations with respect to a template model, theirs is a sophisticated blend-shape model specifically built to capture the deformations of human faces.

Our work has been largely inspired by recent advances in non-rigid tracking **using depth cameras** [18, 37]. Zollhofer *et al.*'s [37] is the most related approach since their setup is directly comparable to ours — a multi-scale template is built first from a rigid sub-sequence, followed by dense non-rigid monocular tracking. However, while their method uses both the depth and the RGB channels, ours only uses RGB images as input and can be seen as its RGB-only equivalent.

Table 1 summarizes our main contributions and the differences with respect to the four most closely related approaches. In summary, ours is the only RGB-only, template-based, monocular, dense and direct approach to non-rigid reconstruction that is sequential and suitable for generic shapes and closed meshes.

3. Problem Formulation

We consider a perspective RGB camera with known internal calibration observing a non-rigid mesh deforming over time. The goal of our algorithm is to estimate, at each time-step t, the current 3D coordinates of the N vertices of the dense non-rigid mesh $\boldsymbol{S}^t = [\dots \boldsymbol{s}_i^t \dots]$, $i = 1..N$, as well as the overall rigid rotation and translation $(\boldsymbol{R}^t, \boldsymbol{t}^t)$ that align the deformed shape and a reference 3D template.

The only inputs to our algorithm are the current RGB image $\boldsymbol{I}^t(x, y)$ observed at time t and a template shape $\widetilde{\mathbf{S}} = [\dots \widehat{\mathbf{s}}_i \dots]$, $i = 1..N$, which is acquired automatically in a preliminary template acquisition step using the multi-view stereo dense volumetric approach of [5]. Typically the user acquires a short rigid sequence to capture the 3D coordinates of the template mesh which is then subsampled to create a multi-resolution hierarchy of coarse-to-fine templates. The template acquisition step is described in detail in section 4. The template is converted to a quadrilateral mesh, consisting of N vertices and M edges.

Once the template has been acquired, our system turns to perform frame-to-frame non-rigid alignment of the 3D shape given only the current frame as input. Although optimization is initialized using the shape from the previous frame $\boldsymbol{S}^{t-1}$, once the template has been generated, the optimization objective does not depend on any other frames. As such, unlike most approaches to non-rigid structure from motion [7, 10, 22, 32], it scales to the streaming of long sequences, with the complexity of optimization guaranteed to

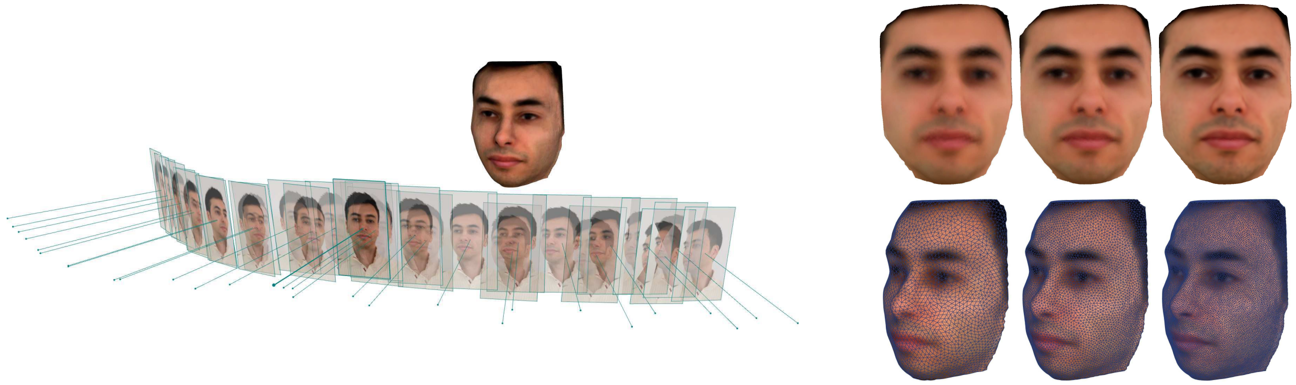

Figure 4: Template acquisition step. **Left:** A volumetric representation is generated from stereo depth maps taken over a rigid subsequence. This is then transformed into a colored mesh. **Right:** The three scales of the template used to robustly estimate deformations.

grow linearly to the number of frames.

4. Step 1: Template Shape Acquisition

The first stage in our process is to obtain a rigid template mesh of the shape. We denote the whole shape as a $3 \times N$ matrix $\widehat{\mathbf{S}}$, and $\widehat{\mathbf{s}}_i$ as the i^{th} vertex on the mesh. We require a set of M images (we used $M \sim 30$) of the shape under a rigid transformation. These are obtained by subsampling a set of frames from a short video where either the object is static and the camera moves or the camera is static and the object is moved under a rigid transformation. Figure 4 provides an example of the output of this process.

The process of the template acquisition is an application of existing multi-view stereo (MVS) techniques; consequently we provide only an overview of the process with appropriate references to the methods used. A more detailed description of the process may be found in the supplementary material.

Extrinsic Calibration The collection of frames from the video were calibrated automatically using an implementation (VisualSFM [36]) of standard rigid structure-from-motion (S*f*M). This was observed to be robust to any incompatible motion in the background. If there is too much background clutter in the image then an automatic segmentation of the foreground can be attempted using a fixation condition (that the center of the image fixates on the object of interest) [6].

Depth-Map Extraction Once we have a calibrated set of frames, we extract a depth-map using the stereo method of [5]. For each (reference) image, we take the two closest viewpoints as neighboring images and extract the best $K = 9$ normalized cross-correlation (NCC) scores matching with 13×13 pixel windows. These are then filtered to provide a single depth estimate (or unknown label) using the default filtering parameters as specified in [5].

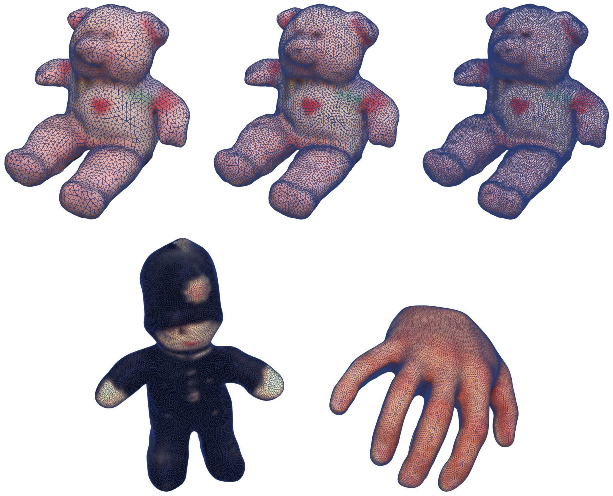

Figure 5: An example of our multi-scale template meshes generated by iterative mesh down-sampling and refinement. **Top:** From left to right the meshes contain approximately 5, 10, and 25 thousand vertices respectively. **Bottom:** The highest levels of the templates for the bobby and hand sequences.

Mesh Estimation The last stage is to extract the template mesh by combining all the individual depth-maps in a single global optimization. As suggested in [5], we combine the depth-maps to recover a single watertight mesh $\widetilde{\mathbf{S}}$ using the volumetric fusion technique of [35] combined with the probabilistic visibility approach of [12].

Template Hierarchy The output of the fusion stage is a watertight mesh $\widetilde{\mathbf{S}}$. From this we build a multi-scale representation of the mesh as shown in Figure 4 (right). This is achieved by iteratively down-sampling and refining the template mesh using the isotropic surface remeshing method

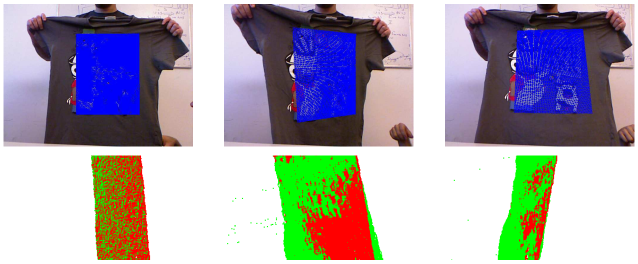

Figure 6: Reconstruction results on T-shirt sequence. Top: input sequence with our tracking. Bottom: Our reconstruction (red) overlayed by ground truth Kinect mesh (green). The scattering of green points visible in the last two columns indicates sparse failures by the Kinect. See table 2 for a quantitative comparison with other approaches.

(and implementation) of Fuhrmann *et al.* [9]. Finally, a color $\widehat{\mathbf{I}}_i$ is associated to each vertex i; this is the median color over all the frames in the rigid subsequence in which the projected vertex is visible.

To avoid aliasing when coloring the low resolution meshes, we blur each of the input images with a length-scale given by the median mesh edge length projected into the corresponding camera view. Figure 5 shows a triangulated example of the multi-scale colored mesh representation.

5. Step 2: Non-Rigid Model Tracking

5.1. Our Energy

Our objective is made of a balanced combination of four terms: *(i)* a *photometric error* which captures the expected color of each *visible* vertex in the template; *(ii)* a *total variation term* on the gradient of the 3D displacements with respect to the template *(iii) as rigid as possible* local regularization – this term allows the mesh to rotate locally without imposing a penalty and *(iv)* a *temporal smoothness* term that penalizes strong frame-to-frame deformations.

The per-frame objective takes the form:

$$\begin{aligned} E(\boldsymbol{S},\boldsymbol{R},\boldsymbol{t}) = {} & E_{\text{data}}(\boldsymbol{S},\boldsymbol{R},\boldsymbol{t}) + \lambda_r E_{\text{reg}}(\boldsymbol{S}) \\ & + \lambda_a E_{\text{arap}}(\boldsymbol{S}) + \lambda_t E_{\text{temp}}(\boldsymbol{S}). \end{aligned} \tag{1}$$

where λ_r, λ_a and λ_t denote the relative weights between the terms. These terms are all required. The first term guarantees that the deformations of the template follow the image; the second term encourages locally smooth deformations while allowing sharp discontinuities which are needed to transition from parts of the object that deform strongly to those that do not; while the third term encourages stronger changes in template depth. Finally, temporal smoothness is needed to avoid flickering.

	error
PCA [34]	18.44
Unconstrained [34]	15.50±1.78
Varol*et al.* [34]	14.79±0.90
Our method	7.05

Table 2: Evaluation results on the T-shirt sequence of [34].

For simplicity's sake, we drop temporal super-scripts where appropriate as much of the formulation does not depend on any other frames. We now define each of the terms of the energy in detail.

5.1.1 Photometric Data Term E_{data}

The **data term** E_{data} encourages a shape such that projection of the vertices into the current image has similar appearance to the template shape. In other words, minimization of this photometric cost encourages brightness constancy with respect to the colors $\widehat{\mathbf{I}} = \{\widehat{\mathbf{I}}_i\}$ of the mesh, built by projecting the images used to build the reference template $\widehat{\mathbf{S}} = \{\widehat{\mathbf{s}}_\mathbf{i}\}$ onto the vertices of the template. As we directly reconstruct closed meshes where much of the object is self-occluded, we first make an initial pass where we estimate the visibility of each vertex in the mesh. For additional robustness, we use a Huber loss.

$$E_{\text{data}}(\boldsymbol{S},\boldsymbol{R},\boldsymbol{t}) = \sum_{i \in \mathcal{V}} |\widehat{\mathbf{I}}_i - \boldsymbol{I}(\pi(\boldsymbol{R}(\boldsymbol{s}_i) + \boldsymbol{t}))|_\epsilon \tag{2}$$

where $\widehat{\mathbf{I}}_i$ is the color of vertex $\widehat{\mathbf{s}}_i$ on the template mesh, $\boldsymbol{I}$ is the current image frame, $\mathcal{V}$ is the set of estimated visible vertices in the frame[1], $\{\widehat{\mathbf{s}}_i\}_1^N$ are the 3D vertices of the template, $\{\boldsymbol{s}_i\}_1^N$ are the 3D vertices of the shape in the current frame, $\pi(\cdot)$ is again the projection from 3D points to image coordinates, known from camera calibration, and $|\cdot|_\epsilon$ denotes the Huber loss, which takes the form

$$|x|_\epsilon = \begin{cases} x^2/(2\epsilon) & \text{if } x^2 \leq \epsilon \\ |x| - \epsilon/2 & \text{otherwise.} \end{cases} \tag{3}$$

5.1.2 Spatial Regularization Term E_{reg}

The **regularization term** E_{reg} is a pairwise term that encourages spatially smooth deformations of the shape $\boldsymbol{S}$ with respect to the template $\widehat{\mathbf{S}}$.

$$E_{\text{reg}}(\boldsymbol{S}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|(\boldsymbol{s}_i - \boldsymbol{s}_j) - (\widehat{\mathbf{s}}_i - \widehat{\mathbf{s}}_j)\|_\epsilon \tag{4}$$

here $\mathcal{N}_i$ is the neighborhood of i, and $\|\cdot\|_\epsilon$ is the vector analog of the Huber loss formed by summing the standard Huber loss over all dimensions.

[1] This is generated by realigning the deformed mesh of the previous frame to minimize photometric error (see section 5.2.1), and z-buffering.

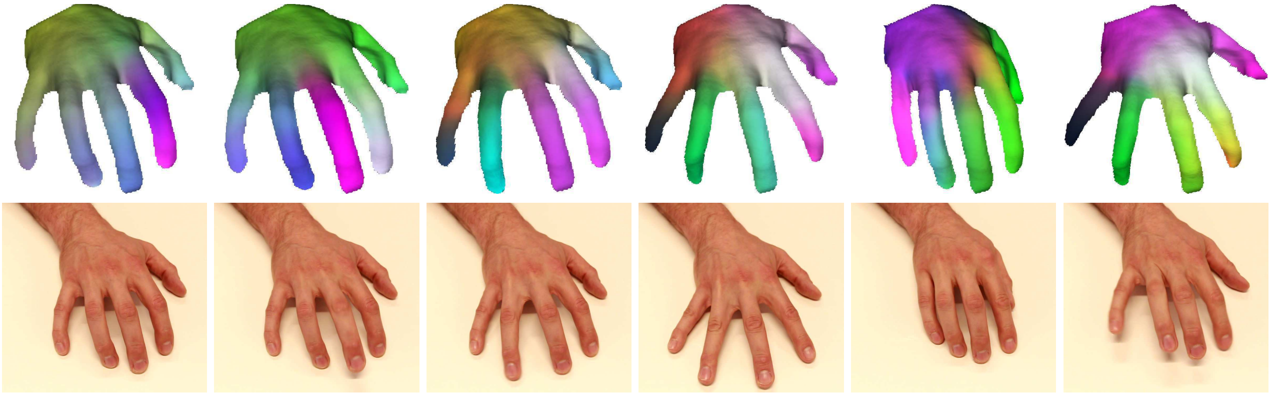

Figure 7: As a corollary of reliable tracking and modeling of deformations, our method generates 3D optical flow that can be used to track the movement of dense points through out the reconstructed sequence. The images above visualize per-pixel 3d flow over a 1,200 frame sequence, where changes in intensity in RGB correspond to movement from the original frame in x,y, and z. The flow strongly distinguishes which finger is tapping (second, and third columns), and reflects movement of the entire hand.

5.1.3 As Rigid as Possible Deformation Term E_{arap}

This cost was first proposed in [27] to allow deformable tracking of an initial mesh against a depth map. It takes the form

$$E_{\text{arap}}(\boldsymbol{S}, \{\boldsymbol{A}_i\}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|(\boldsymbol{s}_i - \boldsymbol{s}_j) - \boldsymbol{A}_i(\widehat{\mathbf{s}}_i - \widehat{\mathbf{s}}_j)\|_2^2 \quad (5)$$

where the variables A_i are per-point local rotations. Essentially this cost allows for local rotations to take place in the mesh without penalty so long as the relative locations between points in the neighborhood of i remain constant. It can be interpreted as a prior that allows for elastic style deformations of meshes.

5.1.4 Temporal Smoothness E_{temp}

The temporal regularization encourages smooth deformations from frame to frame and can be formulated as

$$E_{\text{temp}}(\boldsymbol{S}, \boldsymbol{t}) = \|\boldsymbol{S} - \boldsymbol{S}^{t-1}\|_{\mathcal{F}}^2 + \|\boldsymbol{t} - \boldsymbol{t}^{t-1}\|_2^2 \quad (6)$$

where $\boldsymbol{S}^{t-1}$ and $\boldsymbol{t}^{t-1}$ are the shape and the translation in the previous frame and $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm of a matrix. The need for this term is most apparent when viewing a video of the reconstruction. Although a small amount of temporal regularization only alters the shape a little, it substantially reduces frame-to-frame flickering, while the temporal smoothness in the translation prevents explaining deformations as perspective effects.

5.2. Energy Optimization

For reasons of robustness and efficiency, optimization is performed in a two step form over rotations and translations, and shape separately, and using a 3-layer spatial pyramid.

5.2.1 Initialization

We optimize this objective in a two step form: first the rotations and translation are estimated using the shape from the previous frame.

$$E(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{N} |\widehat{\mathbf{I}}_i - \boldsymbol{I}(\pi(\boldsymbol{R}(\boldsymbol{s}_i^{t-1}) + \boldsymbol{t}))|_\epsilon \quad (7)$$

Then, holding the global rotations and translations constant, $\boldsymbol{s}^t$ is estimated. $\boldsymbol{R}$, $\boldsymbol{t}$, and at the coarsest level of the pyramid, $\boldsymbol{S}^t$ are initialized using their solution taken from the previous frame, and we perform optimization using a conjugate gradient based solver taken from Ceres [1].

5.2.2 Coarse-to-fine optimization

Both the rotation and translation cost (7), and the shape cost (1) are optimized over a set of 3-level coarse-to-fine images and shape templates, with each layer of the pyramid being a factor of two larger than the coarser layer directly above it. As we move down the pyramid from coarse to fine, the 3D vertices are propagated to the next level of the hierarchy using a prolongation step as described in Sumner *et al.* [29]. The weights are precomputed when the template mesh is created. Prolongation was also applied to the ARAP rotations $\{\boldsymbol{A}_i\}$ using the method described in Zollhofer *et al.* [37].

6. Experimental Results

In this section we show qualitative results of our method on a variety of non-planar 3D meshes; a qualitative comparison with the dense NRS*f*M method of Garg *et al.* [10]; and

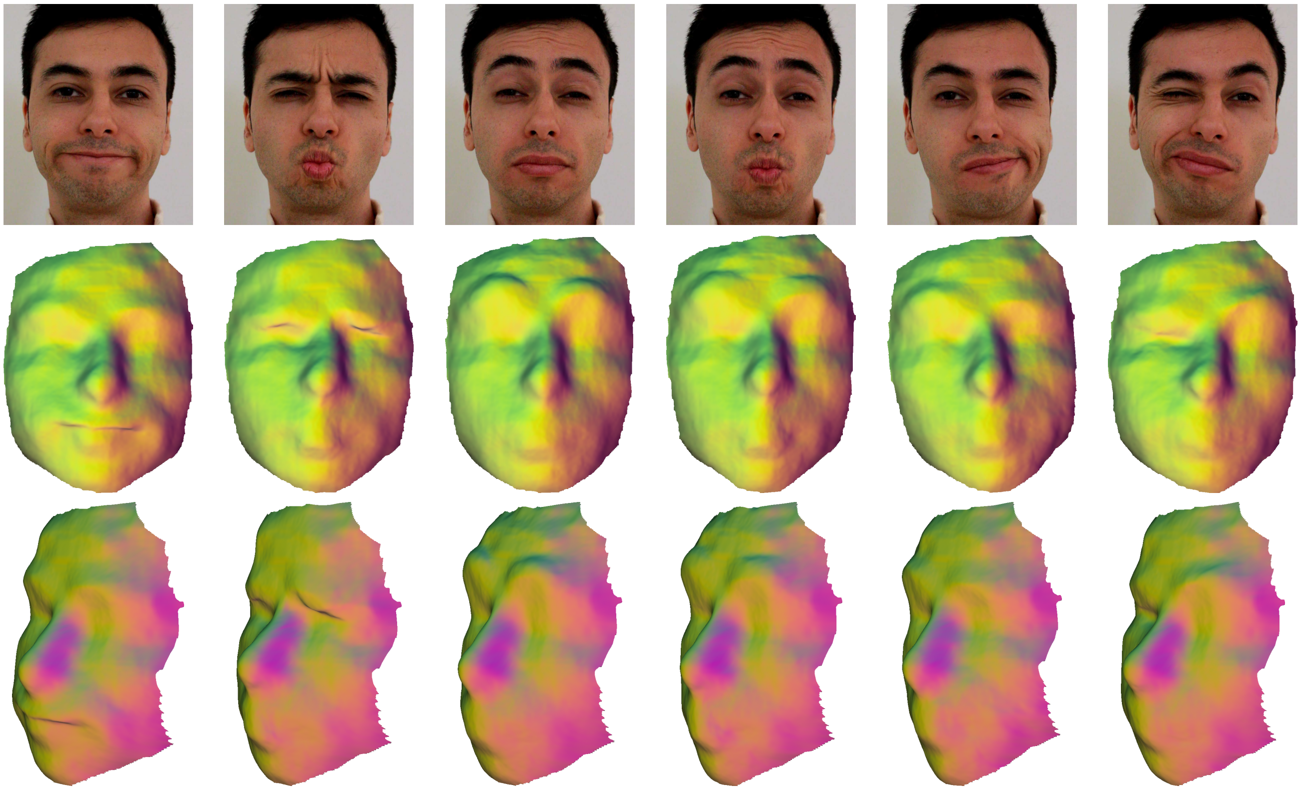

Figure 9: Direct reconstruction from our algorithm on a new face sequence.

we provide a quantitative evaluation on the *t-shirt* ground truth dataset from CVLab [34]. Our results can be best viewed in the accompanying video.[2]

Qualitative results on generic non-planar meshes We show results on some new sequences acquired with a hand-held camera. We show sequences of a *face* (Figure 1) and two soft toys – a *bobby* (Figure 2) and a *pig* (Figure 3). These sequences show a wide range of deformations of a varying set of shapes, with different degrees of elasticity. The reconstructions and deformations generated are convincing. To demonstrate our robustness to occlusions we manually augmented a sequence showing the tracking of a human hand with an artificial occluder (see Figure 10).

Qualitative comparison with Garg *et al.* [10] We show a direct qualitative comparison of the 3D deformable tracking (i.e. excluding the template generation) obtained using our approach with the reconstruction obtained with the dense NRS*f*M method of Garg *et al.* [10]. Our method, as presented, requires both known internal camera parameters and a static subsequence in which to generate the rigid template. Unfortunately neither of these are available for this face sequence. For this one sequence only, we follow [10] in assuming the camera is approximately orthographic, and generate a template by performing rigid S*f*M with bundle adjustment over the multi-frame optic flow used as input and made publicly available by Garg *et al.* [10].

The results can be seen in Figure 8. The advantage in not requiring optic flow as an input can be seen in the improved sharpness of the ear on the left and the smoother skin texture of our results. While the flow parameters must be chosen as a compromise between capturing the sharp movement of the ear, and dampening the flow fluctuations that appear in the face, our model based tracking has no such problems[3]. The other notable difference between our approach and theirs is that our computation is frame to frame instead of batch. Therefore, we naturally handle much longer sequences of thousands of frames while [10] does not scale well above 100 frames, with the dense optic flow method needed as in-

[2] Please see http://visual.cs.ucl.ac.uk/pubs/ddd/ for supplementary material and video.

[3] We speculate that owing to the implicit averaging in taking many frames to generate the rigid template, the template itself shows no such issues.

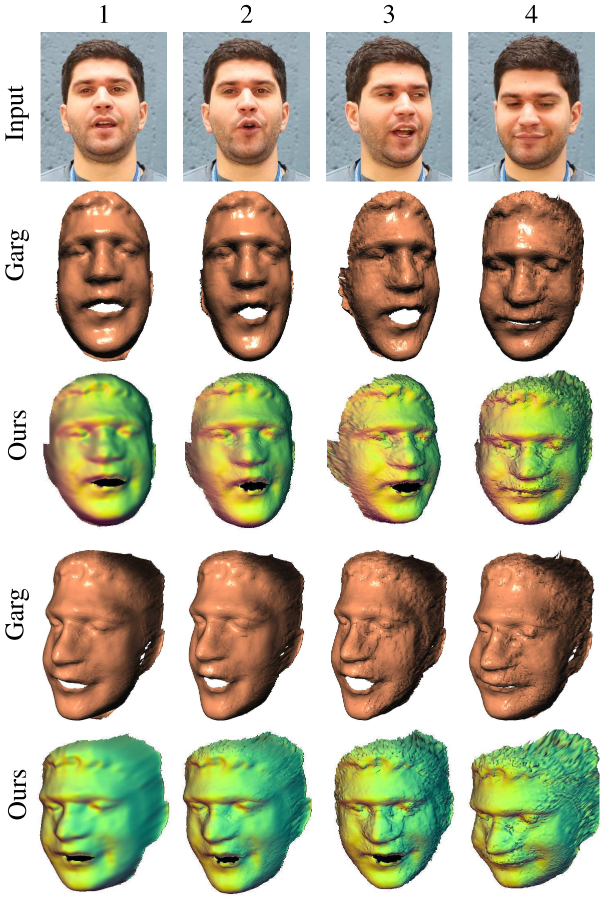

Figure 8: A comparison of our direct model tracking against the flow based reconstruction of Garg *et al.* [10]. In comparison our approach offers sharper reconstruction of the left ear, while most of unnatural skin bumps from columns 3 onwards are less pronounced in our results, which also demonstrate a somewhat wider range of mouth articulations. The apparent spike on the back of the head in columns 3 and 4 arises from a deep template reconstruction capturing the back of the head that unfortunately only extends on one side.

put as the main bottleneck. Despite being a direct approach, the quality of the reconstructions provided by our method is comparable to Garg *et al.* and in some situations arguably better (see Figure 8).

Quantitative evaluation on the t-shirt dataset [34] As a final experiment, we conduct a quantitative comparison against the template-based method of [34], obtaining less than half the numeric error of their approach (see Table 2, and Figure 6). The evaluation is carried out on a sequence of a t-shirt for which ground truth data was collected using a Microsoft Kinect RGB-D camera and made publicly available by [34].

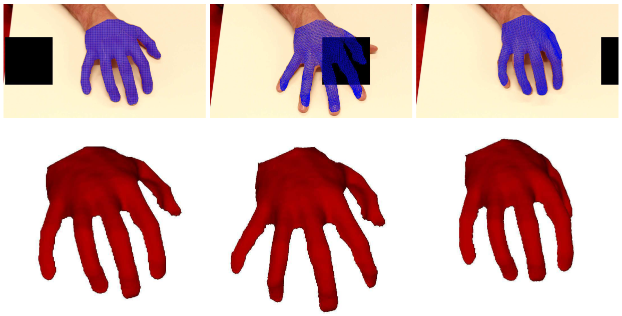

Figure 10: Occlusion experiment on hand sequence. Here it can be seen that the hand motion is reliably estimated despite the presence of a large occluder.

7. Conclusion

We have presented a novel approach to template driven capture of dense detailed non-rigid deformations from video sequences. Our method solves simultaneously the 2D dense registration problem and the 3D shape inference using RGB-video and a pre-acquired template as only input. An additional advantage is that our approach is sequential in nature and can therefore be applied to arbitrarily long sequences. Unlike many other template based methods, our approach can deform complex generic meshes and is not restricted to planar surfaces. We have shown results on real world novel video sequences captured with a hand-held camera which demonstrate the validity of our approach; we compare against an existing method that requires multi-frame optical flow with comparable results; and perform a quantitative evaluation against other template-based approaches on a ground truth dataset where our approach halves the 3D error of competing approaches.

8. Acknowledgments

This work has been partly supported by the Second-Hands project, funded from the European Unions Horizon 2020 Research and Innovation programme under grant agreement No 643950. CR has been funded by a UCL/BBC research fellowship.

References

[1] S. Agarwal, K. Mierle, et al. Ceres solver. http://ceres-solver.org. 6

[2] A. Bartoli, Y. Gerard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *CVPR*, 2012. 2, 3

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000. 2

[5] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, 2008. 3, 4

[6] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Automatic object segmentation from calibrated images. In *CVMP*, 2011. 4

[7] Y. Dai, H. Li, and M. He. A simple prior-free method for non rigid structure from motion factorization. In *CVPR*, 2012. 2, 3

[8] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH*, 2008. 1

[9] S. Fuhrmann, J. Ackermann, T. Kalbe, and M. Goesele. Direct resampling for isotropic surface remeshing. In *Proceedings of Vision, Modeling and Visualization 2010, Siegen, Germany*, 2010. 5

[10] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *CVPR*, 2013. 2, 3, 6, 7, 8

[11] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *SIGGRAPH Asia*, 2013. 3

[12] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007. 4

[13] K. Kolev, P. Tanksanen, P. Speciale, and M. Pollefeys. Turning mobile phones into 3d scanners. In *CVPR*, 2014. 2

[14] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *SIGGRAPH Asia*, 2009. 1

[15] C. Malleson, M. Klaudiny, A. Hilton, and J.-Y. Guillemaut. Single-view rgb-d based reconstruction of dynamic human geometry. In *4DMOD Workshop at ICCV*, 2013. 1

[16] A. Malti, A. Bartoli, and T. Collins. A pixel-based approach to template-based monocular 3d reconstruction of deformable surfaces. In *4DMOD Workshop at ICCV*, 2011. 2, 3

[17] A. Malti, R. Hartley, A. Bartoli, and J.-H. Kim. Monocular template-based 3d reconstruction of extensible surfaces with local linear elasticity. In *CVPR*, 2013. 3

[18] R. Newcombe, D. Fox, and S. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 1, 3

[19] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *ICCV*, 2011. 2

[20] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 1

[21] J. Östlund, A. Varol, D. T. Ngo, and P. Fua. Laplacian meshes for monocular 3d shape recovery. In *Computer Vision–ECCV 2012*, pages 412–425. Springer, 2012. 2, 3

[22] M. Paladini, A. Del Bue, J. Xavier, L. Agapito, M. Stosic, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *IJCV*, 2012. 2, 3

[23] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *CVPR*, 2011. 2

[24] M. Salzmann, R. Hartley, and P. Fua. Convex optimization for deformable surface 3-d tracking. In *CVPR*, 2007. 2, 3

[25] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-Form Solution to Non-Rigid 3D Surface Registration. In *ECCV*, 2008. 3

[26] M. Salzmann, R. Urtasun, and P. Fua. Local deformation models for monocular 3d shape recovery. In *CVPR*, 2008. 2, 3

[27] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP '07, 2007. 6

[28] J. Stuehmer, S. Gumhold, and D. Cremers. Real-time dense geometry from a handheld camera. In *Pattern Recognition (Proc. DAGM)*, pages 11–20, September 2010. 2

[29] R. W. Sumner, J. Schmid., and M. Pauly. Embedded deformation for shape manipulation. In *SIGGRAPH*, 2007. 6

[30] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, 2014. 2, 3

[31] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, 2013. 2

[32] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 2008. 2, 3

[33] L. Valgaerts, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *SIGGRAPH*, 2013. 1

[34] A. Varol, P. F. Mathieu Salzmann, and R. Urtasun. A constrained latent variable model. In *CVPR*, 2012. 5, 7, 8

[35] G. Vogiatzis, C. Hernández, P. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 2007. 4

[36] C. Wu. Visualsfm: A visual structure from motion system. `http://ccwu.me/vsfm/`, 2011. 4

[37] M. Zollhofer, M. Niessner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *SIGGRAPH*, 2014. 1, 2, 3, 6