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

Rui Yu, Chris Russell, Neill D. F. Campbell and Lourdes Agapito

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; and perform a quantitative evaluation on a ground truth dataset.

Index Terms—Dense, Direct, Deformable, Monocular 3D Reconstruction, RGB Video

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 [23] 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 [16], [18] even in real time [21], [42]. At the same time, many multi-camera techniques for marker-less high-end dynamic 3D shape acquisition have been developed over the last decade [9], [38].

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.

Substantial progress has been made in dense 3D reconstruction of rigid shapes or static scenes purely from *RGB* video sequences or image collections, which is now considered a highly mature field. One of the distinguishing features of most dense methods is that they are *direct* approaches in that they simultaneously solve for the dense 2D correspondences and the 3D geometry by minimizing a photometric cost. Multiview stereo systems exist that can recover the dense 3D geometry of rigid meshes accurately from a set of fully calibrated images [31]. Even real-time, live dense 3D reconstruction of static scenes is now possible using a single RGB camera and commodity hardware [22], [33] or even a mobile phone [15], [36]. While spectacular progress has been

- *Rui Yu, Chris Russell and Lourdes Agapito are with the Department of Computer Science, University College London. Neill D. F. Campbell is with Department of Computer Science, University of Bath.*
- *E-mail: r.yu@cs.ucl.ac.uk*
  E-mail: c.russell@cs.ucl.ac.uk
  E-mail: n.campbell@bath.ac.uk
  E-mail: l.agapito@cs.ucl.ac.uk

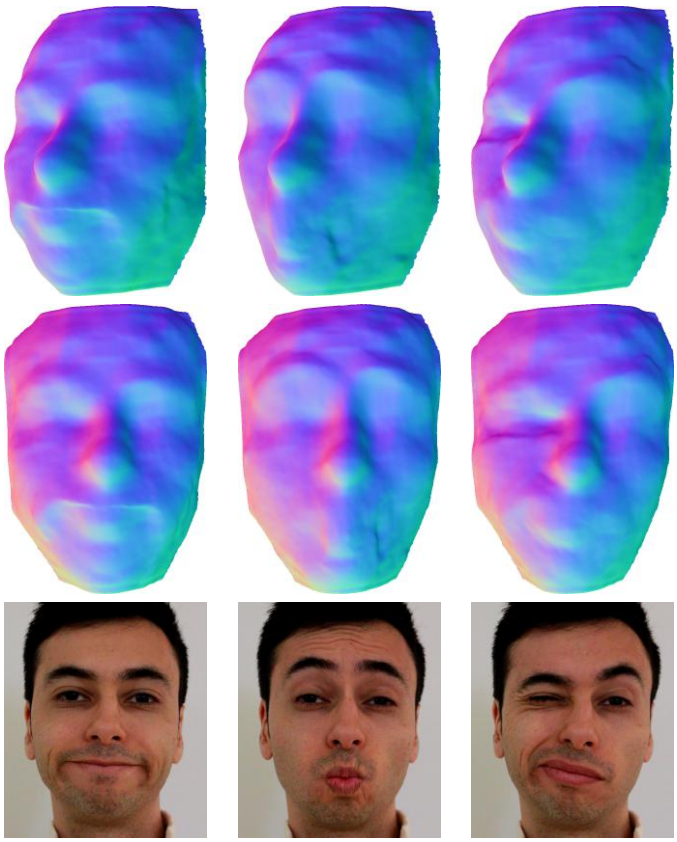

Fig. 1: An automatically generated template is warped (top two rows) in a physically plausible manner consistent with a video sequence (bottom) generating rich dynamic 3D meshes, that capture emotive deformations of the mouth and eyes. Each column corresponds to different views of the same frame.

made in monocular dense 3D reconstruction of static scenes from video [22], [33], [15], [36], 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 [4] 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 [5], [8], [12], [25], [27], [37] and are typically batch methods that process the entire sequence at once. Finally, *shape-from-template* approaches [2], [24], [28], [30] 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.* [12] 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.* [19] and Suwajanakorn *et al.* [35] but they were only demonstrated on planar surfaces (cloth or paper) [19] or worked exclusively for faces [35].

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 [19], [35] 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.* [42]	Malti *et al.*[19]	Suwajanakorn *et al.* [35]	Garg *et al.* [12]	Newcombe *et al.* [21]	Dou *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.

Although clearly inspired by the advances in **non-rigid structure from motion** methods [8], [25], [37], which can typically reconstruct non-rigid surfaces of generic shapes from monocular video while learning a low rank model that explains the deformations, our approach departs from them significantly. In particular, NRS*f*M formulations are batch and require (usually a small number of) point correspondences to be given as input. In contrast, the distinguishing features of our approach are that it is direct, dense, and frame-to-frame.

The most related NRS*f*M method to ours is the dense monocular non-rigid reconstruction algorithm by Garg *et al.* [12]. 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 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], [28], [24], [30]. Many approaches have been proposed mostly taking advantage of the constraints imposed by isometric or conformal deformations [2], [20], [29]. While most template approaches are feature-based and only reconstruct based on a small number of points, Malti *et al.* [19] 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.* [35] 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. In contrast, our template reconstruction step uses a dense volumetric multiview stereo formulation that is generic and can be used for any type of shape. In addition, our energy makes use of robust norms for the data and regularization terms; explores more sophisticated smoothness priors, such as local rigidity (as-rigid-as-possible [32]); and imposes temporal smoothness. Also related is the monocular face capture system of Garrido *et al.* [13]. 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** [21], [42]. Zollhofer *et al.*'s [42] 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. More recently, DynamicFusion [21] takes only the depth point cloud from a Kinect as input and estimates a warp back into a canonical reference scene, where a model is progressively denoised and completed. While [42] makes use of image data to help with frame-to-frame alignment [21] makes no use of any image data. However, since DynamicFusion system uses a fixed reference frame where the volumetric model is incrementally updated, it cannot deal with fast motion, major shape deformation or topology changes. To overcome these limitations, Dou *et al.* [10] proposed a new multi-view real time performance capture system for challenging scenes. By periodically resetting the reference frame to adapt to shape changes over time and robustly fusing data and reference volumes based on correspondence estimation and alignment error, their new Fusion4D system can robustly handle large frame-to-frame motion and topology changes.

While our work is related to and certainly inspired by these depth-based formulations, the underlying estimation problems are fundamentally different. The availability of a depth image for each frame turns the problem of 3D estimation of non-rigid geometry into one of denoising or fusion, while our monocular RGB-only reconstruction problem must infer the 3D deformations of a template purely from 2D image motion data.

Table 1 summarizes our main contributions and the differences with respect to the six most closely related approaches, namely the dense NRS*f*M approach of Garg *et al.* [12], the direct template-based monocular reconstruction approach of Malti *et al.* [19], the total face reconstruction system of Suwajanakorn *et al.* [35], real-time RGB-D non-rigid reconstruction system of Zollhofer *et al.* [42], DynamicFusion system of Newcombe *et al.* [21] and the multi-view Fusion4D system of Dou *et al.* [10].

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 [6]. 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. This template acquisition step is described in more details in section 4. The template is then converted to a triangular 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 [8], [12], [25], [37], it scales to the streaming of long sequences, with the complexity of optimization guaranteed to 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 2 provides an example of the output of this process. As shown in the figure, this step takes sampled images as input and generates a set of colored coarse-to-fine meshes.

The process of the template acquisition is an application of an existing multi-view stereo (MVS) technique [6]; consequently we provide only an overview of the process with appropriate references to the methods used.

Extrinsic Calibration The collection of frames from the video were calibrated automatically using an implementation (VisualSFM [41]) of standard rigid structure-from-motion (S*f*M). This was observed to be robust to some 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) [7].

Depth-Map Extraction Once we have a calibrated set of frames, we extract a depth-map using the stereo method of [6]. 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 [6].

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 [6], we combine the depth-maps to recover a single watertight mesh $\widetilde{\mathbf{S}}$ using the volumetric fusion technique of [40] combined with the probabilistic visibility approach of [14].

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 2 (right). This is achieved by iteratively down-sampling and refining the template mesh using the isotropic surface remeshing method (and implementation) of Fuhrmann *et al.* [11]. 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 3 shows a triangulated example of the multi-scale colored mesh representation.

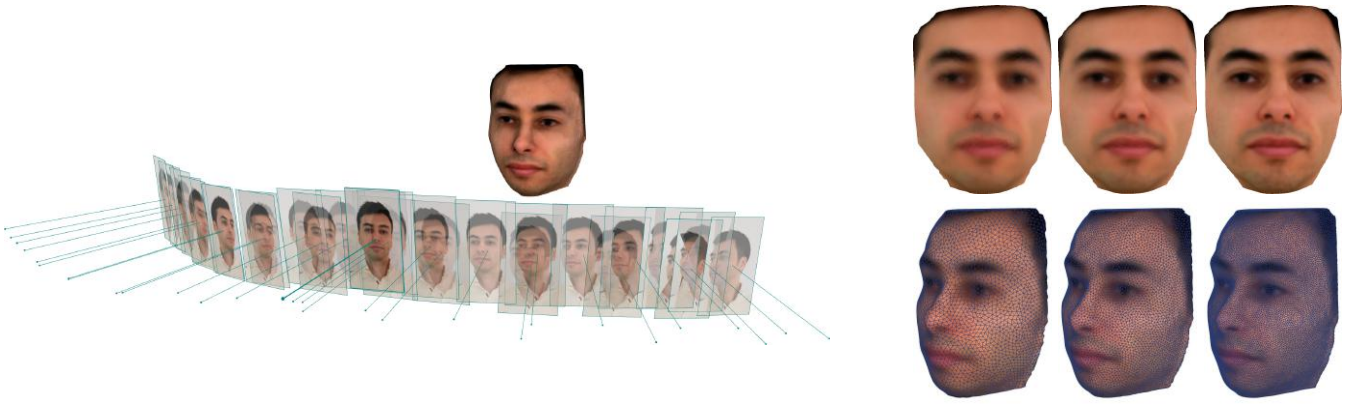

Fig. 2: Template acquisition step. **Left:** A volumetric representation is generated from stereo depth maps taken over a rigid subsequence. This is then transformed into a colored mesh. **Right:** The three scales of the template used to robustly estimate deformations.

Fig. 3: 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 dog and ball sequences.

5 Step 2: Non-Rigid Model Tracking

5.1 Our Energy

Our objective is made of a balanced combination of five 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; *(iv)* a *rotation total variation term* on the gradient of the local 3D rotations of each vertex with respect to the template and *(v)* a *temporal smoothness* term that penalizes strong frame-to-frame deformations.

The per-frame objective takes the form:

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

where λ_r, λ_a , λ_{rr} 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; the third term approximates elastic deformation and encourages the deformation to be locally rigid; while the fourth term encourages large articulation changes in the template shape. Finally, temporal smoothness is needed to avoid flickering.

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 back-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

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

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.

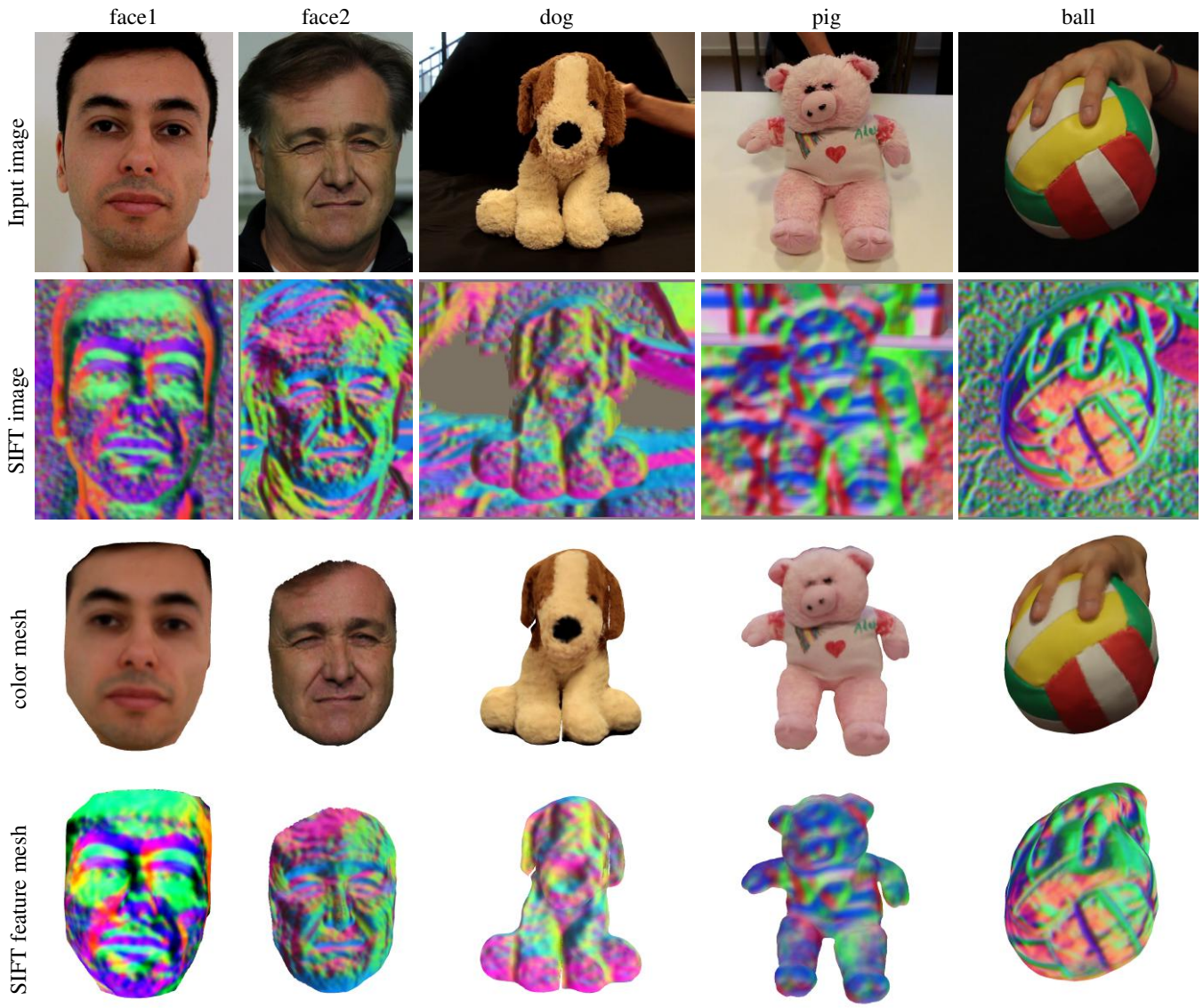

Fig. 4: Input image, dense sift feature image, color mesh and sift feature mesh of 5 different sequences.

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 $\widehat{\mathbf{S}}$.

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

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.

5.1.3 As Rigid as Possible Deformation Term E_{arap}

This cost was first proposed in [32] to allow deformable tracking of an initial mesh against a depth map. It takes the form

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

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. This cost has been widely used in non-rigid motion modelling [42], [21], [10].

5.1.4 Spatial Rotation Regularization Term E_{reg_rot}

As E_{reg} penalizes the gradient of 3D displacements, in the case of large articulation motion, this cost will be relatively large due to the fact that the gradient of 3D displacements will be approximately constant in the whole articulated region. Therefore this term will penalize strong articulated motions. To allow large articulations, we introduce a new regularization term $E_{\text{reg\_rot}}$ on local rotations in addition to the 3D displacements.

$$E_{\text{reg\_rot}}(\{\boldsymbol{A}_i\}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|(\boldsymbol{A}_i - \boldsymbol{A}_j) - (\widehat{\mathbf{A}}_i - \widehat{\mathbf{A}}_j)\|_\epsilon \qquad (6)$$

where $\boldsymbol{A}_i$ is the local arap rotation of vertex i. Unlike E_{reg}, $E_{\text{reg\_rot}}$ will be small in the articulated region, only taking nonzero values around the joints, and therefore encouraging large articulated motion.

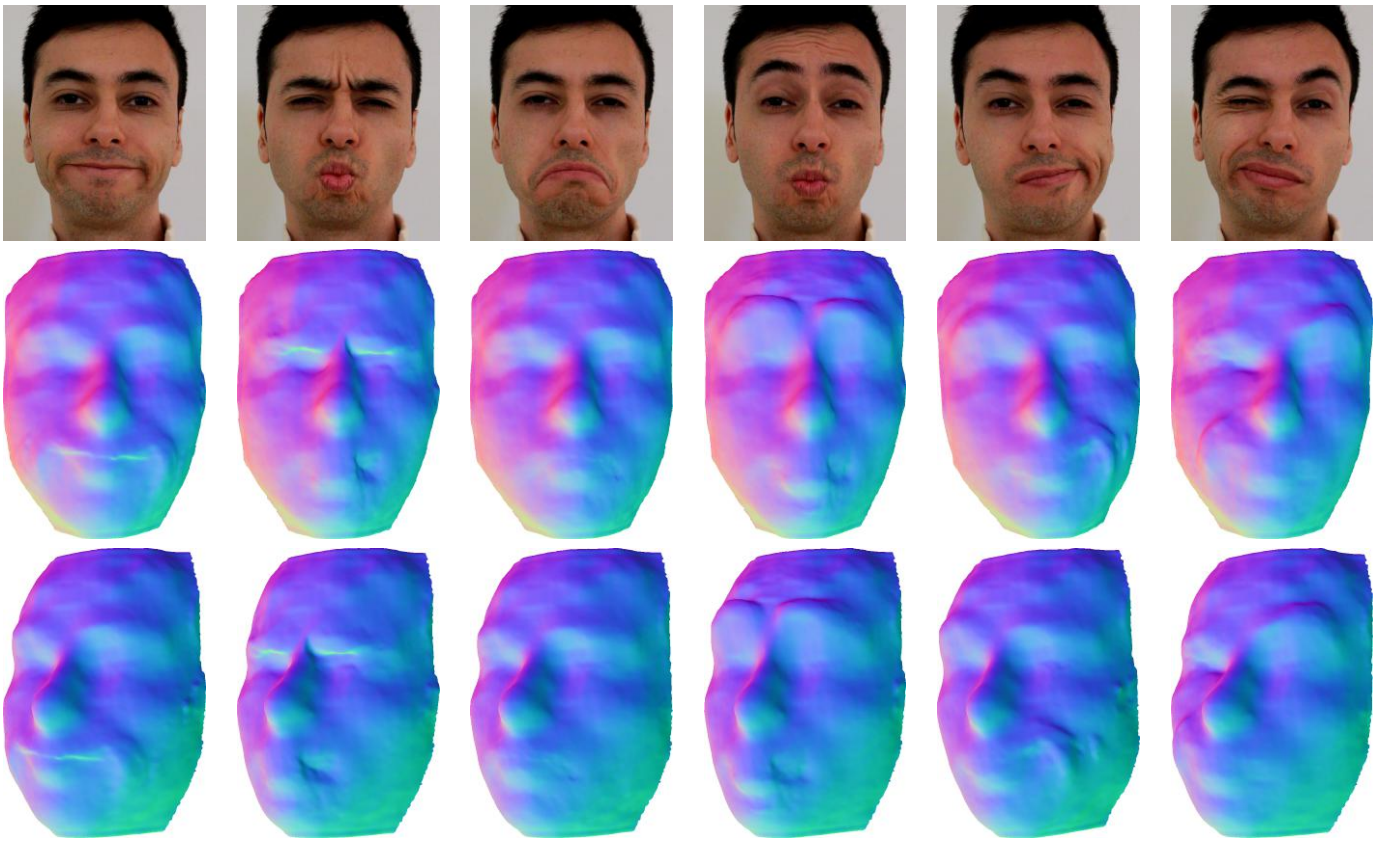

Fig. 5: Direct deformable reconstruction from our algorithm on face1 sequence.

5.1.5 *Temporal Smoothness* E_{temp}

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

$$E_{\text{temp}}(\boldsymbol{S}, \boldsymbol{t}) = \|\boldsymbol{S} - \boldsymbol{S}^{t-1}\|_{\mathcal{F}}^2 + \|\boldsymbol{t} - \boldsymbol{t}^{t-1}\|_2^2 \tag{7}$$

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.

$$E(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{N} |\widehat{\mathbf{I}}_i - \boldsymbol{I}(\pi(\boldsymbol{R}(\boldsymbol{s}_i^{t-1}) + \boldsymbol{t}))|_{\epsilon} \tag{8}$$

Then, holding the global rotation and translation constant, $\boldsymbol{s}^t$ is estimated. $\boldsymbol{R}$, $\boldsymbol{t}$, and $\boldsymbol{S}^t$ (at the coarsest level of the pyramid) are initialized using the solution taken from the previous frame, and optimization is performed using the conjugate gradient based solver from Ceres [1].

5.2.2 *Coarse-to-fine optimization and Deformation Graph*

Both the rotation and translation cost 8, 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 two times 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.* [34]. The weights are pre-computed when the template mesh is created.

$$w_k(i) = \exp(-\|\widehat{\mathbf{s}}_i - \widehat{\mathbf{s}}_k\|_2^2 / 2\sigma^2) \tag{9}$$

where $\widehat{\mathbf{s}}_i$ is the 3D position of vertex i on the finest level template mesh, while $\widehat{\mathbf{s}}_k$ is the position of vertex k on the coarse mesh. σ is given by the largest distance between all the K nearest nodes in the coarse mesh and vertex i, and the weights $w_k(i)$ are then normalized to sum up to 1.

Based on the coarse level mesh $\{\boldsymbol{s}_k\}$, local rotations $\{\boldsymbol{A}_k\}$ and weights $w_k(i)$, we estimate the location of the vertices on the fine level mesh with:

$$\boldsymbol{s}_i = \sum_{k=1}^{K} w_k(i)(\boldsymbol{A}_k(\widehat{\mathbf{s}}_i - \widehat{\mathbf{s}}_k) + \boldsymbol{s}_k) \tag{10}$$

Prolongation is also applied to the arap rotations $\{\boldsymbol{A}_i\}$ by estimating the best local rotations between the fine template mesh and the current mesh.

In our energy formulation 1, the number of variables of the optimization problem is $6N + 6$, where N is the total number

of vertices in the mesh. In our experiments, we use a 3-level mesh pyramid with 5k, 10k and 25k vertices, which gives rise to 30k, 60k and 150k variables respectively. To compute the mesh deformations more efficiently, we could use upper level mesh vertices (not necessarily the one directly above) as deformation nodes and compute the vertex position $\boldsymbol{s}_i$ of the fine level mesh via prolongation 10. Notice that in this case, K is the number of neighbouring deformation nodes of vertex i, $w_k(i)$ is the interpolation weight of node k on vertex i. Assuming there are M deformation nodes, the number of variables of the energy will be $6M+6$ and therefore the algorithm will be much more efficient when M is much smaller than N.

These deformation nodes could be introduced in both data term and other regularization terms. Furthermore, we could use different deformation nodes for the data term and regularization terms. To induce long range regularization over the deformation nodes, we could use a hierarchical deformation graph.

6 Robust Data Term

The photometric data term E_{data} from section 5.1 is based on the *brightness constancy assumption*, i.e. the corresponding point in the image should have similar colour and brightness as the template mesh vertex. In our scenario, this assumption will be violated when there are either illumination changes or shading effects caused by strong local mesh deformations. As shown in Figure 9, there is significant intensity change around the eye and mouth region when the face deforms over the whole sequence and intensity based tracking fails to capture the mesh deformation. In this section, we introduce two new data terms to robustly deal with appearance changes in the tracking.

6.1 SIFT Data Term

Feature descriptors, such as SIFT [17], SURF [3] and ORB [26], have been shown to be robust to large illumination and viewpoint changes. In order to overcome the shortcomings of intensity based tracking, we propose to use the SIFT feature descriptor image for tracking instead of the RGB values.

Specifically, we compute dense SIFT feature images offline using the VLFeat library [39]. Due to memory limitations, we perform PCA (Principal Component Analysis) on the 128 dimensional SIFT features and only keep the first 3 principal dimensions. We compute feature images for the rigid and non-rigid parts of the sequences. The rigid frames are used for template building while the non-rigid one for online tracking. In the template creation stage, instead of attaching RGB colours to mesh vertex, we attach 3-channel SIFT features.

Figure 4 shows the input image, dense SIFT feature image, colour mesh and SIFT feature template mesh for face1, face2, dog, pig and ball sequences.

6.2 NCC Data Term

Normalized Cross-Correlation(NCC) is a widely used distance measure in template matching due to its simplicity and robustness to lighting changes. It is a distance metric between two image patches. Specifically, for two patches $\mathbf{I_p}$ and $\mathbf{I_q}$, the NCC score is defined as follows:

$$NCC(\mathbf{I_p}, \mathbf{I_q}) = \frac{\mathbf{I_p} - \overline{\mathbf{I_p}}}{\|\mathbf{I_p} - \overline{\mathbf{I_p}}\|} \cdot \frac{\mathbf{I_q} - \overline{\mathbf{I_q}}}{\|\mathbf{I_q} - \overline{\mathbf{I_q}}\|} \tag{11}$$

where $\overline{\mathbf{I_p}}$ and $\overline{\mathbf{I_p}}$ are the mean values of patch $\mathbf{I_p}$ and $\mathbf{I_q}$ respectively. NCC measures the similarity of intensities in two neighbourhood regions, and is invariant to the change in average value or intensity range in the regions.

As an alternative to the intensity based photometric data term E_{data}, we compute the NCC score between local template mesh region and corresponding 2D projections on the input image.

7 Frame-to-frame Data Term

The formulation we introduced in section 5.1 is a frame-to-model tracking method, where the data term is based on the matching between a fixed template and an input frame. However, using a fixed template could fail to handle appearance changes, for example, sudden changes in the environment lighting or shading changes due to local deformations, which might be critical as our goal is to track mesh deformation accurately. In order to adapt to possible changes in the intensity over time, we compute the data term based on the difference between the intensities of projections on the previous frame and the current frame. In other words, we update the colours of template mesh vertices for each frame:

$$\widehat{\mathbf{I}}_i = \boldsymbol{I}^{t-1}(\pi(\boldsymbol{R}^{t-1}(\boldsymbol{s}_i^{t-1}) + \boldsymbol{t}^{t-1})) \tag{12}$$

Similarly, we could compute frame-to-frame data terms using NCC and SIFT features by updating the intensities or features based on the projections of tracking results from the previous frame.

8 Experimental Results

In this section we show qualitative results of our method on a variety of non-planar 3D meshes; a qualitative comparison between the results with different combinations of regularization terms and a quantitative evaluation on the face2 sequence from Valgaerts *et al.* [38]. Our results can be best viewed in the video.[2]

Qualitative results on non-planar meshes We show results on some new sequences acquired with a handheld camera. Example sequences include a *face* (Figure 5) , two soft toys – a *pig* (Figure 6) and a *dog* (Figure 7), and a *ball* being squeezed by a hand (Figure 8). 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.

Qualitative results of using different regularization terms To justify the effectiveness of each regularization term, we compared with different combinations of regularizers. Figure 11 and Figure 12 show the tracking results with and without the arap term E_{arap}, spatial rotation regularization $E_{\text{reg\_rot}}$ term and temporal smoothness term E_{temp}.

As shown in the left of figure 11, using only the spatial regularization term E_{reg} does not allow large deformations from the template and cannot capture the large articulation movement when the dog turns sideways its head. While the arap term E_{arap} allows large deformations, it offers too much freedom that the left ear gets curly when the dog rotates its head, as shown in the middle. With the right combination of all three regularization terms (E_{reg}, E_{arap} and $E_{\text{reg\_rot}}$) we show that the energy is able to capture large deformation while not too flexible to induce unnecessary deformations.

Figure 12 illustrates the effectiveness of the temporal smoothness term. Due to the ambiguity in the depth direction, it can be

2. Please see http://visual.cs.ucl.ac.uk/pubs/ddd/ for video.

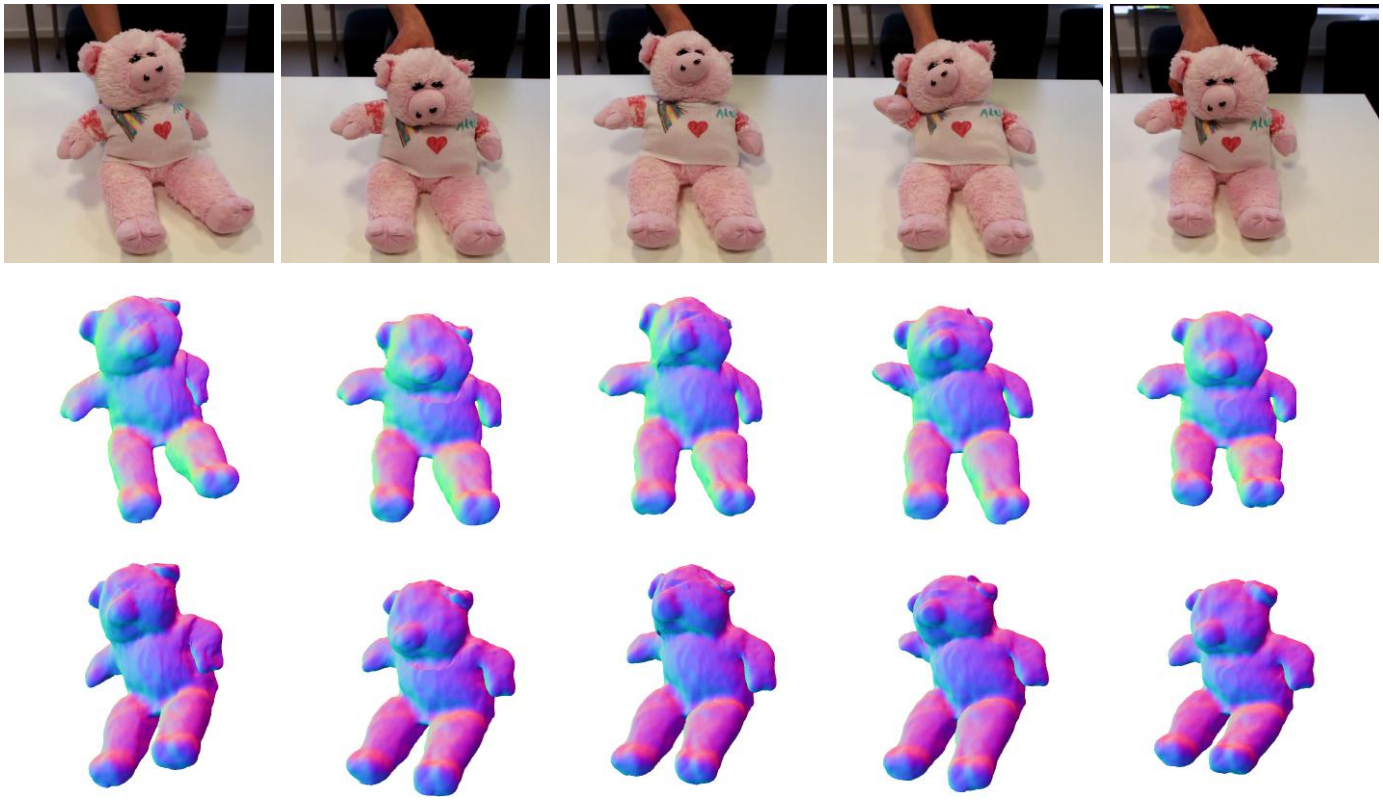

Fig. 6: Direct deformable reconstruction from our algorithm on the pig sequence. Notice that our method is able to capture the motion of the right hand, the deformation of the head and the large articulation between the body and legs.

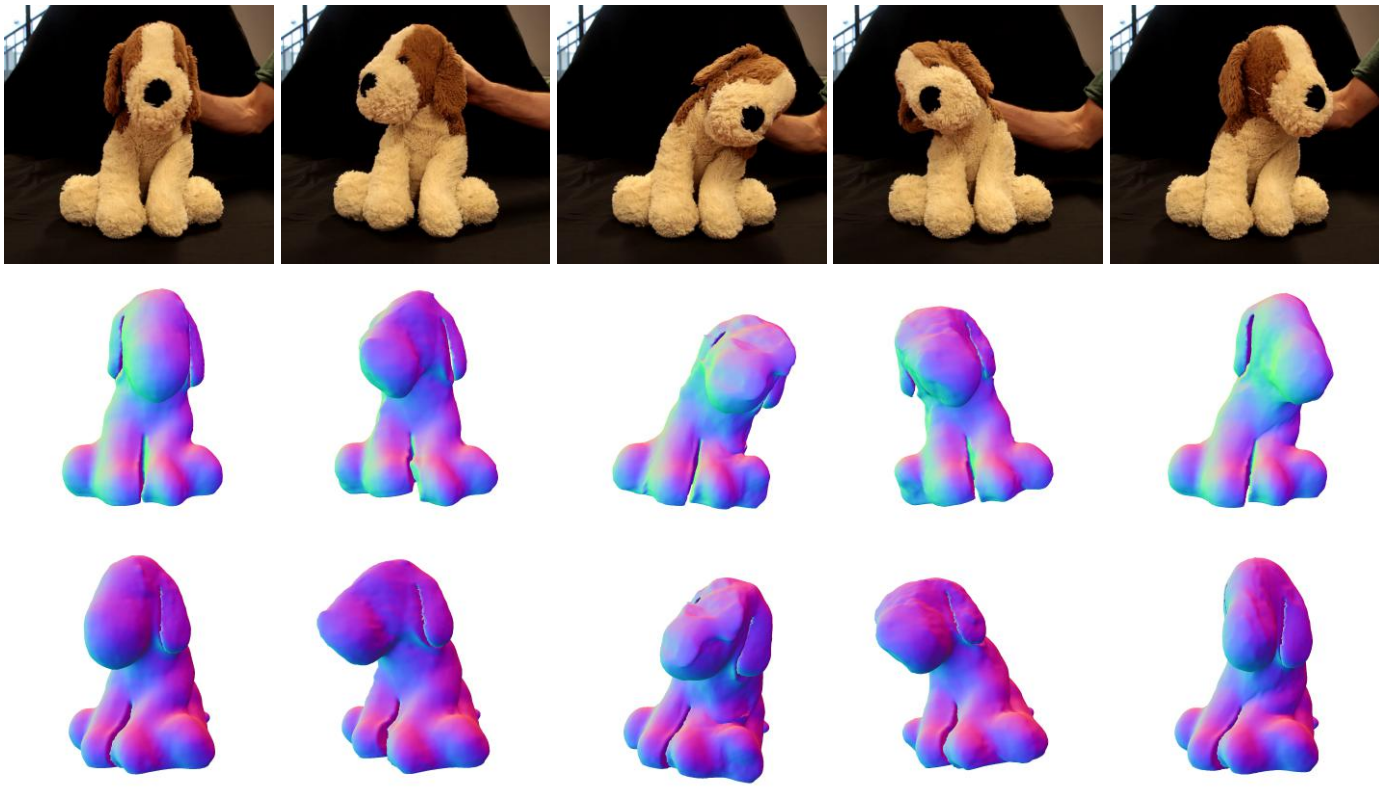

Fig. 7: Direct deformable reconstruction from our algorithm on the dog sequence. It can be seen that despite the large deformation created by the person's hand, our method successfully tracks the motion of the dog's head and the deformation of the neck.

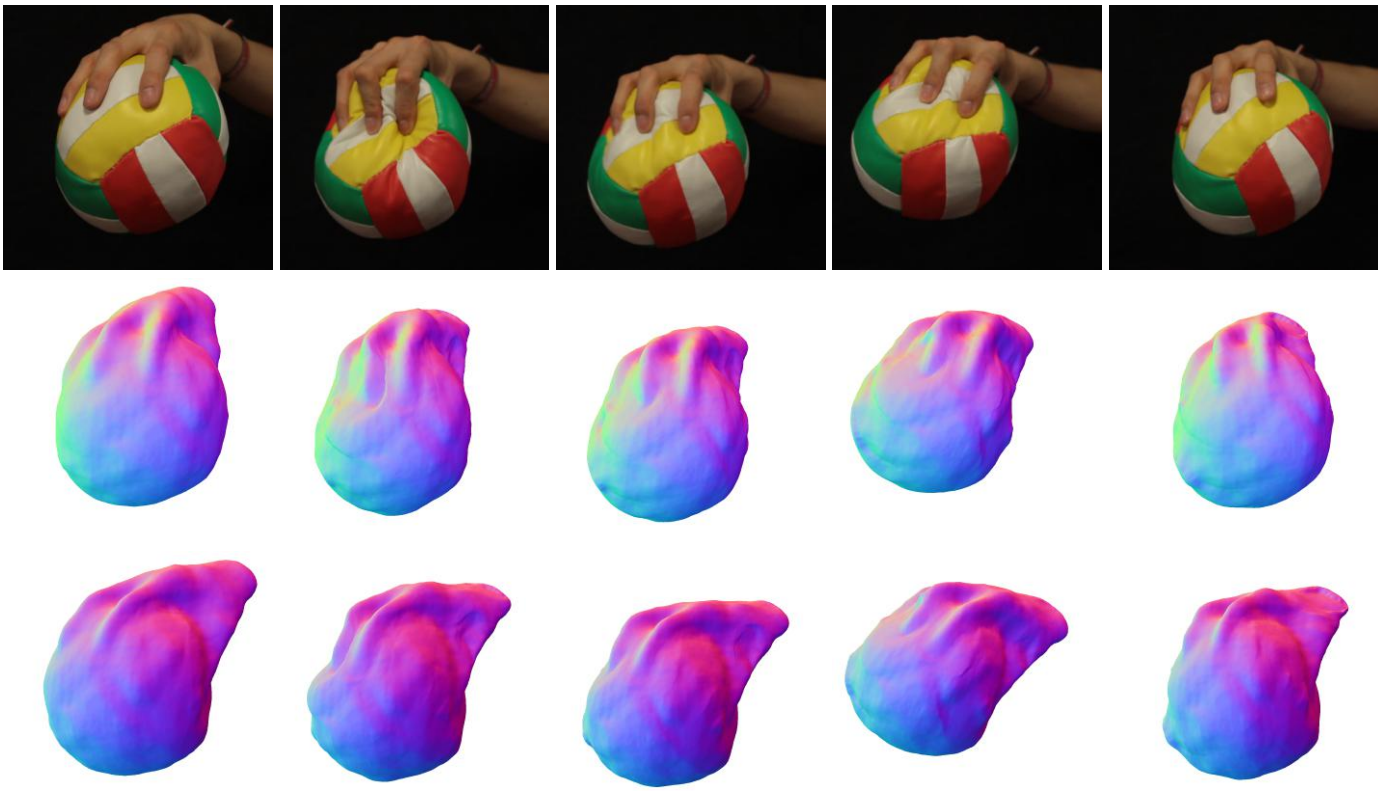

Fig. 8: Direct deformable reconstruction from our algorithm on the ball sequence. Our method is able to track the motion of the ball as well as the deformation induced by hand pressing.

seen that when there is no temporal smoothness term, the mesh tends to move forwards and backwards(as shown in the side view images on the left) without changing the 2D projections much. As shown in the overlaid image in the left, although the 2D projection of the mesh fits well to the input image, it is clear from the side view that the 3D mesh has moved substantially along the depth direction. In contrast, as shown in the right figure, when adding temporal smoothness, i.e. penalising the movement in the depth direction, the mesh tends to stay in place and does not jump forwards and backwards.

Quantitative evaluation with the face sequence from Valgaerts *et al.* [38] We evaluate our results quantitatively by taking the publicly available accurate stereo reconstruction results from Valgaerts *et al.* [38] as ground truth 3D shape.

Figure 9 shows the comparison between ground truth and tracking results as well as the corresponding error heatmaps for intensity, NCC and SIFT feature data terms. Intensity based tracking has high errors as well as more artefacts around the mouth and eyes regions. While NCC generates smoother tracking results than using SIFT, it fails to capture the deformation of the mouth and the details of the deformations. Table 2 shows the average 3D tracking errors using different data terms compared with the stereo ground truth. It is clear that using more robust data terms, such as NCC or SIFT, can improve the tracking performance. The tracking error decreases from 2.38mm to 2.31mm(NCC) and 2.22mm(SIFT).

Figure 10 shows the comparison between frame-to-frame tracking with intensity, NCC and SIFT data terms. Compared to frame-to-model tracking (Figure 9), frame-to-frame gives smoother results. However, due to accumulated errors, its performance is worse compared to frame-to-model tracking results. Table 2 shows a comparison of 3D tracking errors. We can see that in all three cases, the frame-to-frame tracking error is higher than frame-to-model tracking.

	Intensity	NCC	SIFT
Model based tracking error(mm)	2.38	2.31	2.22
Frame to frame tracking error(mm)	2.84	2.76	2.53

TABLE 2: Average tracking errors using intensity, NCC and SIFT features evaluated on the face2 sequence. All the errors are computed with respect to stereo reconstruction results from Valgaerts *et al.* [38] as ground truth. In all cases, frame-to-frame tracking gives higher error due to accumulated errors.

9 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

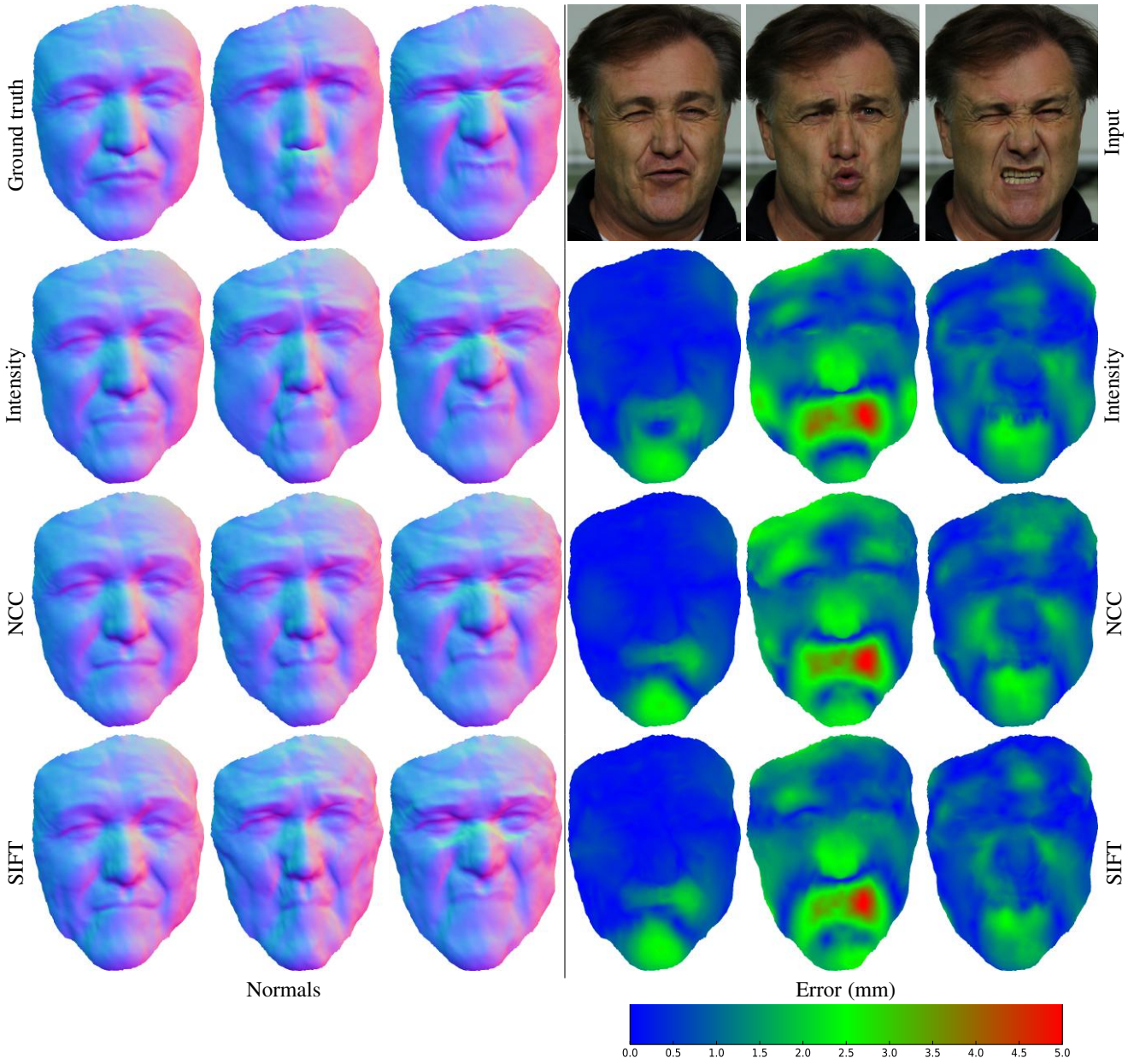

Fig. 9: Comparison results between intensity, NCC and SIFT feature based tracking with frame-to-model data term. Left figure shows the reconstruction results compared with ground truth mesh, right figure shows the error heatmap. For the error heatmap, blue corresponds to low error while red means high error.

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.

10 Acknowledgements

This work has been partly supported by the SecondHands 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] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006. 7

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

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

[6] 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

[7] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Automatic object segmentation from calibrated images. In *CVMP*, 2011. 3

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

[9] 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.

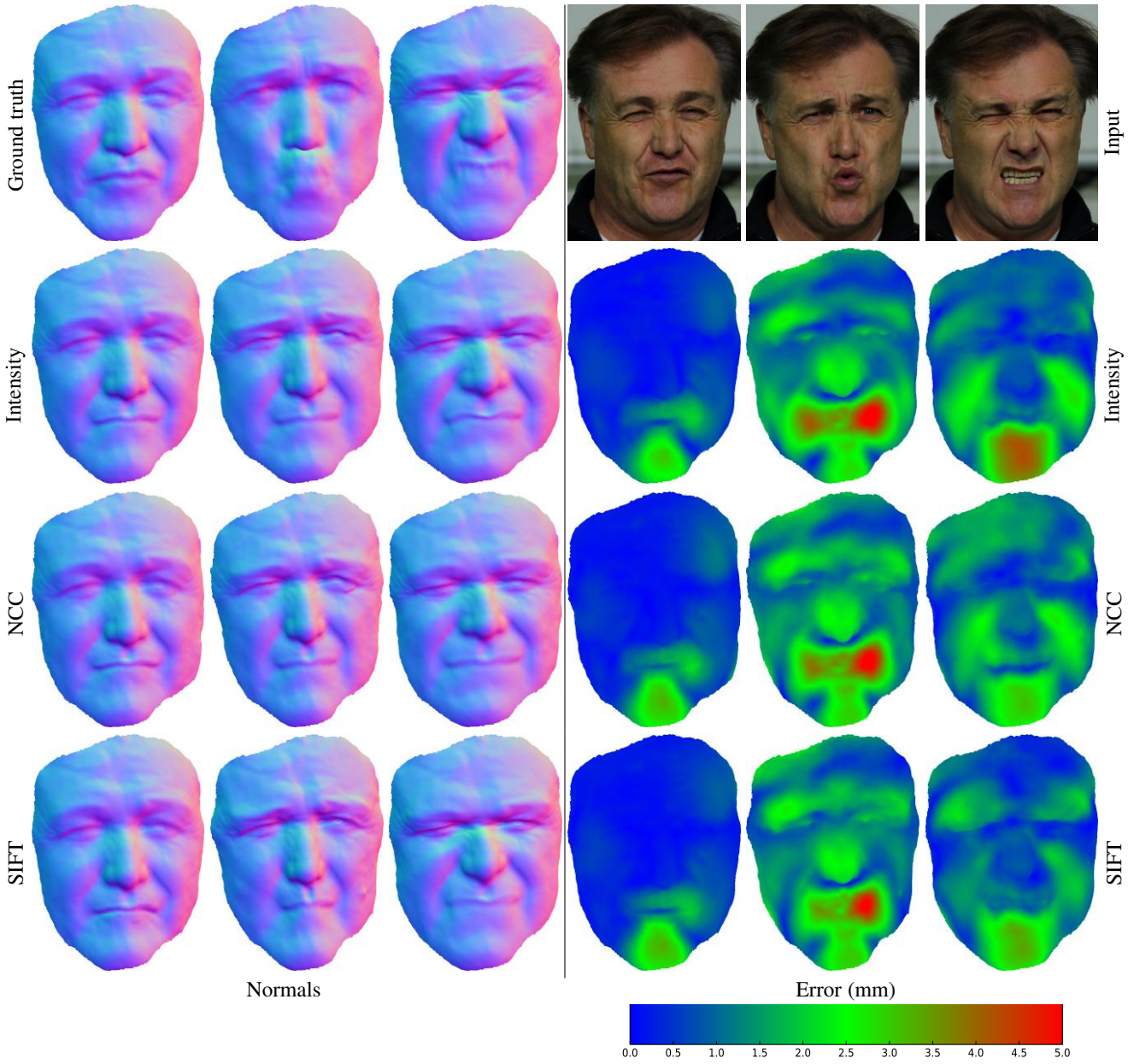

Fig. 10: Comparison results between intensity, NCC and SIFT feature based tracking with frame-to-frame data term. Left figure shows the reconstruction results compared with ground truth mesh, right figure shows the error heatmap. Different from Figure 9, for frame-to-frame data term, NCC gives worse result than intensity based tracking.

1

[10] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016. 2, 3, 5

[11] 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. 3

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

[13] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *SIGGRAPH Asia*, 2013. 2

[14] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. In *CVPR*, 2007. 3

[15] K. Kolev, P. Tanksanen, P. Speciale, and M. Pollefeys. Turning mobile phones into 3d scanners. In *CVPR*, 2014. 1, 2

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

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 7

[18] 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

[19] 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

[20] 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. 2

[21] R. Newcombe, D. Fox, and S. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015. 1, 2, 3, 5

[22] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *ICCV*, 2011. 1, 2

[23] 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

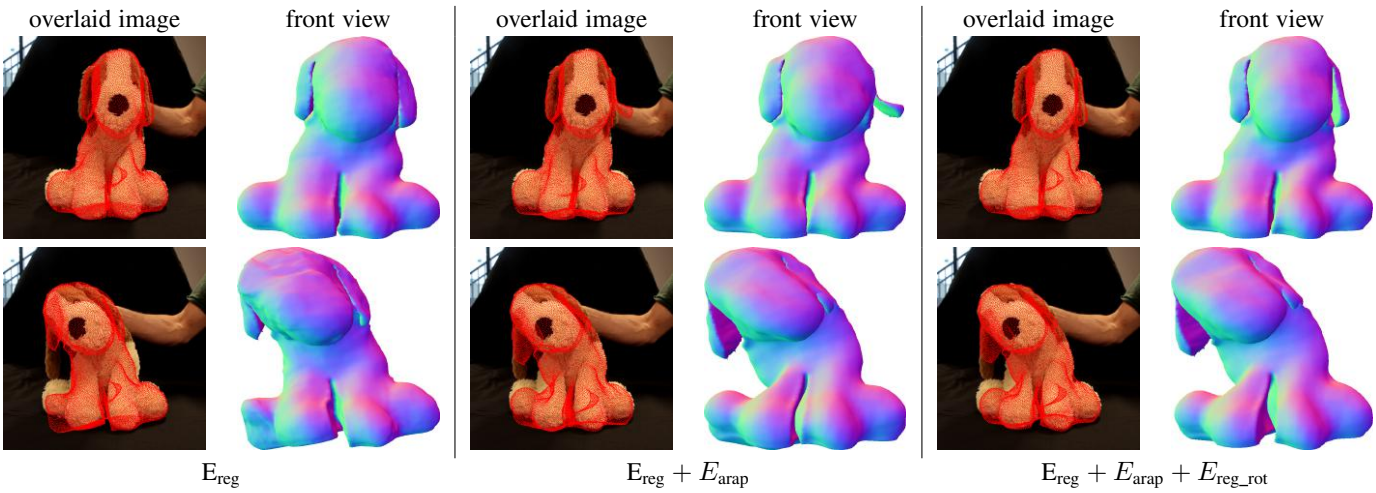

Fig. 11: Comparison results using different combinations of regularization terms, including spatial regularization E_{reg}, arap E_{arap} and rotation regularization $E_{\text{reg\_rot}}$. From left to right, we show the results of using E_{reg}, $E_{\text{reg}} + E_{\text{arap}}$ and $E_{\text{arap}} + E_{\text{arap}} + E_{\text{reg\_rot}}$ respectively. In comparison, the spatial regularization term E_{reg} alone does not allow large deformations from the template and cannot capture the large articulation movement when the dog turns its head sideways. While the arap term E_{arap} allows large deformations, it provides too much freedom. Notice that the dog's ear bends upwards incorrectly. With the right combination of all three regularization terms, we show that the energy is able to capture large deformations while not allowing too much flexibility so as to induce unnecessary deformations.

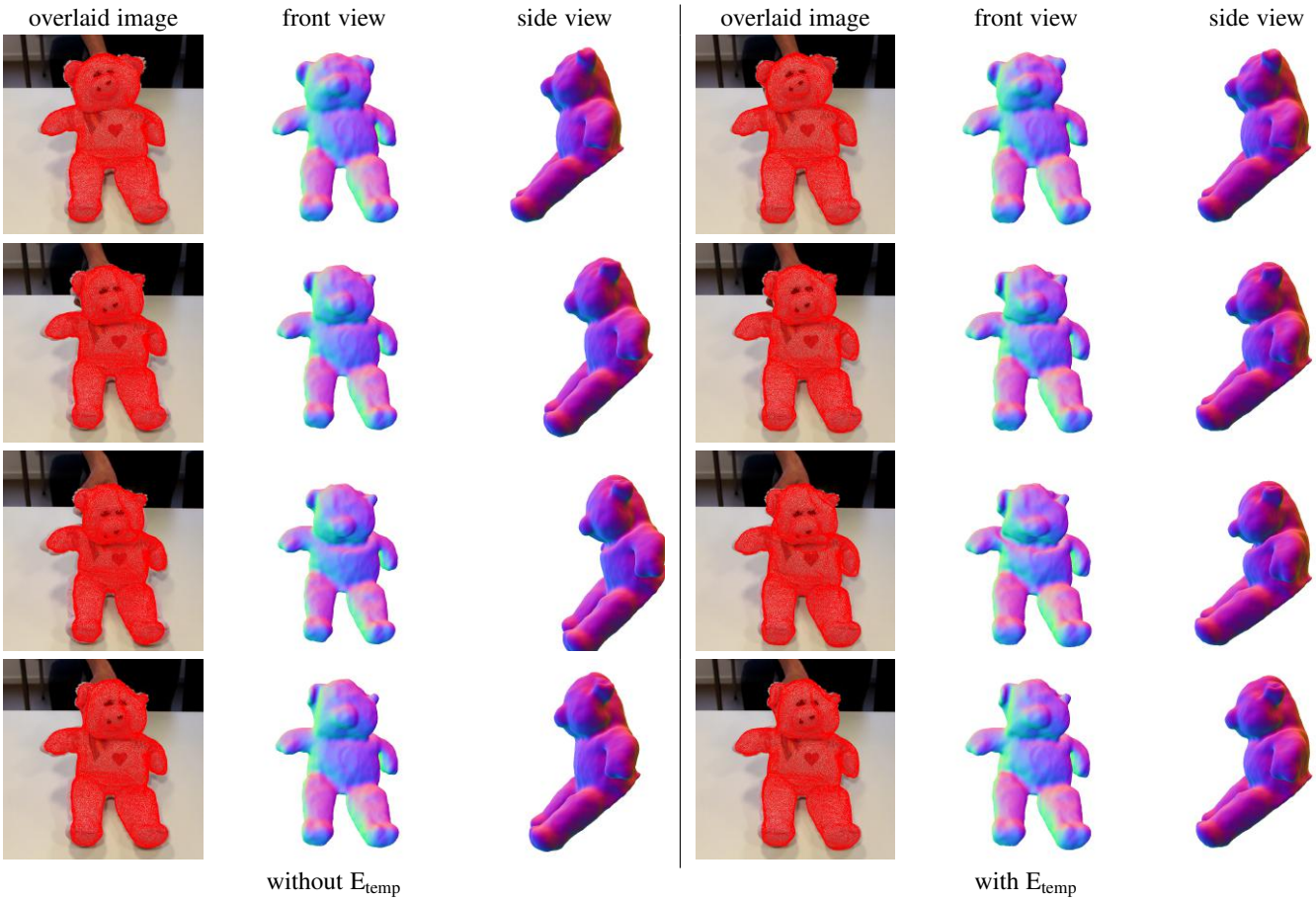

Fig. 12: Comparison results of with and without temporal smoothness regularization term E_{temp}. **Left:** tracking results without using E_{temp}, including input image overlaid with mesh projections, front view and side view of mesh normals. **Right:** corresponding tracking results with E_{temp}. Due to the ambiguity in the depth direction, it can be seen that when there is no temporal smoothness term, the mesh tends to move forwards and backwards (as shown in the side view images of in the left) without changing the 2D projections much. In contrast, in the right we fix this problem by adding temporal smoothness term, in particular by penalising the movement in the depth direction.

[24] 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

[25] 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

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571. IEEE, 2011. 7

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

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

[29] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-Form Solution to Non-Rigid 3D Surface Registration. In *ECCV*, 2008. 2

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

[31] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 1

[32] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP '07, 2007. 2, 5

[33] 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. 1, 2

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

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

[36] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, 2013. 1, 2

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

[38] L. Valgaerts, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *SIGGRAPH*, 2013. 1, 7, 9

[39] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1469–1472. ACM, 2010. 7

[40] G. Vogiatzis, C. Hernández, P. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 2007. 3

[41] C. Wu. Visualsfm: A visual structure from motion system. http://ccwu.me/vsfm/, 2011. 3

[42] 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, 5