Better Together: Joint Reasoning for Non-rigid 3D Reconstruction with Specularities and Shading

Qi Liu-Yin[1]
Qi.Liu@cs.ucl.ac.uk
Rui Yu[1]
R.Yu@cs.ucl.ac.uk
Lourdes Agapito[1]
L.Agapito@cs.ucl.ac.uk
Andrew Fitzgibbon[2]
awf@microsoft.com
Chris Russell[1]
crussell@turing.ac.uk

[1] University College London
London, UK
[2] Microsoft Research Cambridge
Cambridge, UK

Abstract

We demonstrate the use of shape-from-shading (S*f*S) to improve both the quality and the robustness of 3D reconstruction of dynamic objects captured by a single camera. Unlike previous approaches that made use of S*f*S as a post-processing step, we offer a principled integrated approach that solves dynamic object tracking and reconstruction and S*f*S as a single unified cost function. Moving beyond Lambertian S*f*S, we propose a general approach that models both specularities and shading while simultaneously tracking and reconstructing general dynamic objects. Solving these problems jointly prevents the kinds of tracking failures which can not be recovered from by pipeline approaches. We show state-of-the-art results both qualitatively and quantitatively.

1 Introduction

As the quality of 3D reconstructions of dynamic and deformable objects such as animals and faces has improved, robustness and the reconstruction of semantically meaningful details like smile and frown lines become more important. These transient fine details can not be recovered by tracking alone, and require an understanding of the lighting in the environment and a knowledge of how the surface normals of the object affect its illumination.

While these shading artefacts can inform highly-detailed reconstructions, they can also prevent the tracking of objects. In homogeneously textured regions, such as human skin, the variance in the appearance of a patch due to lighting changes can be much greater than the difference in appearance between one patch and the next. A combination of these effects makes it vital that we model illumination changes if we wish to correctly capture facial deformations particularly those of the brow and cheeks.

Pages 42.1-42.12
DOI: https://dx.doi.org/10.5244/C.30.42

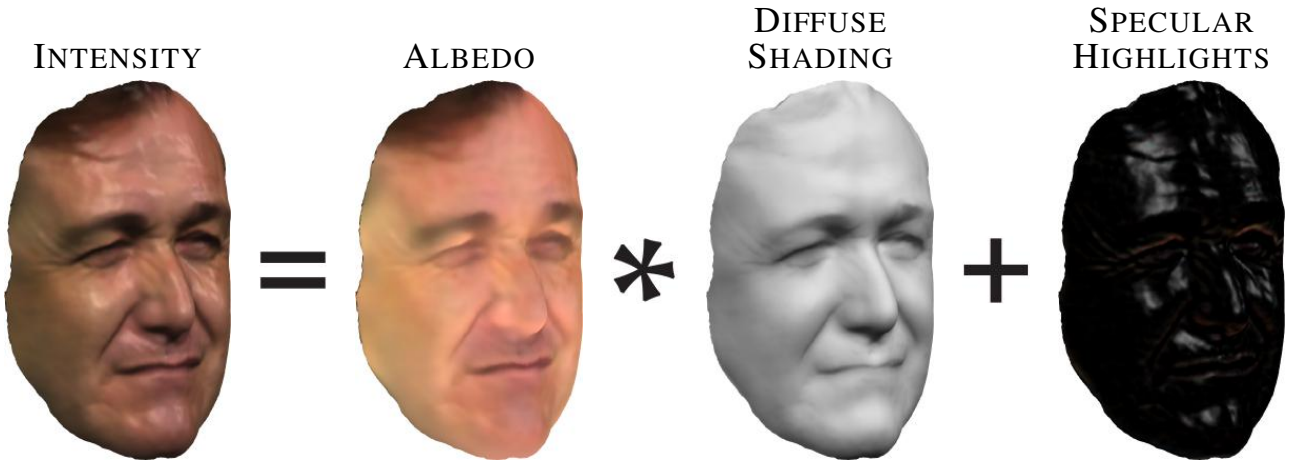

Figure 1: Intensity decomposition into the product of albedo and diffuse shading (as a function of spherical harmonics and 3D shape estimation) plus the specular component.

The instability of colour as a tracking cue is well known and much remarked upon in the literature. Focusing on recent works in dynamic 3D reconstruction using depth or multi-camera capture, it is noticeable how papers such as [5, 13] make use of raw depth maps without colour information in reconstruction. Similarly, although the RGB-D based work *et al*. [27] made use of colour information they only matched appearance between pairs of adjacent frames as over long sequences changes in shadow and illumination made colour matching unreliable. These problems can largely be ignored in the reconstruction of rigid scenes that can be assumed not to be moving relative to the lighting environment. Here shading artefacts remain constant through out the sequence, while specularities typically occur sparsely and can be handled without being explicitly modelled through the use of robust statistics [12].

In the field of non-rigid monocular reconstruction from RGB video, we are not so fortunate. With only a single RGB camera as input, we must make use of colour information. However, without depth information, matching colour only between pairs of frames is prone to drift, with many tracks gradually diffusing away from an object over long sequences [17]. Similarly, moving objects can no longer be assumed to be static with respect to the lighting environment, and outside of a controlled studio-lit environment, changes in the orientation of objects lead to significant changes in appearance. Such changes often lead to the failure of direct image intensity based trackers such as [26]. Building from cutting edge approaches to non-rigid monocular reconstruction from RGB video and SfS, we propose a unified framework for jointly reasoning about shape-from-shading and reconstructing arbitrary deforming objects. Unlike existing methods, our general approach is not object specific and targets non-Lambertian surfaces such as skin while modelling both specularities and shading. Further, we empirically demonstrate that modelling the non-Lambertian properties of surfaces such as skin, and capturing both specularities and shading is vital for the joint integration of SfS with non-rigid reconstruction.

One of the main challenges in non-rigid 3D reconstruction lies in evaluating the quality of reconstructions. It is particularly challenging to capture dense deforming objects of interest with sufficiently high fidelity under real world lighting conditions. For example, depth data from an infra-red structured light source. e.g. the Microsoft Kinect or [27], can not be captured under strong natural light, while multi-camera visible-light techniques such as [19] require relatively uniform lighting to maintain tracking. To validate our approach we both compare on real world sequence captured using the work of [19], and use this data to generate

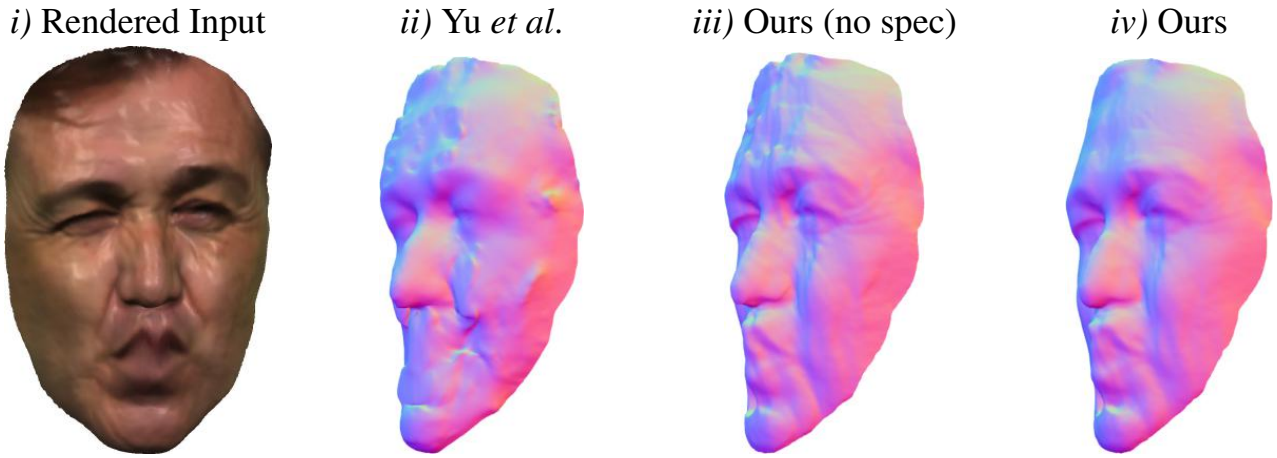

Figure 2: An illustration of our, and rival, approaches on a synthetic face sequence of a specular object under variable lighting. From Left to right: *i)* Sample frame *ii)* The direct reconstruction of Yu *et al*. [26], that does not consider shading artefacts. *iii)* Our new approach that integrates SfS with non-rigid reconstruction but does not consider specularities. *iv)* Our unified framework for SfS, non-rigid structure from motion, and specularity modelling. This framework improves the accuracy of our approach over Yu *et al*. by a factor of nearly 240% reducing the RMS error from 9.28mm to 3.84mm.

realistic synthetic sequences containing severe shading artefacts that could not be tracked by [19]. Our method displays a strong qualitative and quantitative improvement over these rival methods. See figure 2 and section 7 for details.

2 Related Work

All previous attempts to unify shape-from-shading with non-rigid 3D reconstruction from RGB video have been pipeline approaches [7, 11, 18, 19] which first coarsely reconstruct these deformable objects and then apply shape-from-shading to refine the initial reconstruction. Examples of this include, the seminal work Face Reconstruction in the Wild [9], which first made use of automatic point correspondences to compute warps and align images of a variety of celebrities, before reconstructing faces using SfS to build dense face models. This was followed by [18], which refined a coarser intensity based model using SfS.

Varol *et al*. [20] fused shape-from-shading with non-rigid reconstruction, but only performed shape-from-shading on untextured regions of the objects, and non-rigid reconstruction on the textured areas, before fusing these reconstructions as a post-processing step. Moreover, they required a known light field and could not reconstruct high-frequency details such as facial creases. Several works have made use of SfS in refining depth maps, either captured directly using a depth scanner[6, 15, 25] or captured using a multi-camera setup [19]. Of the RGB-D approaches [6] is the most related to ours, and computes both SfS and specularities in order to enhance their depth maps. Previous works have also used SfS to improve tracking: In multi-camera work Beeler *et al*. [3] used a pipeline approach to improve the tracking and refine the shape of an initial reconstruction by both estimating and removing ambient occlusions. While Xu *et al*. [24] defined linear equations for modelling changes of illumination and position that occur when tracking a rigid object in video.

Our work builds on the recent template-based approach to monocular and direct non-rigid 3D reconstruction of Yu *et al*. [26]. This work made no use of SfS, but generated vivid reconstructions of objects by deforming a known template to match direct photometric cost. We extend this direct formulation by augmenting the direct photometric cost with terms that capture the change in appearance that goes with shape and shading, leading to more lifelike and plausible reconstructions.

Algorithm 1: Joint non-rigid 3D reconstruction and shape-from-shading

Input : 3D Template mesh $\widehat{\mathbf{S}}$ + template albedo $\widehat{\boldsymbol{\rho}}_i$ (obtained using Algorithm 2)
Current video frame $\mathbf{I}^t$
Solution to previous frame $\{\boldsymbol{S}^{t-1}, \boldsymbol{R}^{t-1}, \boldsymbol{t}^{t-1}, \mathbf{l}^{t-1}, \boldsymbol{\beta}^{t-1}\}$
Output: Deformed shape $\boldsymbol{S}^t$, rotation $\boldsymbol{R}^t$, translation $\boldsymbol{t}^t$,
spherical harmonic coefficients $\mathbf{l}^t$ and specularities $\boldsymbol{\beta}^t$ for current frame t

for *each new image frame* $\mathbf{I}^t$ **do**
  Initialise $\{\boldsymbol{S}^t, \boldsymbol{R}^t, \boldsymbol{t}^t, \mathbf{l}^t, \boldsymbol{\beta}^t\} \leftarrow \{\boldsymbol{S}^{t-1}, \boldsymbol{R}^{t-1}, \boldsymbol{t}^{t-1}, \mathbf{l}^{t-1}, \boldsymbol{\beta}^{t-1}\}$
  Minimise (3) w.r.t. rigid alignment $\{\boldsymbol{R}^t, \boldsymbol{t}^t\}$ holding $\{\boldsymbol{S}^t, \mathbf{l}^t, \boldsymbol{\beta}^t\}$ constant
  Minimise (3) w.r.t. deformations and lighting $\{\boldsymbol{S}^t, \mathbf{l}^t\}$ holding $\{\boldsymbol{R}^t, \boldsymbol{t}^t, \boldsymbol{\beta}^t\}$ constant
  Minimise (3) w.r.t. specularities $\boldsymbol{\beta}^t$ holding $\{\boldsymbol{S}^t, \mathbf{l}^t, \boldsymbol{R}^t, \boldsymbol{t}^t\}$ constant
end

3 Problem Formulation

Consider a single RGB perspective camera, of known internal calibration, observing a non-rigid object. We propose a sequential, frame-by-frame, approach to capture both the 3D geometry and the reflectance properties of the non-rigid object. We parameterise the object at time-step t as a mesh $\boldsymbol{S}^t$ with N vertices with associated 3D coordinates $\boldsymbol{S}^t = \{\boldsymbol{s}_i^t\}$, $i = 1..N$.

Our proposed approach is summarised in Algorithm 1. The goal for each incoming frame at time t is to estimate the current 3D coordinates of the vertices of the mesh, the light field – parameterized in terms of spherical harmonics – and specularities, 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 method are: the current image frame $\mathbf{I}^t$, the solution to the previous frame, and a 3D template of the object (including its geometry $\widehat{\mathbf{S}}$ and albedo map $\widehat{\boldsymbol{\rho}}_i$) acquired in a preliminary stage described in Section 6 and Algorithm 2. Note that all variables related to the template are denoted with $\widehat{\ }$.

4 Reflectance Model

Modern solutions [18, 19, 26] to *direct* 3D reconstruction of non-rigid objects from RGB video adopt an energy optimisation approach that minimises a robust photometric cost based upon *brightness constancy*. In other words, they jointly estimate dense correspondences alongside non-rigid deformations, by penalising differences in intensity between images and the new deformed shape, which is assumed to be the same colour and brightness of a reference template. As the points on the object change colour in response to differences in illumination or in shading caused by strong deformations these methods need to use robust costs to cope with deviations from the model.

In contrast, our method explicitly models the reflectance properties of non-Lambertian objects and can handle materials which exhibit a mixture of specular and diffuse reflection properties. In practice we adopt an approximation of the Phong reflection model which models light leaving the non-rigid object at point i as the sum of two additive terms: a viewpoint-independent diffuse term and a view-dependent specular term: $\boldsymbol{I}_i = \boldsymbol{I}_i^{\text{diff}} + \boldsymbol{\beta}_i$.

To increase our robustness to changes in lighting and shading and to recover high frequency details of the object geometry, we decouple the diffuse component into the product

Figure 3: Best viewed in colour: A qualitative comparison of our method against that of Yu *et al*. [26], on their face dataset.

of object albedo and the object irradiance or shading (see figure 1). While the albedo is independent of the surface orientation, the shading is a function of the surface normal at each vertex i $\boldsymbol{I}_i^{\text{diff}} = \boldsymbol{\rho}_i r(\mathbf{n}_i(\boldsymbol{S}))$. Here $\boldsymbol{\rho}_i$ is the RGB reflectance or albedo, $\mathbf{n}_i(\cdot)$ is a function that returns the direction of the surface normal at vertex i, and $r(\cdot)$ is an irradiance map function that returns the shading value given the surface orientation vector. We assume white illumination so $r(\cdot)$ returns a single scalar value. Following Basri and Jacobs [2] we model the irradiance map using a spherical harmonic basis

$$r(\mathbf{n}_i(\boldsymbol{S})) = \sum_{n=0}^{N} \sum_{m=-n}^{n} l_{nm} Y_{nm}(\mathbf{n}_i(\boldsymbol{S})) = \mathbf{l} \cdot Y(\mathbf{n}_i(\boldsymbol{S})) \tag{1}$$

where l_{nm} is the coefficient associated with the spherical harmonic function Y_{nm}. We limit our approximation to second order spherical harmonics, i.e. $N = 2$ giving $\mathbf{l}$ nine coefficients.

If we consider a video of a non-rigid object evolving over time $\boldsymbol{S}^t$, our reflection model allows us to write the predicted image intensity of point i observed at time t as

$$\boldsymbol{I}_i^t = \boldsymbol{\rho}_i \mathbf{l}^t \cdot Y(\mathbf{n}_i(\boldsymbol{S}^t)) + \boldsymbol{\beta}_i^t \tag{2}$$

It is clear that our reflection model allows us to cope not only with varying geometry ($\boldsymbol{S}^t$) but also varying illumination coefficients ($\mathbf{l}^t$) and specularities ($\boldsymbol{\beta}_i^t$). Notably, while the image brightness of vertex i might vary over time t due to possible changes in illumination and object surface normals $\mathbf{n}_i(\boldsymbol{S}^t)$ caused by the deformations, its albedo $\boldsymbol{\rho}_i$ is constant over the entire sequence.

Our insight and the main contribution of this work is to track the non-rigid deformations of the object based on *albedo constancy* instead of the more classical *image brightness constancy* constraint which does not hold for non-rigid objects or when the illumination varies over time. In this way, we can take advantage of the changes in illumination and shading to recover high frequency details in non-rigid objects and by increasing 3D tracking accuracy.

5 A Sequential Approach to Joint Non-Rigid 3D Reconstruction and Reflectance Estimation

Much like Yu *et al*. [26], we use a template-based approach to track and reconstruct non-rigid objects. However, while [26] assigned a fixed intensity to each vertex on the template mesh, we decompose the intensity of each vertex (see equation (2)) allowing us to take advantage of the reflectance properties of the object to improve the resulting reconstructions.

5.1 Our Energy

Our per-frame objective takes the form

$$E(\boldsymbol{S},\boldsymbol{R},\boldsymbol{t},\mathbf{l},\boldsymbol{\beta})=E_{\text{data}}(\boldsymbol{S},\boldsymbol{R},\boldsymbol{t},\mathbf{l},\boldsymbol{\beta})+E_{\text{smooth}}(\boldsymbol{S},\boldsymbol{\beta})+E_{\text{arap}}(\boldsymbol{S})+E_{\text{temp}}(\boldsymbol{S},\boldsymbol{t},\mathbf{l},\boldsymbol{\beta})+E_{\text{sparse}}(\boldsymbol{\beta}) \quad (3)$$

Data Term E_{data}: Our data term is a direct photometric cost. Rather than minimising the more commonly used *brightness constancy constraint* we use the more complex reflectance model described in (2) which decomposes the intensity of vertex i into the product of a constant albedo and a time-varying shading term (where variations can be due to changes in illumination or to strong deformations) and explicitly models specularities.

$$E_{data}\left(\mathbf{R},\mathbf{t},\mathbf{S},\mathbf{l},\boldsymbol{\beta}\right)=\sum_{i\in\mathcal{V}}\left\|\mathbf{I}\left(\pi\left(\mathbf{R}(\mathbf{s}_i)+\mathbf{t}\right)\right)-\widehat{\boldsymbol{\rho}}_i\mathbf{l}\cdot Y\left(\mathbf{R}(\mathbf{n}_i(\boldsymbol{S}))\right)-\boldsymbol{\beta}_i\right\|_{\varepsilon} \quad (4)$$

where $\mathcal{V}$ is the set of estimated visible vertices, $\pi(\cdot)$ is the projection from 3D points to image coordinates, known from camera calibration, and $\|\cdot\|_{\varepsilon}$ is the Huber loss.

Spatial Smoothness Term E_{smooth}: This regularisation term encourages spatially smooth deformations of the shape and specularities. In practice the spatial smoothness on the shape is decoupled into two terms: a total variation term that encourages smooth deformations of the shape $\boldsymbol{S}$ with respect to the template $\widehat{\mathbf{S}}$ and a *Laplacian* smoothness term

$$\begin{aligned}E_{\text{smooth}}(\boldsymbol{S},\boldsymbol{\beta}) &= E_{\text{smooth}}(\boldsymbol{S})+E_{\text{smooth}}(\boldsymbol{\beta}) = E_{\text{TV}}(\boldsymbol{S})+E_{\text{Laplacian}}(\boldsymbol{S})+E_{\text{spec}}(\boldsymbol{\beta}) \quad (5)\\ &= \sum_{i\in\mathcal{V}}\left(\sum_{j\in\mathcal{N}_i}\|(\boldsymbol{s}_i-\boldsymbol{s}_j)-(\widehat{\mathbf{s}}_i-\widehat{\mathbf{s}}_j)\|_{\varepsilon}+\frac{1}{|\mathcal{N}_i|}\left\|\sum_{j\in\mathcal{N}_i}(\mathbf{s}_i-\mathbf{s}_j)\right\|_2^2+\sum_{j\in\mathcal{N}_i}\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\right\|_{\varepsilon}\right)\end{aligned}$$

where $\mathcal{N}_i$ is the neighbourhood of i.

ARAP Term E_{arap}: This *as rigid as possible* cost [16] encourages non-rigid objects to preserve locally rigidity while deforming. It allows for local rotations to occur while preserving the relative locations between neighbouring points.

$$E_{\text{arap}}(\boldsymbol{S},\{\boldsymbol{A}_i\})=\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i}\|(\boldsymbol{s}_i-\boldsymbol{s}_j)-\boldsymbol{A}_i(\widehat{\mathbf{s}}_i-\widehat{\mathbf{s}}_j)\|_2^2 \quad (6)$$

where the variables A_i describe per-point local rotations.

Temporal Smoothness E_{temp}: This set of temporal regularisers prevents flickering throughout the sequence

$$E_{\text{temp}}(\boldsymbol{S},\boldsymbol{t},\mathbf{l},\boldsymbol{\beta}_i)=\left\|\boldsymbol{S}-\boldsymbol{S}^{t-1}\right\|_{\mathcal{F}}^2+\left\|\boldsymbol{t}-\boldsymbol{t}^{t-1}\right\|_2^2+\left\|\mathbf{l}-\mathbf{l}^{t-1}\right\|_2^2+\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{t-1}\right\|_2^2 \quad (7)$$

Algorithm 2: 3D Template acquisition

Input : Rigid image subsequence $\{\mathbf{I}^{\mathbf{f}}_{\mathbf{rigid}}\}$ $f = 1, \cdots, F$

Output: 3D coordinates of template mesh vertices $\widehat{\mathbf{S}} = \{\widehat{\mathbf{S}}_i\}$ and Template albedo map $\widehat{\boldsymbol{\rho}} = \{\widehat{\boldsymbol{\rho}}_i\}$ where $i = 1 \cdots N$

1 Obtain rigid camera poses for each frame $\{\mathbf{I}^{\mathbf{f}}_{\mathbf{rigid}}\}$ using VisualSFM [23]

2 Estimate 3D template mesh vertices $\widehat{\mathbf{S}} = \{\boldsymbol{s}^t_i\}$ using MVS [4, 8, 21]

3 Estimate diffuse component $\widehat{\mathbf{I}}^{\mathbf{diff}}_{\mathbf{i}}$ $\forall i$ template vertices as median colour over all frames

4 Solve for the illumination coefficients $\widehat{\mathbf{l}}$ minimising (9) assuming white albedo

5 Solve for albedo map of the template $\widehat{\boldsymbol{\rho}} = \{\widehat{\boldsymbol{\rho}}_i\}$ minimising (10)

where $\boldsymbol{S}^{t-1}$, $\boldsymbol{t}^{t-1}$, $\mathbf{l}^{t-1}$, $\boldsymbol{\beta}^{t-1}$ are the shape, translation, spherical harmonic coefficients and specularities in the previous frame and $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm of a matrix.

Sparsity Term $E_{\mathbf{sparse}}$: This prevents the entire image being "explained away" as a specularity, by penalising the use of specularities.

$$E_{\text{sparse}}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{V}} \|\boldsymbol{\beta}_i\|_{\varepsilon} \tag{8}$$

Energy Optimisation: For reasons of efficiency, we adopt a multi-stage optimisation similar to the approach taken by real-time SLAM approaches such as [10]. Starting from the solution given by the previous frame, we hold all other coefficients fixed and optimise first over rotations and translations, followed by jointly optimising shape and spherical harmonic coefficients, and finally re-estimating the specularities. The first two of these optimisations are performed coarse-to-fine over a 3-layer spatial pyramid, providing robustness against sudden movements and deformations of the object while for efficiency reasons, specularities are only estimated at the finest level, and propagated to the coarser levels of the pyramid, ready for the next iteration. Algorithm 1 summarises our optimisation strategy. We use the Levenberg-Marquardt implementation from the Ceres solver [1] for all continuous optimisation, applying preconditioned conjugate gradient for the linear solver.

6 Template Capture

This section describes how we capture the static geometry and the reflectance properties of the object of interest – or in other words how we build the template model used for tracking. We achieve this by moving a hand-held camera around the object while it remains rigid for a few seconds, to observe it from different angles. During the template capture step, we assume that the illumination remains constant.

Template geometry: For each frame we estimate the relative pose of the camera with respect to the object using a standard off-the-shelf structure from motion approach (VisualSFM [23]). We then use an existing multi-view stereo approach [4] to produce individual depth maps for each frame. Finally the depth maps are fused using the volumetric technique of [21] and the probabilistic visibility approach of [8] to produce as output a watertight mesh of the template shape parameterized as the set of 3D vertex coordinates $\widehat{\mathbf{S}} = \{\widehat{\mathbf{s}}_{\mathbf{i}}\}$, $i = 1..N$.

Template reflectance properties: The next step of the template acquisition stage is to assign a colour value to each vertex i on the mesh. Our implicit assumption is that the light leaving

Table 1: Comparison results with Yu *et al.* [26] on 4 different synthetic sequences. We report average RMS error (in mm.) over all frames w.r.t ground truth.

	LF (mm)	SF (mm)	LC (mm)	SC (mm)
Yu *et al.* [26]	7.29	7.93	9.18	9.28
Ours (not modelling specularities)	2.91	3.28	3.50	4.21
Ours (modelling specularities)	2.73	2.89	3.42	3.84

the surface of the template is the sum of a viewpoint-independent diffuse term and a view-dependent specular term. We estimate the diffuse term $\widehat{\mathbf{I}}_{\mathbf{i}}^{\mathbf{diff}}$ as the median colour over all the frames in the rigid subsequence in which the projected vertex is visible. While some previous approaches favoured the use of the minimum observed intensity value [14], we choose to use the median as proposed by Wood *et al.* [22] since it provides robustness to shadows and errors in the camera tracking. We decompose the diffuse component of the template $\widehat{\mathbf{I}}_{\mathbf{i}}^{\mathbf{diff}}$ further into the product of an *albedo map* and an irradiance function parameterized in terms of *spherical harmonics* to approximate the illumination and the surface normals as described in (2). First we solve for the *spherical harmonic coefficients* by optimising the following photometric objective function with respect to $\hat{\mathbf{l}}$:

$$E_{template}(\hat{\mathbf{l}}) = \sum_{i \in \mathcal{V}} \left\| \widehat{\mathbf{I}}_{\mathbf{i}}^{\mathbf{diff}} - \hat{\boldsymbol{\rho}}_{\boldsymbol{i}} \mathbf{l} \cdot Y\left(\mathbf{n}_i(\widehat{\mathbf{S}})\right) \right\|_{\varepsilon} \tag{9}$$

where $\hat{\boldsymbol{\rho}}_{\boldsymbol{i}}$ is an initial assumption of the albedo map (e.g. white, uniform colour, or the result from k-means).

The *albedo map* is estimated by minimising the same cost with a small variant – we give a lower confidence to points with low shading. Also, a weighted local smoothing term is added based on the difference in intensity.

$$E'_{template}(\hat{\boldsymbol{\rho}}) = \sum_{i \in \mathcal{V}} w_i^a \left\| \widehat{\mathbf{I}}_{\mathbf{i}}^{\mathbf{diff}} - \hat{\boldsymbol{\rho}}_{\boldsymbol{i}} r(\mathbf{n}_i(\widehat{\mathbf{S}})) \right\|_{\varepsilon} + \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} w_{ij}^{a'} \left\| \hat{\boldsymbol{\rho}}_{\boldsymbol{i}} - \hat{\boldsymbol{\rho}}_{\boldsymbol{j}} \right\|_2^2 \tag{10}$$

where $w_i^a = r(\mathbf{n}_i)$ is chosen to decrease the importance placed on regions of low shading and $w_{ij}^{a'} = \exp\left(-\frac{\|\mathbf{I}_i - \mathbf{I}_j\|_2^2}{2\sigma_s^2}\right)$ to encourage points with similar appearance to share the same albedo.

7 Experimental Evaluation

We tested the proposed method on ground truth sequences generated using the reconstructed face shapes from [19], down-sampled ten times in order to reduce the runtime of our method. Each of the vertices of the mesh will have a constant albedo over time, which was estimated using the first reconstructed shape provided in a similar way as described in section 6. Using this albedo map, we render four different scenarios combining Lambertian or specular surfaces on a scene lit by two white directional lights with constant or changing intensity. Thus, we have sequences with a perfect Lambertian surface with fixed (LF) or changing illumination (LC), and a specular surface with fixed (SF) or changing illumination (SC) (see figure 2 and 4). To test these ground truth sequences we use a coarse-to-fine pyramid with a template mesh of $\sim 6,000$ vertices at the coarsest level and, for the finest level, one with the same amount of vertices as the ground truth ($\sim 24,000$ vertices). We run our method on a computer with an Intel Core i7-5930K CPU, which takes around one minute to process each

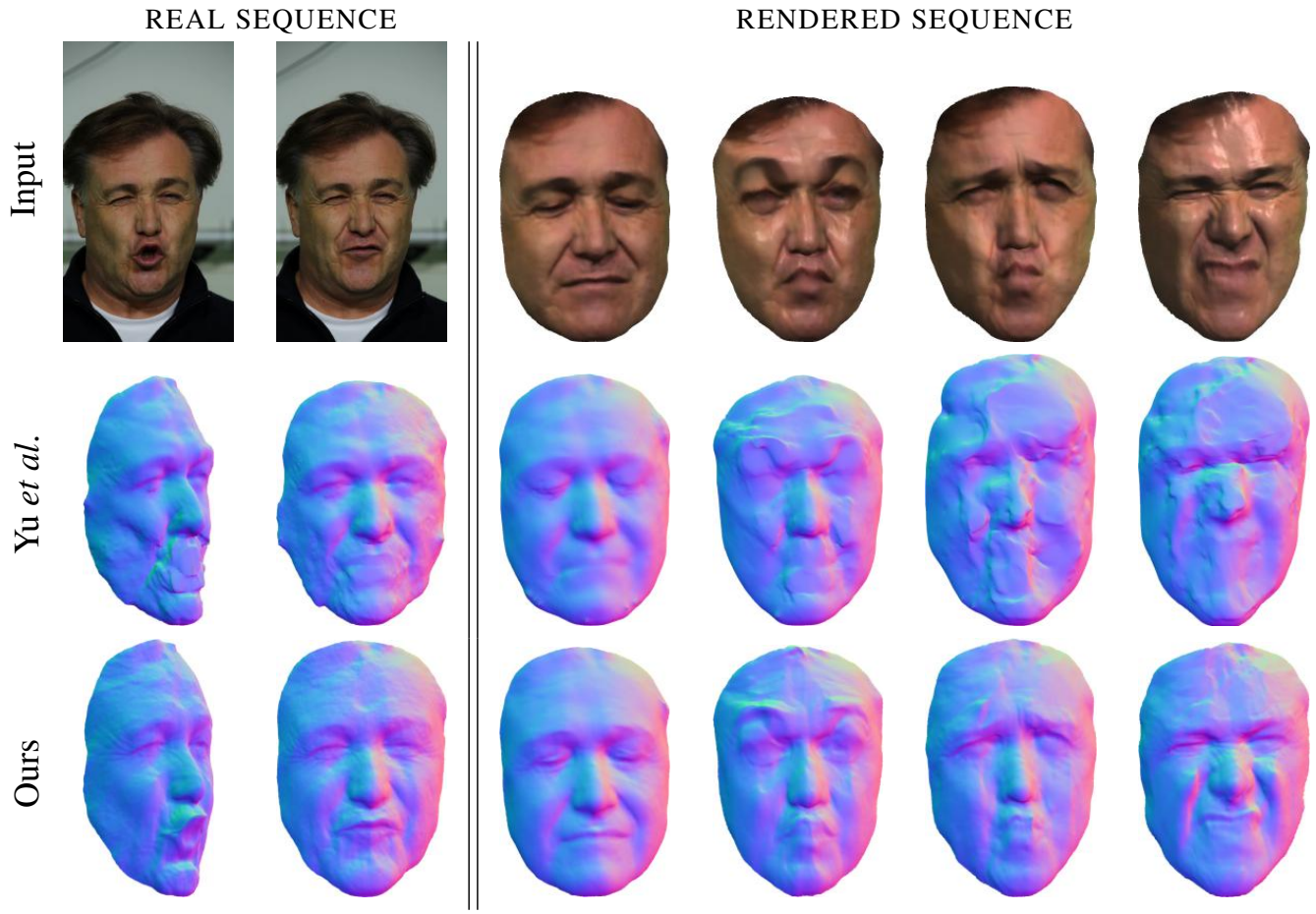

Figure 4: Best viewed in colour: A qualitative comparison of our method against that of Yu *et al*. [26], on (LEFT) the real-world sequence of [19] and (RIGHT) our four rendered sequences (from left to right: LF, SF, LC and SC).

frame. Table 1 shows the comparison results against the recent template-based method of Yu *et al*. [26] using their publicly available code. It can be seen that our approach is significantly better in all four cases, with or without modelling specularities. We reduce the baseline error by a factor of 220%-260% when the specular component is not estimated and around 240%-280% when it is. It should also be noticed how estimating the specular component improves the results for pure Lambertian sequences. This is due to the fact that the estimated specularities are also compensating for the errors in the initially computed albedo. In figure 5 it can be seen the decomposition of the results from two frames of the SC sequence and shows how our method can handle this challenging scenario with big changes on intensity. Our results can be best viewed in the accompanying video[1].

We further evaluate qualitatively on three real sequences: the original face from [19] (figure 4), the face from [26](figure 3), and new sequence of a hand deforming a ball (figure 6). In the case the first one, notice the improvement on the reconstructed deformation of the mouth thanks to our diffuse shading model while [26] is only able to recover a flat surface.

8 Conclusion

We have presented a principled approach jointly reasoning about non-rigid structure from motion and shape-from-shading, and provided strong empiric evidence that it is required

[1]Please visit http://www0.cs.ucl.ac.uk/staff/Qi.Liu/bmvc16/better_together.html to check video results and to access our publicly available code and datasets.

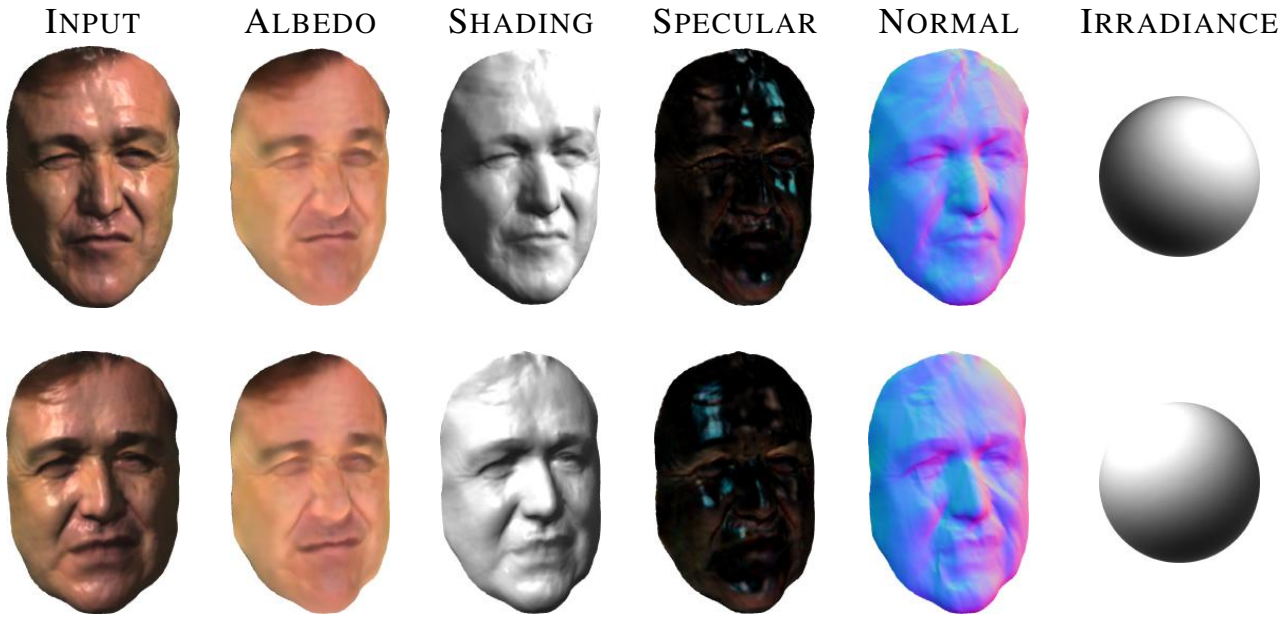

Figure 5: Results of two different frames from the SC sequence and the corresponding intrinsic decomposition.

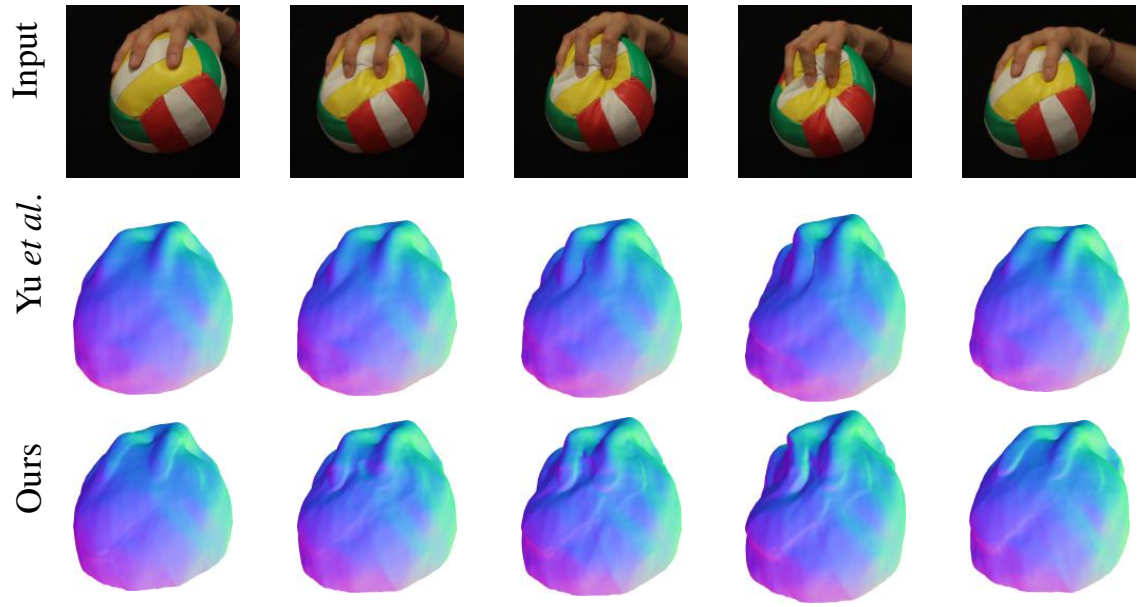

Figure 6: Comparison between the results of Rui *et al.* [26] and our method on a real sequence of a hand holding and deforming a ball.

to avoid systematic tracking failures, and that it significantly improves the reconstruction quality of fine semantic details. Although, we focused upon the challenging problem of reconstruction from a single RGB camera, such joint reasoning could readily be applied to RGB-D and multi-camera based approaches, and the increased robustness and detailed reconstructions it brings is likely to be of use to the wider community.

Acknowledgements

This work was partly supported by the SecondHands project, funded from the European Unions Horizon 2020 Research and Innovation programme under grant agreement No 643950. Qi Liu-Yin was funded by a UCL/Microsoft Research studentship. Chris Russell was funded by a UCL/BBC research fellowship.

References

[1] Sameer Agarwal, Keir Mierle, et al. Ceres solver. http://ceres-solver.org.

[2] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233, 2003.

[3] Thabo Beeler, Derek Bradley, Henning Zimmer, and Markus Gross. Improved reconstruction of deforming surfaces by cancelling ambient occlusion. In *12th European Conference on Computer Vision*, pages 30–43, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[4] Neill D. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, 2008.

[5] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi. 3d scanning deformable objects with a single rgbd sensor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] R. Or El, R. Hershkovitz, A. Wetzler, G. Rosman, A.M. Bruckstein, and R. Kimmel. Real-time depth refinement for specular objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 2013.

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

[9] Ira Kemelmacher-Shlizerman and Steven M Seitz. Face reconstruction in the wild. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1746–1753. IEEE, 2011.

[10] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007.

[11] Abed Malti, Adrien Bartoli, and Toby Collins. Template-based conformal shape-from-motion-and-shading for laparoscopy. In *Information Processing in Computer-Assisted Interventions*, 2012.

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

[13] Richard Newcombe, Dieter Fox, and Steve Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015.

[14] Ko Nishino, Zhengyou Zhang, and Katsushi Ikeuchi. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 599–606. IEEE, 2001.

[15] Roy Or El, Guy Rosman, Aaron Wetzler, Ron Kimmel, and Alfred M. Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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

[17] N. Sundaram, Thomas Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

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

[19] Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2012)*, volume 31, pages 187:1–187:11, November 2012.

[20] Aydin Varol, Appu Shaji, Mathieu Salzmann, and Pascal Fua. Monocular 3d reconstruction of locally textured surfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1118–1130, 2012.

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

[22] Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 287–296. ACM Press/Addison-Wesley Publishing Co., 2000.

[23] Changchang Wu. Visualsfm: A visual structure from motion system. http://ccwu.me/vsfm/, 2011.

[24] Yilei Xu and Amit K Roy-Chowdhury. Integrating the effects of motion, illumination and structure in video sequences. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1675–1682. IEEE, 2005.

[25] Lap-Fai Yu, Sai-Kit Yeung, Yu-Wing Tai, and Stephen Lin. Shading-based shape refinement of rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1422, 2013.

[26] Rui Yu, Chris Russell, Neil Campbell, and Lourdes Agapito. Direct, dense, and deformable: Non-rigid 3d reconstruction from rgb video. *ICCV*, 2015.

[27] 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.