Using Specularities to Boost Non-Rigid Structure-from-Motion

Agniva Sengupta
EnCoV-Institut Pascal
CNRS/Université Clermont Auvergne
Clermont-Ferrand, France
i.agniva+sengupta@gmail.com

Karim Makki
EnCoV-Institut Pascal
CNRS/Université Clermont Auvergne
Clermont-Ferrand, France
Karim.MAKKI@uca.fr

Adrien Bartoli
Department of
Clinical Research & Innovation,
CHU Clermont-Ferrand, France
adrien.bartoli@gmail.com

***Abstract*— Non-Rigid Structure-from-Motion (NRS*f*M) reconstructs the time-varying 3D shape of a deforming object from 2D point correspondences in monocular images. Despite promising use-cases such as the grasping of deformable objects and visual navigation in a non-rigid environment, NRS*f*M has had limited applications in robotics due to a lack of accuracy. To remedy this, we propose a new method which boosts the accuracy of NRS*f*M using sparse surface normals. Surface normal information is available from many sources, including structured lighting, homography decomposition of infinitesimal planes and shape priors. However, these sources are not always available. We thus propose a widely available new source of surface normals: the specularities. Our first technical contribution is a method which detects specular highlights and reconstructs the surface normals from it. It assumes that the light source is approximately localised, which is widely applicable in robotics applications such as endoscopy. Our second technical contribution is an NRS*f*M method which exploits a sparse surface normal set. For that, we propose a novel convex formulation and a globally optimal solution method. Experiments on photo-realistic synthetic data and real household and medical data show that the proposed method outperforms existing NRS*f*M methods.**[1] [2] [3]

I. Introduction

NRS*f*M is a challenging problem that has been extensively researched over the past two decades, leading to classical [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] and deep-learning based [11], [12], [13], [14], [15] methods. Unfortunately, their accuracy remains inadequate for many practical applications. A concrete way to improve accuracy is to exploit additional information from the image data, beyond the mere 2D point correspondences used by all NRS*f*M methods.

We propose to exploit the specularities, which are the highlights formed by the reflection of the light sources on the scene surface. The specularities are widespread in many domains. The main advantage of using normals from specularities is that they tend to occur at high surface curvature points where the chances of the local surface normal aligning with the viewing direction are higher [16], giving a sparse normal set that is more informative about the surface geometry than other surface normal sets. This thus deals with highly folded deformable surfaces reflecting a considerable number of small specularities. We propose a novel method to exploit them. As with all NRS*f*M methods, ours takes monocular images and 2D point correspondences as inputs. It then follows two steps: step one reconstructs independent surface normals from the images, and step two includes these surface normals in NRS*f*M.

Our first technical contribution addresses step one. We assume that the light source is approximately localised, for instance in light-equipped robots, photographic setups, mobile phones, and endoscopy. More formally, we introduce the Close Light Camera (CLC) setup, which includes a perspective camera and a point light source whose distance from the camera is small against the scene-camera distance. We propose a learning-based specularity detector and a geometry-based surface normal reconstruction method.

Our second technical contribution addresses step two. Existing NRS*f*M formulations and methods do not accommodate surface normal data. We thus propose a new formulation and solution method. A bigger challenge that we successfully address is to devise a convex formulation, which can be efficiently solved with global optimality. Our normal-boosted NRS*f*M method is general, as it is not bound to the CLC setup and may exploit surface normals from arbitrary sources. Figure 1 presents an overview of our proposed pipeline.

We provide extensive experimental validation including quantitative accuracy evaluation for synthetic data obtained from *Blender*, including baseline comparison, and validation on several real datasets.

II. Existing Work

We review existing work on specular normal reconstruction and NRS*f*M.

A. Specular normal reconstruction

Specularities have traditionally been considered as a nuisance [17], both in classical and learning-based reconstruction methods. In single-image reconstruction, [18] shows the important sensitivity of Shape-from-Shading (S*f*S) to specularities and [19] shows that depth perception suffers high uncertainty under specular reflection. In multi-image reconstruction, [20] shows their detrimental effect on establishing feature correspondences. The most common approach to deal with specularities is thus to discard them, by estimating specular masks and inpainting before running the reconstruction method. Owing to their omnipresence in contexts such as endoscopy, some recent work however proposed to exploit

[1] Code and dataset: encov.ip.uca.fr/ab/code_and_datasets

[2] Supplementary video: tinyurl.com/nrsfmspecularyoutube

[3] This work was funded by the FET-Open grant 863146 Endomapper.

them. In particular, [21], [22] detects elliptical specular blobs under the local planarity assumption and reconstruct the surface normal by rectifying the ellipse to a circle. This has strong limitations, as specularities tend to occur at high curvature points, breaking the planarity assumption [23]. While [21] uses an explicit fixed intensity to choose the isophotes, [22] uses a neural segmentation approach and fits the isophote curve to each mask component.

B. Non-Rigid Structure-from-Motion

Classical NRSfM methods models the scene by either low-rank shape bases [4], [3], [2] or by physics-based constraints such as isometry [6], [7]. The main challenge of NRSfM is its dependency on reliable point correspondences, which are challenging to establish in many contexts. Mitigation attempts were taken in physics-based NRSfM with shape priors [24], [25]; these priors are however unavailable in many cases. NRSfM from higher-order derivatives of the point correspondences [5], [8], [26] suffer from similar drawbacks. The performance of Deep NRSfM [11], whose architecture is derived from a classical sparse coding algorithm, also crucially depends on the point correspondences. Additionally, learning-based NRSfM methods [13], [27], [28] suffer from the domain-shift problem, making them unsuited to many robotic applications. Any additional available information should certainly be used to improve accuracy. Surface normals are one such important additional information. Unfortunately, none of the existing NRSfM methods make provision for including this additional sparse surface normal information.

III. Methods

We first give our independent specular normal reconstruction method and then our normal-boosted NRSfM method.

A. Specularity detection and normal reconstruction

1) General points: We first formalise the CLC setup, which, recall, hinges on the key assumptions that the camera performs perspective projection and that the light source is close to the camera. Our model of the CLC setup is thus a perspective camera and a point light source located at the camera centre [29]. We use an existing fully convolutional neural network trained in a supervised manner to segment the specularities [22]. We propose a new normal reconstruction method, called the *sightline-based reconstruction method*, which we combine with an existing *isophote-based reconstruction method* to form a new *combined reconstruction method*, illustrated in figure 2. Briefly, the isophote-based reconstruction method [21] assumes that the specularity has an elliptical shape in the image, which it recovers by ellipse fitting and then uses the pose-from-circle method [30] to reconstruct the normal.

2) Sightline-based normal reconstruction: In the CLC setup, the lighting direction is collinear with the sightline, which is the viewing direction of any scene point. In addition, this is also the direction of perfect mirror reflection when the surface normal is also collinear. Hence, the intuition we exploit is that, at a strongly specular pixel, the surface normal is given by the sightline, which can be directly estimated from the pixel coordinates and the camera model.

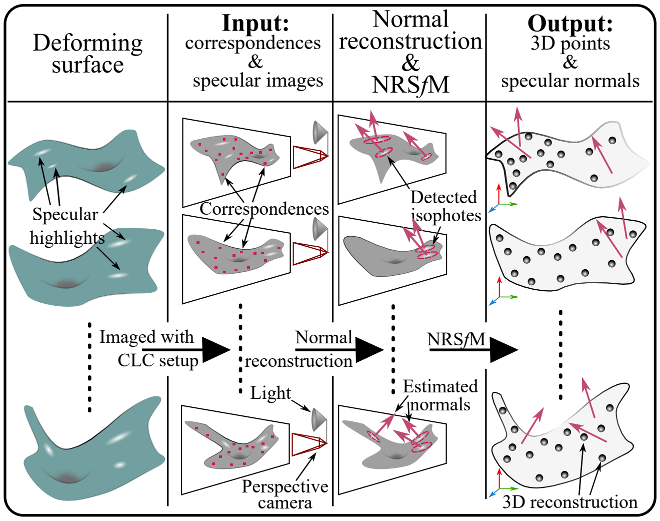

Fig. 1: Overview of the proposed normal-boosted NRSfM pipeline. The inputs are n images of a deforming surface under the CLC setup with m point correspondences. First, we detect the specularities, with a count of l_i for image $i \in [1, n]$, from which we reconstruct l_i surface normals. Second, we reconstruct n 3D point clouds, of size $m\,l_i$ for image $i \in [1, n]$, containing all input correspondences and normals.

Considering a single specular blob, which is a set of connected pixels detected as specular by the neural network, we propose to find the image point which has the highest intensity, called the Brightest Point (BP). Unfortunately, the camera generally saturates all over the specular blob pixels, which thus all have the maximum possible intensity. This makes the task of localising the BP difficult. We propose a method in three steps. First, we detect a level-set of the specular blob at some prescribed intensity value, using the marching squares algorithm. The resulting curve is named an *isophote*. The zero-intensity isophote is simply the segmentation boundary. Second, we smooth the isophote to cancel noise using a smoothing cubic B-spline, from which we sample isophote with a high number of points, typically 1000 points. Third, we determine the BP coordinates (x_0, y_0) as the median point of the isophote sample points. Given the BP, the sightline can be trivially found from the perspective camera model, giving the surface normal $\mathbf{n} \in S^2$ as $\mathbf{n} \propto \mathbf{K}^{-1}(x_0, y_0, 1)^\top$, where S^2 is the set of direction vectors in 3-space and $\mathbf{K} \in \mathbb{R}^{3\times3}$ holds the camera intrinsics.

3) Combined normal reconstruction: The existing isophote-based method has the advantage of using a whole isophote, bringing stability, but requires a locally flat surface. In contrast, the proposed sightline-based method does not require this hypothesis but only exploits a single point. The pros and cons are shown in table I. We propose a combined method exploiting the strength of both and introducing additional tests, shown in figure 2.

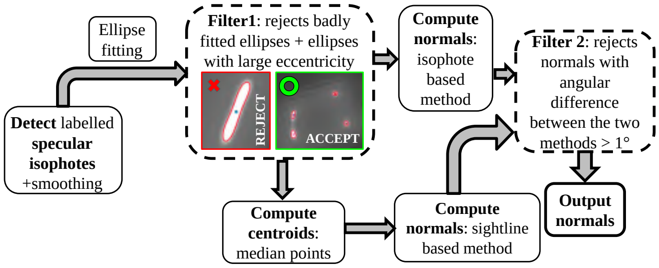
Ellipse fitting
Detect labelled specular isophotes +smoothing
Filter1: rejects badly fitted ellipses + ellipses with large eccentricity
REJECT
ACCEPT
Compute normals: isophote based method
Filter 2: rejects normals with angular difference between the two methods ≥ 1°
Compute centroids: median points
Compute normals: sightline based method
Output normals

Fig. 2: Proposed specular normal reconstruction pipeline.

First, we perform an *ellipticity* test for each candidate specular blob as filtering criterion. Formally, this uses a simple thresholding of the residual error of ellipse fitting and discard many unstable specular blobs. Second, we reject the blobs whose elliptical approximation have large *eccentricity*, *i.e.* close to 1, which are mainly due to neighbouring blobs erroneously merged in the segmentation. Third, we run both the isophote-based and sightline-based normal reconstruction methods and filter results by reconstruction *agreement*. In short, we only keep the normals whose estimate have an angular difference lower than a threshold, which we choose as 1 degree.

B. A convex method for normal-boosted NRSfM

We describe our normal-boosted NRS*f*M.

1) Problem setup: Given n images with m point correspondences across the images, NRS*f*M reconstructs the 3D position of these points. In the standard NRS*f*M setup, the image points are denoted by their image coordinates $\{\mathbf{p}_{i,j} \in \mathbb{R}^2\}$ for image $i \in [1, n]$ and correspondence $j \in [1, m]$. Visibility is modelled by binary indicators $\{\mathcal{V}_{i,j} \in \{0, 1\}\}$, where $\mathcal{V}_{i,j} = 1$ if $\mathbf{p}_{i,j}$ is visible and $\mathcal{V}_{i,j} = 0$ otherwise. Invisibility occurs when some points cannot be tracked owing to occlusions or tracking failure. The corresponding unknown 3D points are denoted by $\{\mathbf{P}_{i,j} \in \mathbb{R}^3\}$. The calibrated perspective camera is modelled by its intrinsics parameters in matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, with frames height H and width W pixels respectively. To this standard setup, we add the sparse surface normal information, modelled as $\{\mathbf{n}_{i,r} \in S^2\}$. In this notation, we have the image index $i \in [1, n]$ and the normal index $r \in [1, l_i]$, where l_i is the normal count in image i. These sparse normals may arise from various sources and modalities. We denote as $\{\mathbf{p}_{i,r} \in \mathbb{R}^2\}$ the image coordinates at which these normals are given.

TABLE I: Pros and cons of the normal reconstruction methods, where BP stands for Brightest Points.

Method	Pros	Cons
Sightline-based	• Low computational complexity	• Difficulty to find the BP
Isophote-based [21]	• No need for the BP • Handles occluded isophote and invisible BP	• Slight increase of computational complexity • Assumes local planarity • Difficulty to detect the isophote
Combined	• High accuracy • Stability	• Computational complexity • Reduces the number of reconstructed normals

2) Problem statement: The main difficulty of using the normals is that NRS*f*M reconstructs sparse points, as opposed to a surface, whereas the notion of normal is necessarily related to a surface. We give the key idea to tackle this difficulty and then its formal implementation as a convex cost function.

Key formulation idea. We introduce a piecewise planar representation, similar to the commonly used surface mesh representation. Our proposal however differs. Indeed, we could straightforwardly triangulate the point correspondences and force each prescribed normal to be collinear to the normal of its containing triangle, expressed in terms of the unknown 3D points. This, however, leads to a nonconvex formulation, stemming from the cross-product required to obtain the triangle's normal from its vertices. We propose instead to measure the orthogonality between each prescribed normals and each of the three edges of its containing triangle. This, fortunately, leads to a convex formulation. Let the image point where the normal is prescribed be $\mathbf{p}_{i,r_1}$ and the containing triangle vertices be $\mathbf{p}_{i,j_1}$, $\mathbf{p}_{i,j_2}$ and $\mathbf{p}_{i,j_3}$. The formulation is illustrated in figure 3 and stated formally in the next section.

Neighbourhood graphs. The proposed formulation requires two pieces of information: 1) the association of correspondences to form a graph-based neighbourhood structure, and 2) the association of correspondence pairs to prescribed normals. Point 1) can be solved as classically in NRS*f*M, using the Nearest-Neighborhood Graph (NNG) already required in NRS*f*M methods [6], [7]. The NNG is a graph connecting the image points which are consistently close within the image set, and thus likely to be close on the unknown reconstructed surface. Point 2) however requires a new structure, which we achieve as a modified NNG named the Nearest Surface-patch Graph (NS*p*G). The NNG is denoted by $\mathcal{N} \in \{0, 1\}^{m\times m}$ and computed using the method from [6], with k_{N} number of nearest neighbours sought for each image points. We have

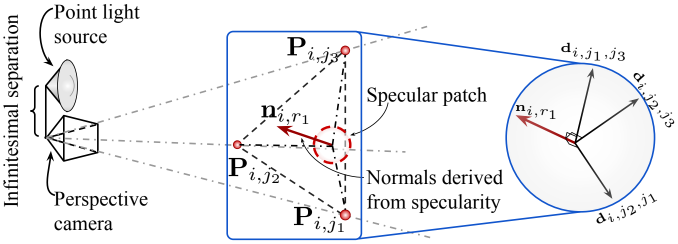

Fig. 3: Key idea to formulate normal-boosted NRS*f*M. For a prescribed normal $\mathbf{n}_{i,r_1}$, we maximise the orthogonality of each edge connecting its neighbouring keypoints. The figure shows the case for three arbitrary points, $(\mathbf{P}_{i,j_1}, \mathbf{P}_{i,j_2}, \mathbf{P}_{i,j_3})$. Our convex implementation maximises the orthogonality of $\mathbf{n}_{i,r_1}$ and the three edges, given as $\mathbf{d}_{i,j_x,j_y} = \mathbf{P}_{i,j_x} - \mathbf{P}_{i,j_y}, (x, y) \in \{1, 2, 3\}, x > y$. The proposed inner product based cost function given in section III-B penalises deviation from orthonormality. Importantly, it is convex and thus globally solvable.

$\mathcal{N}_{j,q} = 1$ if $\mathbf{p}_{i,j}$ and $\mathbf{p}_{i,q}$ are considered as neighbours, for all $i \in [1, n]$. The per-image NS*p*G is derived by associating $\{\mathbf{p}_{i,r}\}$ to $\{\mathbf{p}_{i,j}\}$ based on their consistent image proximity, denoted by $\mathcal{S}_i \in \{0, 1\}^{m \times m \times l_i}$. This has similarities to the method for deriving the NNG, with the difference that the NS*p*G is computed for each image separately as the normals are not tracked across the images. Our precise methods is given in algorithm 1. We have $\mathcal{S}_{i,j,q,r} = 1$ if the neighbouring point-pairs $(\mathbf{p}_{i,j}, \mathbf{p}_{i,q})$ are considered as neighbours with the normal at $\mathbf{p}_{i,r}$ for image $i \in [1, n]$.

Algorithm 1 Computing the NS*p*G for image i

```
Input: Correspondences $\{\mathbf{p}_{i,j}\}$, normals $\{\mathbf{p}_{i,r}\}$, NNG $\mathcal{N}$
Output: NSpG $\mathcal{S}_i$
  $\mathcal{S}_i \leftarrow \mathbf{0}_{m \times m \times l_i}$
  $\Delta \leftarrow \min(H, W)/\lambda$                    ▷ $\lambda$ is a tunable parameter
  for $r \in [1, l_i]$ do                  ▷ each prescribed normal in turn
    for $j \in [1, m]$ do                  ▷ each correspondence...
      for $q \in [1, m], \mathcal{N}_{j,q} = 1$ do       ▷ ...and neighbours
        if $d(\mathbf{p}_{i,r}, \mathbf{p}_{i,j}) \leq \Delta \wedge d(\mathbf{p}_{i,r}, \mathbf{p}_{i,q}) \leq \Delta$ then
          $\mathcal{S}_{i,j,q,r} \leftarrow 1$
```

Deformation model. We use a physics-based deformation model, specifically the inextensible model [6], [7], which is widely applicable to a broad category of surfaces. We denote by $\{g_{j,q} \in \mathbb{R}\}$ the unknown geodesic distances between $\mathbf{P}_{i,j}$ and $\mathbf{P}_{i,q}$ for all $i \in [1, n]$. Using a discrete approximation of the geodesic distance taken with inextensibility implies that the Euclidean distance between $\mathbf{P}_{i,j}$ and $\mathbf{P}_{i,q}$ must be lower or equal to $g_{j,q}$ across all images.

Parameterisation. The obvious method for parameterising an NRS*f*M problem is to represent 3D points $\{\mathbf{P}_{i,j}\}$ by their 3D coordinates. But this is an overparameterisation and is not tractable. The 3D points are actually restricted along the Sight-Line (SL), from the camera centre to the image point on the retina, $\mathbf{K}^{-1}\mathbf{p}_{i,j}$. Hence, the 3D points can be parameterised by just their depths along SL. We denote these unknown depths by $\{\delta_{i,j} \in \mathbb{R}\}$ and the known direction SL directions as $\{\mathbf{q}_{i,j} \in S^2\}$.

NRS*f*M without normals. We begin by recalling the NRS*f*M with the isometric-inextensible model from [6]:

$$\begin{aligned} &\underset{\{\delta_{i,j}\},\{g_{j,q}\}}{\arg\min} \; -\sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{V}_{i,j}\delta_{i,j} \\ \text{s.t.} \quad & \mathcal{V}_{i,j}\mathcal{V}_{i,q}\mathcal{N}_{j,q}\|\delta_{i,j}\mathbf{q}_{i,j} - \delta_{i,q}\mathbf{q}_{i,q}\| \leq g_{j,q}, \mathcal{V}_{i,j}\delta_{i,j} \geq 0, \\ & \sum_{j}\sum_{q} \mathcal{N}_{j,q} g_{j,q} = 1, \quad \forall\ i \in [1, n], \quad (j, q) \in [1, m]. \end{aligned} \tag{1}$$

The problem in eq. (1) is solved as a Second-Order Cone Programming (SOCP). This formulation, from [6], has three important aspects. Firstly, it has a cost that maximises the depth of all points. Depth maximisation follows the Maximum Depth Heuristic (MDH) proposed in [31], favouring a deeper solution to point depths to avoid convex-concave ambiguities, a well-known problem in NRS*f*M. Secondly, it has an inextensible-isometric constraint, which enforces the deformation model of our choice. Thirdly, there is the normalisation constraint which requires the geodesic distances to sum to unity. This is required as both the geodesic distances and point depths could otherwise freely scale while minimising and satisfying all costs and constraints respectively. Equation (1) also includes a positive depth constraint to avoid degenerate solutions.

	[5]	[4]	[3]	[2]	[7]	[6]	Ours
Sparse	0.41	0.28	0.50	0.31	0.78	0.28	**0.26**
Dense	0.41	0.26	0.49	0.31	0.58	0.22	**0.20**

TABLE II: Comparison of our proposed method with state-of-the-art approaches across two densities, sparse with $m = 40$ and dense with $m = 80$, repeated over 100 times, and the mean Root Mean Square Error (RMSE) are reported in the table above, all values in Arbitrary Units (au). The corresponding error histograms are in fig. 4

Normal-boosted NRS*f*M. We introduce an extra cost term to the formulation (1) to penalise the deviation of edges from orthogonality to their nearest sparse normal in the NS*p*G:

$$\begin{aligned} &\underset{\{\delta_{i,j}\},\{g_{j,q}\}}{\arg\min} \\ &\lambda_{\text{S}} \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{q=1}^{m}\sum_{r=1}^{l_i} \mathcal{Q}_{i,j,q,r} |\langle \delta_{i,j}\mathbf{q}_{i,j} - \delta_{i,q}\mathbf{q}_{i,q}, \mathbf{n}_{i,r}\rangle| \\ &\qquad - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{V}_{i,j}\delta_{i,j} \\ \text{s.t.} \quad & \mathcal{V}_{i,j}\mathcal{V}_{i,q}\mathcal{N}_{j,q}\|\delta_{i,j}\mathbf{q}_{i,j} - \delta_{i,q}\mathbf{q}_{i,q}\| \leq g_{j,q}, \mathcal{V}_{i,j}\delta_{i,j} \geq 0, \\ & \sum_{j}\sum_{q} \mathcal{N}_{j,q} g_{j,q} = 1, \quad \forall\ i \in [1, n], \quad (j, q) \in [1, m], \end{aligned} \tag{2}$$

where $\mathcal{Q}_{i,j,q,r} = \mathcal{V}_{i,j}\mathcal{V}_{i,q}\mathcal{N}_{j,q}\mathcal{S}_{i,j,q,r}$. Importantly, this formulation remains an SOCP, which is a convex problem, that we solve by off-the-shelf solvers [32], [33].

IV. Experimental Results

We present our experimental validation. We begin with synthetic data. We compare our method with baseline approaches using challenging deformations using the deformable paper model of [34]. We follow up by qualitative evaluation on real data. The first dataset uses a handheld mobile device with the flash-gun turned filming a deforming paper. The second dataset is a colonoscopic video from the EndoMapper [35] database.

A. Synthetic data

We begin with a brief description of our simulation setup implemented in Blender followed by validation of, first, the accuracy of estimation of normals from specularity, and, second, the accuracy of normals boosted NRS*f*M, the normals being obtained from both the simulated 3D model and from actual specularities.

Blender simulation. We synthesise images using a wide-angled perspective camera of 10 mm focal length with a simulated spotlight light source collocated with the camera center, the spotlight has a radius of 0.1 m and projected to a $120°$ cone. A deforming surface is generated with free-form deformation [36], the surface is imparted with a strong 'clearcoat' to boost specular reflections. To generate a realistic shading, we use the anisotropic principled Bidirectional Reflectance Distribution Function (BSDF) model, popularised by Disney [37], [38], mimicking real-world specularities on the surface of the deforming object. The generated surface has 10404 vertices and 20402 faces and deforms over 250 frames. We implement a data-subsampler that picks up n random frames from this sequence, where $n \ll 250$, and uses the projected vertices of the mesh to generate m random tracks of correspondences across these n frames, where $m \ll 10404$.

1) Evaluation of normal reconstruction from specularities: We evaluate both qualitatively and quantitatively, the normal reconstruction method introduced in [21], [22] on highly curved surfaces for which ground truth normals are available (see Figure 5 for some qualitative results). Six Blender simulated sequences with different spotlight diameters have been used. Each sequence contains about 1400 specular blobs arbitrarily distributed across 250 images of the deforming plane. The ratio of useful specular blobs leading to accurate normal estimation for each sequence is given in Figure 6c. From which we notice that increasing the spotlight radius leads to a decrease of useful blobs which is mainly due to an increase in number of merged elliptical blobs that have been discarded by the first filter because of their large eccentricity, as illustrated in Figure 2.

Despite the fact the illumination model did not follow the Phong model, which was the baseline in [21], the obtained results confirmed that specularities should be elliptical blobs in the CLC setup. Furthermore, the distance between ellipse centers and the isophote centroid (its median point) are too close in the case of accurately estimated normals. A good approximation of specularity's brightest point could be one of them.

An accurate estimation of a large number of useful normals follows. This set of normals is selected by ensuring a good agreement between two different methods (with an angular error less than $1°$), showing a high accuracy compared to the ground truth. Figure 6 shows quantitative results for the evaluation of normal estimation, defining the ratio of useful specularites as the total number of detected elliptical blobs divided by the total number of specular blobs including saturated ones.

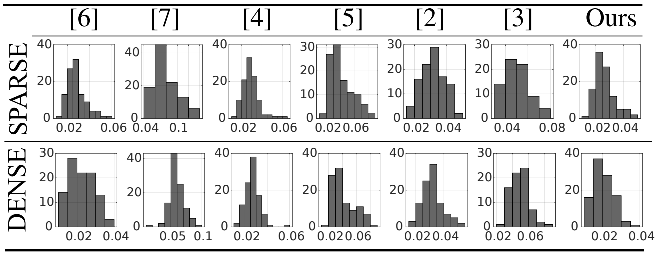

Fig. 4: Histogram of RMSE comparing our proposed approach to baseline methods over 100 repeated experiments, the mean values are shown in table II.

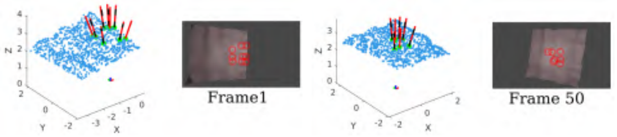

Fig. 5: Evaluation of normal reconstruction from specularities in synthetic data. Qualitative results for two different poses are presented. 3D renderings show ground truth normals in red while the estimated normals are represented by black dashed lines.

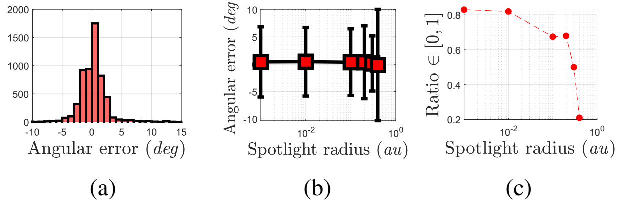

Fig. 6: Quantitative evaluation of normal estimation error: (a) histogram for angular difference in orientation for all the sequences combined, (b) error bars for all the sequences, and (c) ratios of useful specular blobs. The ratio is obtained by dividing the number of useful blobs by the total number of specular blobs (uniformly for all sequences).

2) Evaluation of NRSfM reconstruction: First, we provide an initial validation of our NRSfM method proposed in section III-B. Using the Blender simulated data of a deforming planar shape, as in section IV-A.1, we vary the number of sampled feature correspondences m and the number of sparse normals used $s = l_i$ in our NRSfM, given that l_i is held constant across all the frames. At this step, we use synthetic normals available from the mesh model, not restricting ourselves to normals from specularity, to better characterise our reconstruction pipeline. The quantitative results are shown in fig. 7, demonstrating that increasing s indeed results in higher reconstruction accuracy, validating

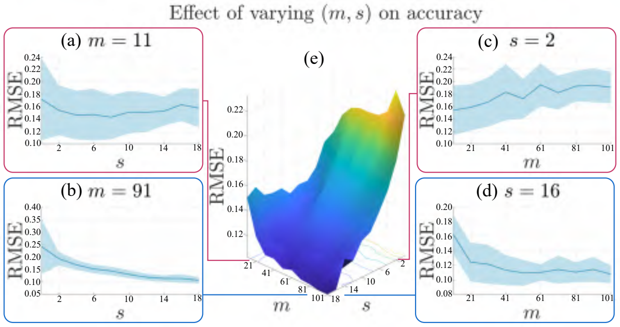

Fig. 7: Reconstruction accuracy. (a) and (b) show the effect of increasing s for $m = 11$ and $m = 91$, (c) and (d) show the effect of increasing m for $s = 2$ and $s = 16$.

our approach, the improvements being more pronounced when correspondences are denser. This is intuitive since a large number of sparse normals with few keypoints will inevitably result in some incorrect assignment of NS*p*G neighbours.

Next, we simulate challenging deformable surfaces by following the paper model proposed in [34]. We simulate deformable surfaces with two different densities of correspondences, the *sparse* data with $m = 40$ and the *dense* data with $m = 80$. The baseline is maintained uniformly at $n = 7$ and all frames are assigned the same number of $s = 7$ specular normals, randomly assigned from the groundtruth surface. The presence of accurate groundtruth allows us to compare our method with baseline approaches, including [5], [6], [2], [3], [4], and [7]. We repeat each randomised experiment 100 times, the resulting histogram of RMSE, expressed in terms of the au, is shown in fig. 4. The accuracy of our proposed method is ahead of all other compared methods, [4] and [6] being the closest competitors. But the mean value of all the RMSE across the 100 experiments, shown in table II, confirms our advantage over all methods.

B. Real data

We now present our results on real data. The real data used by us do not have groundtruth, so we present qualitative results and analysis.

1) Colonoscopy: We use an image sequence extracted from the Endomapper Dataset [35], composed of 37 images from [24], each of resolution 1248 × 1080. A total of about 8000 elliptical specular blobs (and therefore useful normals) were detected. An angular error less than 1° between the normals estimated using methods 1 and 2 is obtained showing a very good agreement between the two methods and confirming the correctness of these estimated normals. We extracted 10 equally spaced imaged from this dataset and upgraded the point correspondences to $m = 84$. Combined with the estimated specular normals, then NRS*f*M achieved good reconstruction accuracy in the region with correspondences. The reconstructed 3D points are densified with [39] and texturised, resulting in the qualitative images of fig. 8. Qualitative comparison with baseline methods has been presented for one randomly sampled frame in fig. 10, our approach is the only one capable of recovering the conical/trough shape of the visible segment of the colon.

2) Mobile phone data: A set of 13 images of a paper sheet reflecting specularities is captured using a Samsung Galaxy A14 mobile phone camera. Acquisitions are performed in the dark with the flash gun activated. The paper sheet has been deformed by hand with no constraints on the camera viewing angles. A sparse set of 53 labeled keypoints are tracked (semi-automatically, which is a standard practice in non-robust NRS*f*M) throughout the entire sequence and then used as input for the normal-boosted NRS*f*M. These landmarks are reasonably evenly distributed across the deforming plane surface. We give some qualitative results for the reconstruction in figure 9. Qualitative comparison with two other baseline methods has been presented for one randomly sampled frame in fig. 10, all the other baseline methods failed to produce results on this dataset. The proposed method takes 2.01 seconds for solving NRS*f*M on this dataset, a small increase over the 1.65 seconds taken by [6].

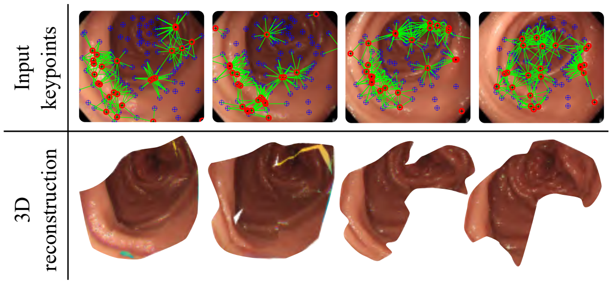

Fig. 8: Results from reconstruction of the colonoscopic sequence. Note that the reconstructed patch becomes smaller in the last two frames since the correspondences track a smaller region of the colon. The blue markers are feature correspondences, the black/red markers are specularity centroids and the green lines are the edges of NS*p*G.

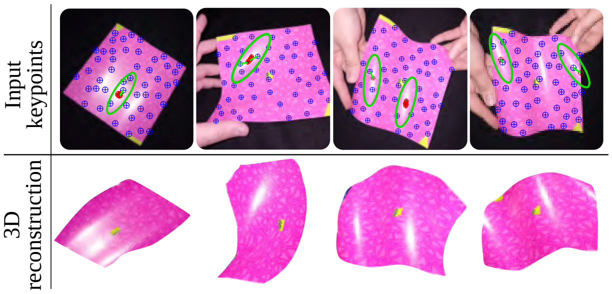

Fig. 9: Reconstruction of the paper surface captured with a handheld mobile device with flash gun turned on. The approximate region of detected specularities are in green.

V. Conclusion

We presented a novel method which exploits prescribed sparse normals in NRS*f*M. We show how such surface normals can be automatically obtained from specular highlights, with mild hypotheses on the light source. Our experiments show that the proposed method outperforms in synthetic and real data. We plan to investigate the use of additional prescribed normal sources and to develop deformable object grasping. Eventually, our method forms a strong opportunity to combine NRS*f*M and S*f*S.

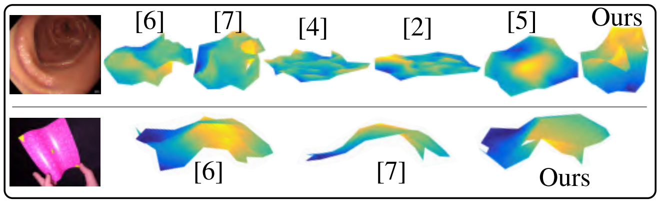

Fig. 10: Comparison of our proposed NRS*f*M method with baseline approaches over one randomly sampled frame from the colonoscopy and mobile phone data

References

[1] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Robust isometric non-rigid structure-from-motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6409–6423, 2021.

[2] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014.

[3] Paulo FU Gotardo and Aleix M Martinez. Kernel non-rigid structure from motion. In *2011 International Conference on Computer Vision*, pages 802–809. IEEE, 2011.

[4] Onur C Hamsici, Paulo FU Gotardo, and Aleix M Martinez. Learning spatially-smooth mappings in non-rigid structure from motion. In *European Conference on computer vision*, pages 260–273. Springer, 2012.

[5] Ajad Chhatkuli, Daniel Pizarro, and Adrien Bartoli. Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity. In *British Machine Vision Conference*, 2014.

[6] Ajad Chhatkuli, Daniel Pizarro, Toby Collins, and Adrien Bartoli. Inextensible non-rigid structure-from-motion by second-order cone programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2428–2441, 2017.

[7] Pan Ji, Hongdong Li, Yuchao Dai, and Ian Reid. " maximizing rigidity" revisited: a convex programming approach for generic 3d shape reconstruction from multiple perspective views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 929–937, 2017.

[8] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Isometric non-rigid shape-from-motion in linear time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4679–4687, 2016.

[9] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Local deformable 3d reconstruction with cartan's connections. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):3011–3026, 2019.

[10] Shaifali Parashar, Yuxuan Long, Mathieu Salzmann, and Pascal Fua. A closed-form solution to local non-rigid structure-from-motion. *arXiv preprint arXiv:2011.11567*, 2020.

[11] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1558–1567, 2019.

[12] Ayça Takmaz, Danda Pani Paudel, Thomas Probst, Ajad Chhatkuli, Martin R Oswald, and Luc Van Gool. Unsupervised monocular depth reconstruction of non-rigid scenes. *arXiv preprint arXiv:2012.15680*, 2020.

[13] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019.

[14] Matteo Pedone, Abdelrahman Mostafa, and Janne Heikkilä. Learning-based non-rigid video depth estimation using invariants to generalized bas-relief transformations. *Journal of Mathematical Imaging and Vision*, 64(9):993–1009, 2022.

[15] Erik Johnson, Marc Habermann, Soshi Shimada, Vladislav Golyanik, and Christian Theobalt. Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6597–6606, 2023.

[16] Andrew Zisserman, Peter Giblin, and Andrew Blake. The information available to a moving observer from specularities. *Image and vision computing*, 7(1):38–42, 1989.

[17] Liang Li, Evangelos Mazomenos, James H Chandler, Keith L Obstein, Pietro Valdastri, Danail Stoyanov, and Francisco Vasconcelos. Robust endoscopic image mosaicking via fusion of multimodal estimation. *Medical Image Analysis*, 84:102709, 2023.

[18] Gastone Ciuti, Marco Visentini-Scarzanella, Alessio Dore, Arianna Menciassi, Paolo Dario, and Guang-Zhong Yang. Intra-operative monocular 3d reconstruction for image-guided navigation in active locomotion capsule endoscopy. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 768–774. IEEE, 2012.

[19] Javier Rodriguez-Puigvert, David Recasens, Javier Civera, and Ruben Martinez-Cantin. On the uncertain single-view depths in colonoscopies. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–140. Springer, 2022.

[20] Rema Daher, Francisco Vasconcelos, and Danail Stoyanov. A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence. *Medical Image Analysis*, 90:102994, 2023.

[21] Karim Makki and Adrien Bartoli. Normal reconstruction from specularity in the endoscopic setting. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[22] Karim Makki, Kilian Chandelon, and Adrien Bartoli. Elliptical specularity detection in endoscopy with application to normal reconstruction. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–6, 2023.

[23] Karim Makki and Adrien Bartoli. Reconstructing the normal and shape at specularities in endoscopy. *arXiv preprint arXiv:2311.18299*, 2023.

[24] Agniva Sengupta and Adrien Bartoli. Colonoscopic 3d reconstruction by tubular non-rigid structure-from-motion. *International Journal of Computer Assisted Radiology and Surgery*, 16(7):1237–1241, 2021.

[25] Agniva Sengupta and Adrien Bartoli. Totem nrsfm:

Object-wise non-rigid structure-from-motion with a topological template. *International Journal of Computer Vision*, pages 1–42, 2024.

[26] Shaifali Parashar, Daniel Pizarro, and Adrien Bartoli. Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2442–2454, 2017.

[27] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.

[28] Faisal Khan, Saqib Salahuddin, and Hossein Javidnia. Deep learning-based monocular depth estimation methods—a state-of-the-art review. *Sensors*, 20(8):2272, 2020.

[29] Takayuki Okatani and Koichiro Deguchi. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Computer vision and image understanding*, 66(2):119–131, 1997.

[30] David Forsyth, Joseph L Mundy, Andrew Zisserman, Chris Coelho, Aaron Heller, and Charlie Rothwell. Invariant descriptors for 3D object recognition and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):971–991, 1991.

[31] Mathieu Salzmann and Pascal Fua. Linear local models for monocular reconstruction of deformable surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):931–944, 2010.

[32] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[33] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[34] Mathieu Perriollat and Adrien Bartoli. A computational model of bounded developable surfaces with application to image-based three-dimensional reconstruction. *Computer Animation and Virtual Worlds*, 24(5):459–476, 2013.

[35] Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O León Barbed, Javier Morlana, David Recasens, Víctor M Batlle, Juan J Gómez-Rodríguez, et al. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data*, 10(1):671, 2023.

[36] Pushkar Joshi, Mark Meyer, Tony DeRose, Brian Green, and Tom Sanocki. Harmonic coordinates for character articulation. *ACM Transactions on Graphics (TOG)*, 26(3):71–es, 2007.

[37] Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez. Practical physically-based shading in film and game production. In *ACM SIGGRAPH 2012 Courses*, pages 1–7. 2012.

[38] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, volume 2012, pages 1–7. vol. 2012, 2012.

[39] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Cláudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics*, 5(4):349–359, 1999.