Improving 3D Surface Reconstruction from Endoscopic Video via Fusion and Refined Reflectance Modeling

Rui Wang*[a], True Price[a], Qingyu Zhao[a], Jan-Michael Frahm[a], Julian Rosenman[b,a], Stephen Pizer[a,b]
[a]Dept. of Computer Science, [b]Dept. of Radiation Oncology; Univ. of North Carolina at Chapel Hill

ABSTRACT

Shape from shading (SFS) has been studied for decades; nevertheless, its overly simple assumptions and its ill-conditioning have resulted in infrequent use in real applications. Price et al. recently developed an iterative scheme named shape from motion and shading (SFMS) that models both shape and reflectance of an unknown surface simultaneously. SFMS produces a fairly accurate, dense 3D reconstruction from each frame of a pharyngeal endoscopic video, albeit with inconsistency between the 3D reconstructions of different frames. We present a comprehensive study of the SFMS scheme and several improvements to it: (1) We integrate a deformable registration method into the iterative scheme and use the fusion of multiple surfaces as a reference surface to guide the next iteration's reconstruction. This can be interpreted as incorporating regularity of a frame's reconstruction with that of temporally nearby frames. (2) We show that the reflectance model estimation is crucial and very sensitive to noise in the data. Moreover, even when the surface reflection is not assumed to be Lambertian, the reflectance model estimation function in SFMS is still overly simple for endoscopy of human tissue. By removing outlier pixels, by preventing unrealistic BRDF estimation, and by reducing the falloff speed of illumination in SFS to account for the effect of multiple bouncing of the light, we improve the reconstruction accuracy.

1. INTRODUCTION

1.1. Problem and proposed solution

Endoscopy is an in-body examination method that provides direct and high-resolution visualization of human organs to physicians. However, due to large camera distortion and the lack of spatial reference, it is difficult for a physician to interpret the 3D geometry and position of an object of interest, thus limiting the usefulness of endoscopic video in treatment planning. In addition, due to the large amount of redundant information, lack of an efficient method to do video-based comparison, and most importantly inability to provide a full view of the target object, endoscopic video is almost never used for review.

Price et al. [1] developed an algorithm for shape-from-motion-and-shading (SFMS) that can reconstruct a textured interior tissue surface in 3D from each endoscopic video frame, and Zhao et al. [2] developed a group-wise surface registration algorithm that can fuse such single-frame-based 3D textured geometries into one complete textured surface. Through an overall pipeline described in [2], an endoscopic video that contains many flat and redundant views is transferred into a single complete 3D textured surface that we call the *endoscopogram*. The model provides (1) complete 3D anatomical geometry, which facilitates tumor localization; (2) efficient visualization, which provides a full overview of the scoped area and provides comparison within and between patients; and (3) the opportunity to register endoscopy data with other modalities, such as CT, thereby enabling transfer of the tumor information into CT spaces for treatment planning. Price et al. and Zhao et al. applied their work to pharyngeal endoscopic videos and showed success.

However, their combined method is still far from perfect: (1) Since the reconstruction method of Price et al. is frame-by-frame, there are no temporal constraints between successive images, which leads to inconsistent reconstructions and even failure to reconstruct some frames. In addition, due to such inconsistency, very few partial surface reconstructions can be selected for fusion. (2) One reason that Price's method creates more realistic reconstruction than many other state-of-the-art endoscopic reconstruction techniques [3,[4] is that it uses a reflectance model that is more sophisticated than simple Lambertian reflectance. However, this model is still not powerful enough to characterize the complex inner body environment with liquids causing specularities, absorption by multilayer tissues, and multiple bouncing of the light.

In this work, we focus on solving the inconsistency problem of the SFMS method so that longer sequences can be fused together. In addition, we further explore the reflection model estimation module and refine it so as to produce a more

Medical Imaging 2017: Image Processing, edited by Martin A. Styner, Elsa D. Angelini, Proc. of SPIE
Vol. 10133, 101330B · © 2017 SPIE · CCC code: 1605-7422/17/$18 · doi: 10.1117/12.2254587

accurate reconstruction result. We use the same pharyngeal phantom and evaluation scheme as in Price's paper to show the improvement of reconstruction accuracy. We also apply the improved SFMS on colonoscopic video to show its improved reconstruction consistency in this new domain.

1.1 Related works

A variety of partial solutions to the difficult problem of 3D reconstruction of endoscopic video have been published. The majority of them are based on two types of computer vision techniques: shape-from-shading [9,10,11] and structure-from-motion [11,12]. SFS is useful for 3D reconstruction of the endoscopic video because (1) it operates on single image and is thus unaffected by tissue deformations from frame to frame; (2) it gets simplified by the fact that the light source and the camera are co-located in the endoscopic environment. SFM can produce accurate 3D reconstruction by matching salient feature points from multiple different views and more importantly estimate the camera positions for each frame, which is useful for fusing multiple 3D reconstructions into a whole. The complementary properties of these two methods make the combination of them a natural choice for medical vision tasks. Kaufman et al. [3] use SFS to reconstruct a 3D surface from the endoscopic image and then use 2D feature points to find a transformation matrix to align consecutive 3D surfaces. Malti et al. [15],[16] use SFM to build a rigid 3D template of the scene and then use such template together with SFS to refine the existing surface. Wu et al. [4] proposed a multi-frame SFS algorithm that can obtain consistent and complete shape reconstruction by leveraging trackers in the endoscope and identifying common occlusion boundaries across different frames. However, their method only works for rigid objects with a surface exhibiting Lambertian reflectance. Besides these SFS and SFM-based methods, Hong et al. [6] proposed a virtual colon reconstruction scheme based on identifying pre-designed features. In Nadeem et al.'s work [5], machine learning techniques are used to estimate depth from a single endoscopic image. They use a virtual endoscopic dataset as training data and applied it on real images. However, none of these methods can produce a full 3D reconstruction from endoscopic video with a poorly known shape prior, arbitrary surface reflectance, and large tissue deformation.

2. METHODS

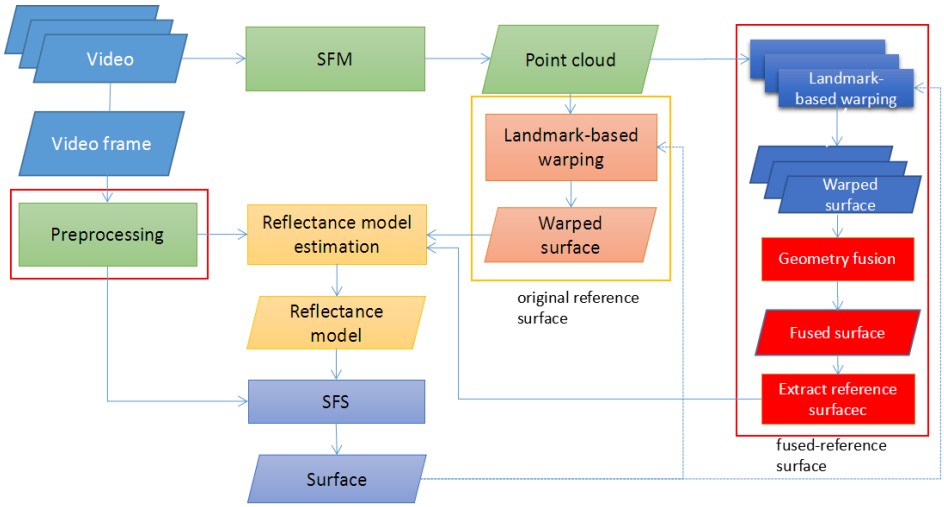


Figure 1. Original SFMS and the proposed improved SFMS pipeline for single frame reconstruction. Modifications are shown in red boxes. The underlying algorithms of SFS and SFM are detailed in [1].

Figure 1 shows the original [1] as well as the improved SFMS pipeline. New contributions are outlined in red. In summary, starting from sparse point clouds obtained by structure-from-motion, the overall pipeline iteratively estimates both the reflectance model and the shape of an unknown object, in this case the anatomical surface viewed by an endoscope. This is an EM-type problem, to estimate the latent data (depth) and a vector of unknown parameters (reflectance model), given observed data (the input image) and a sparse surface representation obtained via structure-from-motion (SFM) [7] from a few nearby frames. By utilizing data from SFM, SFMS is able to perform guided per-frame reconstruction without relying

on a pre-estimated reflectance model. However, due to a lack of temporal constraints, the method often induces inconsistent reconstructions across successive individual frames.

2.1 Improving shape reconstruction and reflectance model estimation by fusion

As mentioned before, one of the problems that we are addressing in this paper is the inconsistency of the frame-by-frame SFMS method. We assume that even the tissues in endoscopic video are deformable, in adjacent frames they should still be relatively close to each other. Therefore, we expect the reconstructed surfaces from adjacent frames to have small deformations from each other. In other words, we want the reconstruction to be consistent across time. In the original SFMS method, each frame uses its own SFM points to generate a warped surface, which is used as a prior shape for reflectance model estimation. During the iterative reconstruction, the method makes no interconnection between different frames. Therefore, initialization errors easily lead to different reconstruction results. We solve this problem by introducing a *fused reference surface*. This fused reference surface can be seen as a summary estimation from multiple frames, which is more robust than a single frame estimation. In addition, by leveraging the deformable registration and outlier geometry trimming in the geometry fusion [2], this fused reference surface is much more reliable than a simple average. Finally, all the frames use this fused reference surface to estimate their reflectance models and guide the SFS reconstruction as well. Our experimental result shows that this fused-reference surface provides not only more consistent but also more accurate geometry for each frame.

The modified algorithm is as follows. Lines in boldface indicate the new contributions.

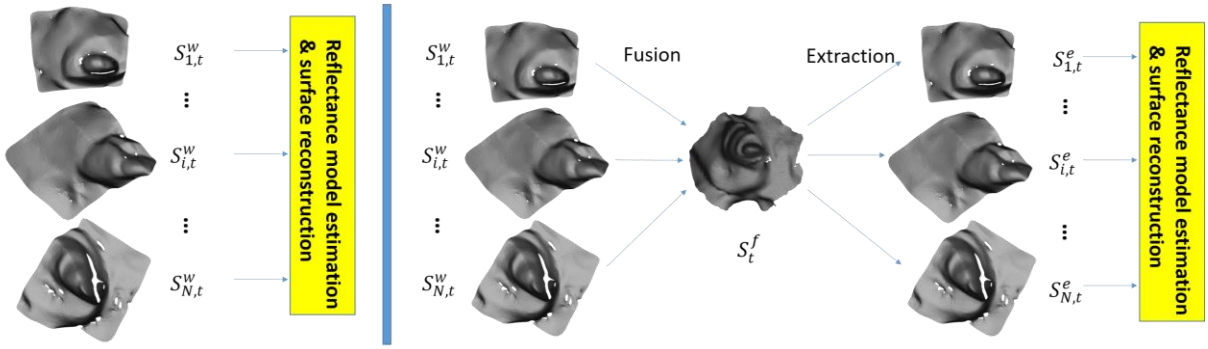
Algorithm 1 Fusion-guided SFMS

Input: A sequence of endoscopic video frames $\{F_i | i = 1 \dots N\}$

1: Generate a sparse 3D point cloud P and camera positions $C_{i,t}$ from the input frames using SFM
2: Initialize estimated surface of each frame with constant depth
3: Warp the estimated surface $S_{i,t}^{w}$ using its corresponding SFM 3D points P_i
4: Fuse the warped surfaces into a fused reference surface $\boldsymbol{S_t^f}$
5: For each frame F_i
6: Extract a reference surface $S_{i,t}^{e}$ from the fused reference surface $\boldsymbol{S_t^f}$
7: Warp the reference surface $S_{i,t}^{e}$ using its corresponding SFM 3D points P_i
8: **Remove saturated and under-illuminated pixels $\boldsymbol{F_i'}$**
9: Estimate the reflectance model *BRDF* using the extracted reference surface $S_{i,t}^{e}$ and the preprocessed image F_i'
10: Perform SFS to generate a better estimate surface $S_{i,t+1}^{w}$
11: Repeat steps 3-10 until convergence

In this algorithm the subscript i indicates the frame or camera index and t is the iteration index. The superscript f indicates the fused reference surface, w is the warped surface, and e is the extracted surface. A sequence of endoscopic video frames $\{F_i | i = 1 \dots N\}$ is the only input to our system. At step 1, a sparse 3D point cloud P is generated using a software named *Colmap* [7]. *Colmap* implements a structure-from-motion (SFM) algorithm that simultaneous estimates both camera pose and 3D scene structure from multiple frames. In our system, the point cloud P is used as a prior for reflectance model estimation and surface reconstruction.

In comparison to the original SFMS, where $S_{i,t}^{w}$ (step 3) is directly used for reflectance model estimation (step 9) and surface reconstruction (step 10), our fusion-guided SFMS uses a single fused reference surface S_t^f. That surface is generated by fusing all warped surfaces $\{S_{i,t}^{w} | \, i = 1 \dots N\}$ at iteration t using Zhao's [2] registration method (step 4). Since each endoscopic image is taken at a different time, such fusion provides temporal regularity across all the frames. Figure 2 shows an example of $S_{i,t}^{w}$ and S_t^f. We could directly incorporate temporal regularity into the SFS equation by computing optical flow between successive fames, but that would result in an extremely complex optimization system. Separating the temporal regularity and SFS makes the overall problem more solvable and stable.



(a) Frame-by-frame reference surface (b) Fused reference surface

Figure 2. In the original pipeline, the reference surface is generated for each reconstruction separately. In fusion-guided SFMS, the fused reference surface generated using the deformable registration is shared by all reconstructions by extracting a surface visible only to each specific camera pose from that fused reference surface.

Step 6 illustrates how the fused reference surface is being used. Given camera position $C_{i,t}$, obtained from SFM (step 1), the corresponding surface $S^e_{i,t}$ that is visible to $C_{i,t}$ is extracted from S^f_t as the initial guidance surface for the following reflectance model estimation and surface reconstruction. Since SFM points are treated as ground-truth, a warping is performed in step 7 to ensure that the reference surface $S^e_{i,t}$ won't deviate too much from those points.

The SFS equation is a parameterization of radiance of light I_r reaching the observer and irradiance of light I_i hitting the surface.

$$BRDF(\theta) = \frac{I_r}{I_i},\ I_i = I\frac{A}{r^2}cos\theta \tag{1}$$

The *BRDF* here is a function with only one variable, the detailed derivation can be found in [1]

$$BRDF(\theta) = \sum_{k=0}^{K-1}\left(\alpha_k + \beta_k \sin\left(\frac{\theta}{2}\right)\right)\cos^k\theta \tag{2}$$

Given the extracted reference surface $S^e_{i,t}$, the distance r to the light source, and the angle between the incident light and surface normal θ can easily be computed. Thus estimating the *BRDF* becomes solving a linear system with *2K* parameters.

More details about the SFM, reflectance model estimation, and SFS are provided in [1]. More details about the group-wise deformable surface registration algorithm used for fusion are provided in [2].

2.2 Improving reflectance model estimation by outlier removal and approximation of multiple light bouncing

Many assumptions and constraints are needed for SFS to be solvable. Among those assumptions, Lambertian surface reflection is one of the most popular. In [1], Price et al. proposed a more flexible reflectance model (equation 2) for modeling the surface in endoscopic environment, which is suitable for any kind of surface property. Furthermore, the reflectance model estimation process is simplified by utilizing SFM points as prior information, and the co-location of the light and camera.

Price's reflectance model estimation uses a linear regression yielding the BRDF coefficients ω given the reflectance model X and image y. This regression is sensitive to noise. In the original SFMS, the whole frame is used to estimate the reflectance model. However, saturated and under-illuminated pixels do not provide much useful information on surface depth. Such pixels can easily be filtered out using a predefined threshold. Doing so prevents corruption of the reflectance model by these outliers. In addition, because large BRDF coefficients are unrealistic, we also introduce a term preferring small coefficients ω in estimation of the reflectance model, thus improving its robustness against noisy data:

$$\min_{\omega} ||X\omega - y||_2^2 + \alpha||\omega||_2^2. \tag{3}$$

Furthermore, the use of the fused surface instead of the reference surface from each single reconstruction induces further consistency of the reflectance model across different frames.

We noticed that the original SFMS formulation tends to underestimate surface depth for points farther away from the camera compared to the average depth of the scene. We suspect this is because the single-reflection assumption inherent in [1] does not hold in endoscopic video. Points farther away from the camera are additionally illuminated by light bouncing off nearer points. Figure 2 shows that the original reflectance model (on the right) expects the far surface to be very dark, while it is much brighter in the actual image (on the left). We solve this problem by reducing the falloff speed of illumination in the SFS model and thus roughly approximate the multiple light bouncing effect where the overall environment is brighter. Equation 4 is the new reflectance model, where m controls the rate of light attenuation:

$$I_r = I_i \frac{A}{r^m} \cos(\theta)\, BRDF(\theta). \quad (4)$$

I_r is the observed radiance, I_i is the light source intensity, A is related to the projected area of the light source, and θ is the angle between incident light and surface normal. Table 1 shows the total squared error in intensity, averaged over 12 images, using a variable falloff term versus using a fixed falloff of $m = 2$. It is apparent that intensity over the entire image is much better modeled when a variable falloff is used.

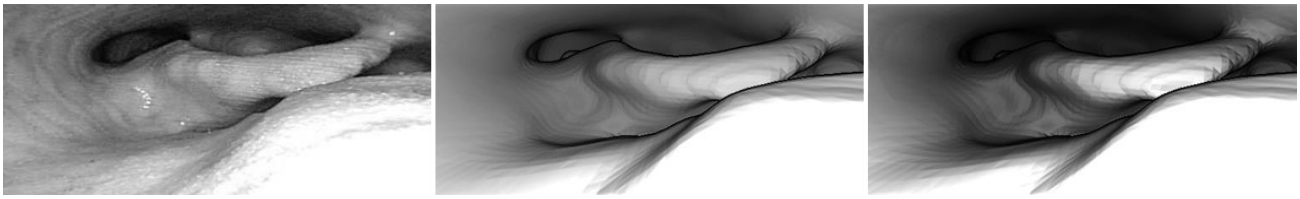


Figure 3. Estimated image from the refined and original reflectance models. From left to right: original image, estimation according to the refined reflectance model, and estimation according to the original reflectance model.

	Variable Intensity Falloff (proposed)	Fixed Intensity Falloff
Mean squared error in intensity over 12 images	3261.293	7983.519

Table 1.The mean squared error in intensity between the original input intensity image and a rendered version of that image using a reflectance model fit to that image with the underlying ground-truth surface. Error is averaged over 12 images of the phantom model.

3. RESULTS

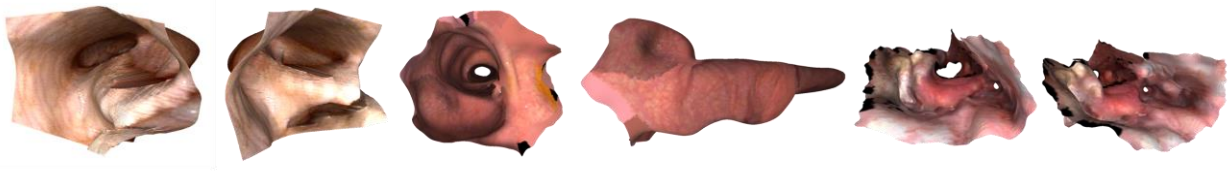


Figure 4. Example results from our improved SFMS method. Left: Phantom. Middle: Colonoscopy video. Right: Throat.

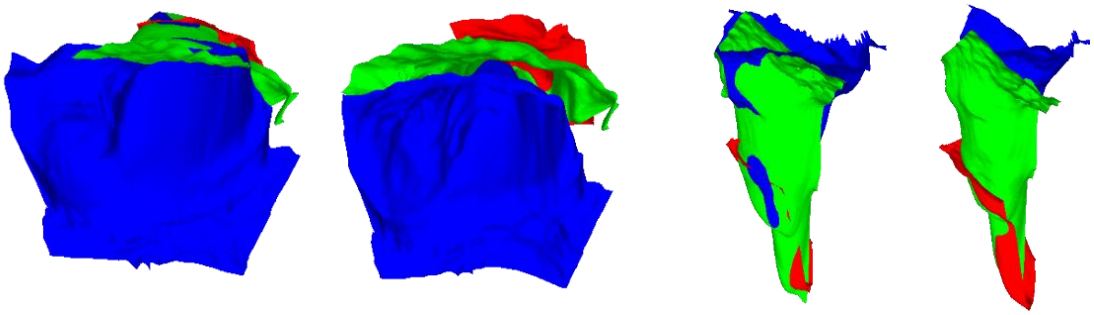
To evaluate SFMS, Price et al. used endoscopy of a 3D-printed phantom model. A CT image of that model provided a ground-truth 3D mesh of the throat (on the left of Figure 3). We use the same data and evaluation scheme to show the superiority of our fusion-guided SFMS. The closest distance of SFMS estimation to phantom surface is used to measure the reconstruction accuracy. We uniformly picked 50 frames from a sequence of 100 frames as testing data. Table 2 shows the percentage of average distance of each pixel to the ground-truth surface that falls within 0.5, 1.0, 1.5, 2.0, and 2.5 mm. These show improvements due to our modifications.

Methods	**Mean (Std. Dev.) Proportion of Pixels within D mm of Ground Truth**

D	0.5mm	1.0mm	1.5mm	2.0mm	2.5mm
SFMS	0.148 (0.066)	0.273 (0.093)	0.386 (0.100)	0.485 (0.108)	0.573 (0.115)
SFMS with improved refl. model	**0.169 (0.044)**	0.314 (0.071)	0.427 (0.100)	0.519 (0.116)	0.593 (0.123)
SFMS with fusion and improved refl. model	0.158 (0.024)	**0.319 (0.054)**	**0.453 (0.090)**	**0.560 (0.112)**	**0.637 (0.110)**

Table 2. Comparison result between original and improved SFMS methods using ground-truth endoscopic data.

Since the phantom is rigid, the SFM algorithm already produces a fairly dense point cloud, which leads to rather consistent surface reconstructions between adjacent frames. However, in real endoscopic video, SFM produces only a sparse and sometimes inaccurate point cloud due to tissue deformation. Therefore, without temporal regularities, the original SFMS generates reconstruction results that have larger deformations than pure tissue deformations between adjacent frames, which we called inconsistent reconstructions. We have used real patient data to visually compare the reconstruction consistency between the original and the fusion-guided SFMS methods. Besides the pharyngeal dataset, we also applied the improved SFMS on colonoscopic video as a new application. Figure 4 shows the comparison result on a colonoscopic video sequence. Those three surfaces (in red, green, and blue) are reconstructed from three adjacent frames. As we can see, fusion-guided SFMS (a and c) produces a more consistent reconstruction (surfaces are closer to each other) than the original SFMS (b and d).



(a) (b) (c) (d)

Figure 5. Demonstrating increased reconstruction consistency of improved SFMS method. Three single-frame reconstructions of colonoscopic video, shown respectively in blue, green, and red are superimposed. (a) top view of improved SFMS results. (b) top view of original SFMS results. (c) side view of improved SFMS results. (d) side view of original SFMS results.

4. DISCUSSION & FUTURE WORK

This paper communicated an effective method for reconstructing a 3D textured surface from endoscopic video named SFMS. We have presented several improvements to it. By using a fused reference surface, we incorporated a temporal constraint into the frame-by-frame SFMS that leads to more accurate and consistent reconstructions. This allows the method to be applied to longer sequences, such as colonoscopic videos. We also refined its reflectance model by outlier pixel removal, preference for realistic BRDFs, and an approximation for multiple bouncing light. We have demonstrated via both phantom and real endoscopic videos that our fusion-guided SFMS produces more accurate and consistent results.

The target object of our system is deformable; therefore, using a rigid phantom for evaluation could not show its full capacity. Therefore, we are developing a new evaluation scheme that use a deformable phantom as ground-truth. We will use one of the endoscopograms that has realistic texture and 3D structure as a base model. Such an endoscopogram is shown in figure 4. Then we will learn elasticity parameters of this model from a sequence of frame-by-frame 3D reconstructed partial surfaces. Afterwards, the learned elasticity parameters will be applied to the base model to create realistic deformations. Finally, a synthetic camera and light source can be simulated using computer graphics techniques to produce a synthetic endoscopic video. We believe this deformable ground-truth data can produce a more comprehensive evaluation than our current 3D printed phantom model. Furthermore, since the whole rendering is based

on computer graphics techniques, different BRDF can be applied to evaluate the generality of the surface reflectance modeling.

We are also applying convolution neural network (CNN) techniques to further improve our current pipeline. We have applied CNN to automatically select informative frames (no motion blur, no saturated illumination, clear view of target object, etc.) from a raw endoscopic video. We are currently working on using CNN to directly infer depth from endoscopic images, which will bypass the complex surface reflectance modeling problem.

REFERENCES

[1] Price, T., et al., "Shape from Motion and Shading in Uncontrolled Environments," Under review. Preliminary draft available at http://cs.unc.edu/~jtprice/papers/sfms_preliminary_2016.pdf

[2] Zhao, Q., et al., "The Endoscopogram: a 3D model Reconstructed from Endoscopic Video Frames," To Appear in MICCAI (2016). Paper available at http://midag.cs.unc.edu/pubs/papers/qingyu_miccai_16.pdf

[3] Kaufman, A., Wang, J., "3D surface reconstruction from endoscopic videos," Visualization in Medicine and Life Sciences. Springer Berlin Heidelberg, 61-74 (2008).

[4] Wu, C., Narasimhan, S.G., Jaramaz, B., "A multi-image shape-from-shading framework for near-lighting perspective endoscopes," International Journal of Computer Vision 86(2) (2010) 211–228.

[5] Nadeem, Saad, and Arie Kaufman. "Depth Reconstruction and Computer-Aided Polyp Detection in Optical Colonoscopy Video Frames." arXiv preprint arXiv:1609.01329 (2016).

[6] Hong, D., Tavanapong, W., Wong, J., Oh, J., and de Groen, P. C., "3D reconstruction of virtual colon structures from colonoscopy images," Computerized Medical Imaging and Graphics 38(1), 22–33 (2014).

[7] Schönberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." CVPR, 2016.

[8] Schönberger, Johannes L., et al. "Pixelwise View Selection for Unstructured Multi-View Stereo."European Conference on Computer Vision. Springer International Publishing, 2016.

[9] Ahmed, Abdelrehim H., and Aly A. Farag. "A new formulation for shape from shading for non-Lambertian surfaces." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.

[10] Horn, Berthold KP. "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view." (1970).

[11] Zhang, Ruo, et al. "Shape-from-shading: a survey." IEEE transactions on pattern analysis and machine intelligence 21.8 (1999): 690-706.

[12] Durou, Jean-Denis, Maurizio Falcone, and Manuela Sagona. "Numerical methods for shape-from-shading: A new survey with benchmarks." Computer Vision and Image Understanding109.1 (2008): 22-43.

[13] Hartley, Richard, and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

[14] Pollefeys, Marc, et al. "Visual modeling with a hand-held camera." International Journal of Computer Vision 59.3 (2004): 207-232.

[15] Malti, Abed, and Adrien Bartoli. "Combining Conformal Deformation and Cook–Torrance Shading for 3-D Reconstruction in Laparoscopy." IEEE Transactions on Biomedical Engineering61.6 (2014): 1684-1692.

[16] Malti, Abed, Adrien Bartoli, and Toby Collins. "Template-based conformal shape-from-motion-and-shading for laparoscopy." International Conference on Information Processing in Computer-Assisted Interventions. Springer Berlin Heidelberg, 2012.