Endoscopic-CT: Learning-Based Photometric Reconstruction for Endoscopic Sinus Surgery

A. Reiter[a], S. Leonard[a], A. Sinha[a], M. Ishii[b], R. H. Taylor[a], and G. D. Hager[a]

[a]Johns Hopkins University, Dept. of Computer Science, Baltimore, MD, USA
[b]Johns Hopkins Medical Institutions, Dept. of Otolaryngology - Head and Neck Surgery, Baltimore, MD, USA

ABSTRACT

In this work we present a method for dense reconstruction of anatomical structures using white light endoscopic imagery based on a learning process that estimates a mapping between light reflectance and surface geometry. Our method is unique in that few unrealistic assumptions are considered (i.e., we do not assume a Lambertian reflectance model nor do we assume a point light source) and we learn a model on a per-patient basis, thus increasing the accuracy and extensibility to different endoscopic sequences. The proposed method assumes accurate video-CT registration through a combination of Structure-from-Motion (SfM) and Trimmed-ICP, and then uses the registered 3D structure and motion to generate training data with which to learn a multivariate regression of observed pixel values to known 3D surface geometry. We demonstrate with a non-linear regression technique using a neural network towards estimating depth images and surface normal maps, resulting in high-resolution spatial 3D reconstructions to an average error of 0.53mm (on the low side, when anatomy matches the CT precisely) to 1.12mm (on the high side, when the presence of liquids causes scene geometry that is not present in the CT for evaluation). Our results are exhibited on patient data and validated with associated CT scans. In total, we processed 206 total endoscopic images from patient data, where each image yields approximately 1 million reconstructed 3D points per image.

Keywords: 3D reconstruction, structure from motion, shape from shading, video-CT registration

1. INTRODUCTION

Sinus surgery is typically performed under endoscopic guidance, through a procedure termed Functional Endoscopic Sinus Surgery (FESS), a large percentage of which employ surgical navigation systems to visualize critical structures that must not be disturbed during surgery. Because the sinuses contain structures that are smaller than a millimeter in size and delineate very critical anatomy such as the optic nerve and the carotid artery, navigation accuracy is critical. Even though in ideal conditions navigation accuracy typically reaches only 2mm,[1,2] we have previously developed video-CT registration capabilities that are demonstrated to CT accuracy,[3,4] averaging from 0.83-1.2mm, depending on the viewing conditions.

Beyond navigation, 3D metrology and reconstruction is an important capability for in-situ FESS because as surgeries progress, often anatomy is purposely disturbed, making correspondence of endoscopic imagery to pre-operative CT scans difficult to near-impossible. Though it is possible to perform intra-operative CT, the additional exposure to radiation is an unnecessary side effect. In ideal conditions, the endoscopic imagery can be used for high quality reconstruction to serve as an intra-operative "Endoscopic CT", without the added risks to radiation exposure. This work demonstrates a method that combines video-CT registration with a light reflectance model to estimate a depth image for every image in the endoscopic sequence (i.e., the distance from the camera to every point in the scene, for every pixel in the image), ultimately resulting in a photo-realistic 3D reconstruction of the surgical site.

Further author information: (Send correspondence to A. Reiter)
A Reiter: E-mail: areiter@cs.jhu.edu

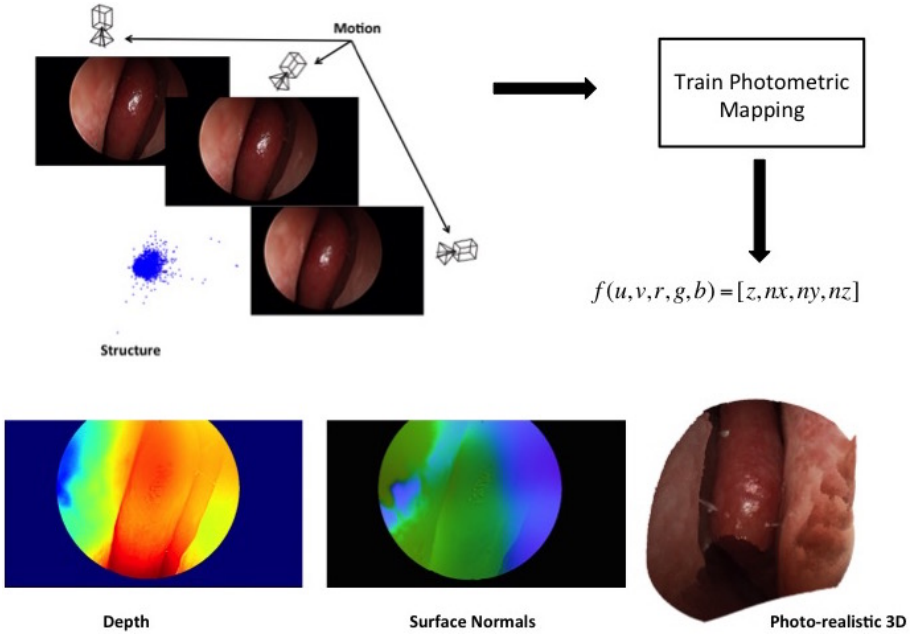

Figure 1. Flowchart for our photometric reconstruction framework.

2. METHODS

Our methodology relies on using training data gathered from our SfM process to be able to associate 3D scene points with pixel colors and locations in an image for the purposes of learning a mapping through statistical regression (see Fig. 1 for a diagram overview of our procedure). This approach is unique in that the regression inherently accounts for typically difficult-to-model physical processes, such as sub-surface scattering and absorption, inter-reflections, light intensity profile, position, and direction, and so on. The Bidirectional Reflectance Distribution Function (BRDF) is a function that estimates these parameters, however in practice this is very difficult to model. We begin by a discussion of gathering training data for the purposes of training our regression.

2.1 Gathering Training Data

Figure 1 demonstrates the proposed framework for photometric reconstruction from training data gathered by our video-CT SfM process.[3] In short, SfM is a computer vision technique that estimates the 3-dimensional "structure" of a scene using a series of images in which the camera is moving throughout a static scene. As a byproduct, SfM also yields the 3D camera positions and orientations (the "motion") from which each image was taken, allowing us to correspond the 3D scene points to pixels in the images through standard projective geometry.[5] Our framework begins by collecting **N** 3-D scene points $\mathbf{s} = [s_1, s_2, \ldots, s_N]$, where each $s_i = (x_i, y_i, z_i)$, the result of the SfM process. Additionally, for each of the **M** images used to produce our SfM result, we collect the associated **M** camera poses $\mathbf{p} = [p_1, p_2, \ldots, p_M]$, where each $p_j = [R_j, t_j]$, the orthonormal 3x3 rotation matrix R representing the orientation of the camera as well as the position of the camera t.

Given these 3D points and camera poses, for each point and camera we wish to collect the associated pixel position **x** and color **c**, by applying perspective geometry. First, each 3D point s_i is transformed from the SfM coordinate system to the given camera p_j coordinate frame:

$$\begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{bmatrix} = R_j \begin{bmatrix} x_{SfM} \\ y_{SfM} \\ z_{SfM} \end{bmatrix} + t_j \tag{1}$$

We apply projective normalization to retrieve the pixel position $\mathbf{x_{ij}} = [u_{ij}, v_{ij}]$:

$$[u,v] = [f_x \frac{x_{cam}}{z_{cam}} + c_x, f_y \frac{y_{cam}}{z_{cam}} + c_y] \tag{2}$$

where (f_x, f_y) and (c_x, c_y) are the camera's intrinsics focal length and principal point, respectively (recovered through a one-time camera calibration technique prior to the procedure,[6] along with distortion coefficients to rectify the images). Finally, given each $\mathbf{x_{ij}}$, we then look up color $\mathbf{c_{ij}} = [r_{ij}, b_{ij}, c_{ij}]$ for point s_i in image j. In the end, the SfM process yields $N * M$ observations with which to train our photometric regression. Though SfM is able to produce high accuracy reconstructions,[3] they are typically sparse at best because they rely heavily on texture in the scene, which is rare and difficult in endoscopic video. Therefore, even with high-resolution imagery, it is typical to acquire only a small ratio of reconstructed points relative to the total number of pixels in the image.

2.2 Reflectance Regression

The typical imaging setup describes the scene as a camera with a point light source that images a 3D surface. In this case, the image is formed by an understanding of how the light is emitted from the source, reflects off the viewed surface, and absorbs into the image sensor of the camera. The BRDF is a 4-dimensional function that relates the incoming light direction, outgoing viewing direction, surface normal direction, and reflected radiance to describe the imaged light of the surface. In practice, there are many variables that affect this, as mentioned above. Modeling an accurate BRDF is a difficult task, where much work has been done in the past, however still lacking a fully general solution. However, in the case of endoscopic video, there is a unique situation that occurs in which the light source is statically fixed to the camera, and so as the camera moves throughout the scene, so does the light in order to form the images. We use this fact to simplify our reflectance model so that the light source and viewing directions are the same. However, different from prior work,[7] we make no assumptions about the actual position of the light source (nor that it is a point source) as well as no assumptions on how light reflects off of surfaces.

Typically, a Lambertian model is employed, whereby the apparent brightness of a surface is the same to any observer regardless of the viewing angle. In practice, especially with endoscopic imaging, physical issues such as tissue absorption and scattering as well as the presence of liquids invalidate this assumption. We alleviate these challenging issues by formulating the reflectance model in the form of a regression learned directly from the data. Because our SfM process yields 3D points related to 2D pixels across several images, we seek to estimate the following function:

$$f(u,v,r,g,b) = [z, n_x, n_y, n_z] \tag{3}$$

where (u,v) in a pixel in an image, (r,g,b) is the associated red-green-blue color at this pixel, z is the 3D depth of this viewed pixel (as obtained by the SfM sparse reconstruction), and (n_x, n_y, n_z) is the surface normal (obtained through recovering the closest triangle of a mesh produced using video-CT registration, the description of which is out of the scope of this paper and has been demonstrated previously[3]). The trick here is that because the light moves with the camera, each time a scene point is viewed, the appearance of that same point varies as the camera/light moves. By observing and modeling how appearances change across many scene points and mapping these to geometry with respect to the camera (i.e., all depths and surface normals are expressed relative to the camera frame), we understand how light reflectance relates to scene geometry without explicitly modeling the difficult physical characteristics mentioned earlier. Furthermore, by incorporating the pixel position (u,v) as a part of the parameterization of f, the light profile and position is inherently addressed without understanding the explicit location, intensity profile, or direction. This is because the amount of light emanating from the camera as viewed by a particular pixel is constant (because it moves statically with the camera), and what changes then is how the scene geometry reflects/absorbs this light. We also note that because we only utilize light reflectance, there is no reliance on texture.

2.3 Non-Linear Regression Model

In this work we experimented with a non-linear regression technique using a neural network to estimate f. Here, we setup a feed-forward neural network with 5 input nodes (u,v,r,g,b) through a single hidden layer with 8 nodes (empirically determined so as to avoid data over-fitting), and 4 output nodes (z, nx, ny, nz). Each of the nodes in the hidden and output layers employed hyperbolic tangent functions.

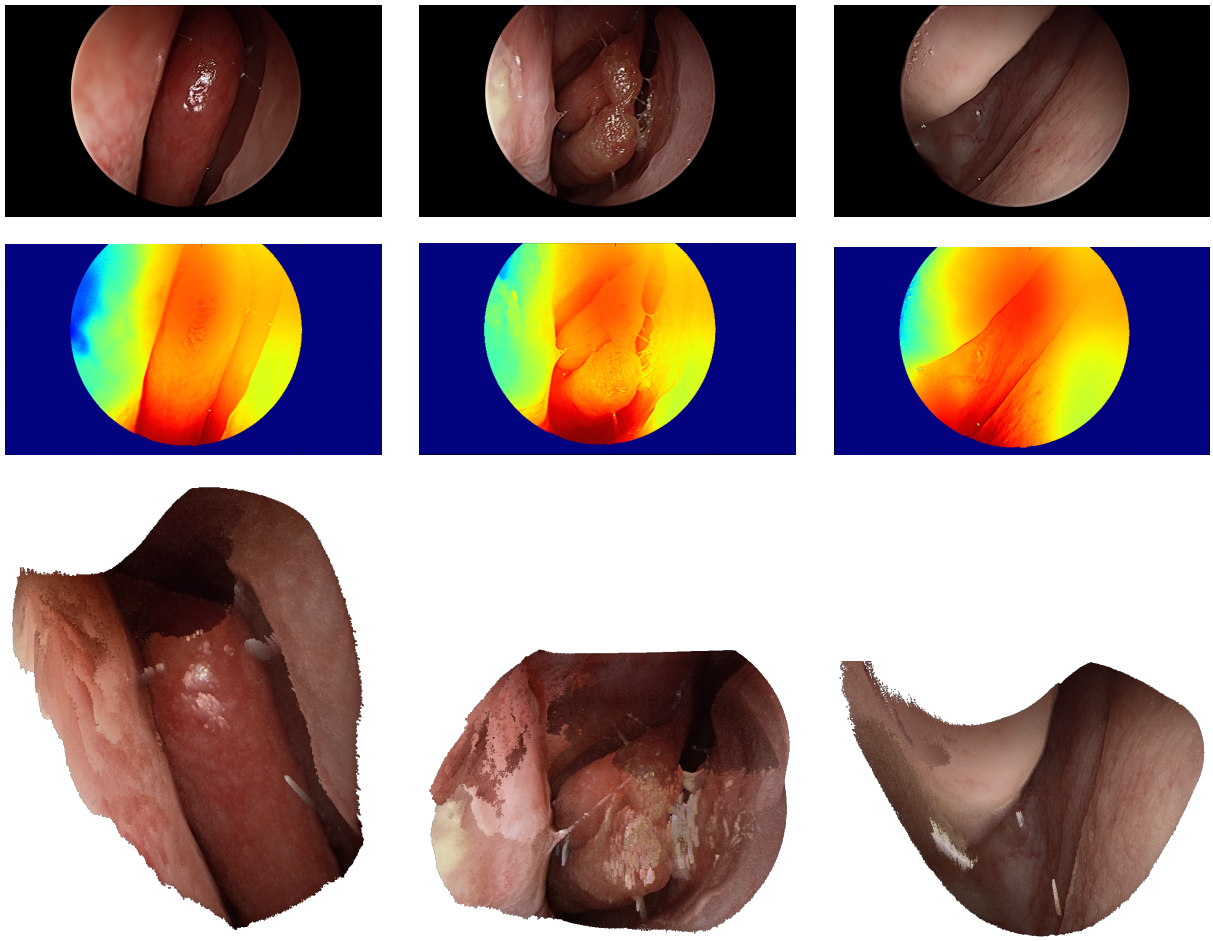

Figure 2. Examples of dense photometric reconstruction using our neural network technique. Each column is a different sequence. The first row is the color endoscopic image, chosen from a video sequence; second row is the corresponding depth image (red maps to high values, and blue to low); third row is the surface normal color mapped image; last row is the photo-realistic 3D reconstructed point cloud, with RGB colors assigned to each 3D point.

3. RESULTS

In this section we describe results for our non-linear regression approach. Here we seek to demonstrate the capabilities of the proposed approach by training the photometric regression on a small subset of images, and then applying the trained model throughout images from the rest of the sequences at different, novel anatomical sites (i.e., the models were not updated over time, but just trained once).

3.1 Training

We used 103,665 SfM points collected from 36 images to train the regression, where each image has a pixel resolution of 1920x1080. For each of the 36 images, we have camera matrices relating the SfM world to the cameras as well as video-CT registration results so we can evaluate the accuracy. For our neural network regression, we used 77,748 (75%) of these SfM points to train the network using back propagation with a learning rate of 0.01, and obtained an accuracy of 0.36mm in depth and 29.5 degrees in surface normal angle accuracy on the remaining 25,917 SfM validation points.

3.2 Evaluation

In order to evaluate the accuracy of our reconstructions, we estimate a depth image for each endoscopic image separately using our trained model, now for images not used in the training process (and at different areas of the sinuses not visible

during training). Each depth point is transformed to a 3D scene point in the camera coordinate system using the reverse of the projective equations described above. Then, using our video-CT registration, we express each 3D point in the CT coordinate system. After producing a mesh of triangles from our patient CT scans, we locate the nearest 3D triangle to each 3D point, and compute the point-to-plane distance to this triangle to express the best average measure of reconstruction accuracy over all test points.

3.3 Testing

Figure 2 shows several examples of our dense photometric reconstruction technique. The top row displays the color image from 3 different endoscopic video sequences. The second and third rows are the depth and surface normal mapped images, respectively, and the last row is a 3D plot of the colorized photo-realistic 3D reconstruction as a point cloud. The surface normal color mapped images (3^{rd} row) are constructed by mapping each of the 3 elements of the unit surface normal vector to a value in the range [0-255] (n_x for red, n_y for green, and n_z for blue).

The accuracy of our reconstructions, as obtained by comparing to the CT scans, differs based on how well the anatomy matches the CT. For example, the first column shows a fairly clean anatomy, and so the mean accuracy we obtained was 0.53mm. However, the second column shows a situation with significant amounts of mucus, which is not present in the CT, and in this case our mean accuracy was 1.06mm. Nevertheless, looking carefully at the last row of column 2, one can clearly see that the mucus is reconstructed and showcased fairly accurately, illustrating the detail of the reconstruction. Conversely, in both the first and third columns, one may notice spikes within the reconstruction, and these are due to poor behavior from specularities in the image, a common imaging artifact with endoscopic imaging. Addressing this challenge is a topic of future research.

Sequence	Number of Images	Number of total reconstructed points	Mean Accuracy (mm)	Standard Deviation
01	36	36085676	0.53	0.38
04	36	36058547	1.06	0.68
05	33	32600950	1.12	0.79
06	34	33388650	0.74	0.50
09	33	33109030	1.09	0.65
11	34	34089486	1.07	0.69

Table 1. Table of mean accuracy and associated standard deviation in 3D position across several sequences

Table 1 shows our full results, where we chose several sub-sequences to reconstruct for an exhaustive evaluation. Using a total of 206 images, we reconstruct every point in every image (except outside the circular area defined by the endoscope as well as extremely bright regions that serve as specularities, though some still filter through incorrectly). The third column shows the total number of points reconstructed for each sequence, given the endoscopic region and specularity constraints, across all of the images shown in column 2. The third and fourth columns show the mean accuracy and associated standard deviation for 3D point reconstruction as related to the CT.

4. CONCLUSIONS

In this work we presented a novel method for dense photometric reconstruction of endoscopic imagery by learning a mapping of light reflectance to scene geometry through a non-linear regression with a neural network. We showed that by learning this function on a per-patient basis, we are able to achieve very high accuracy on novel images throughout the scene by assessing the point reconstructions against the CT scan. We made no assumptions about the position or shape of the light profile or how the surface reflects and absorbs light. Future work includes further enhancing, and perhaps simplifying, the regression model as well as addressing current challenges such as light specularities.

REFERENCES

[1] Fried, M., Kleefield, J., Gopal, H., Reardon, E., Ho, B., and Kuhn, F., "Image-guided endoscopic surgery: Results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system," *Laryngoscope* **107**(5), 594–601 (1997).

[2] Metson, R., Gliklich, R., and Cosenza, M., "A comparison of image guidance systems for sinus surgery," *Laryngoscope* **108**(8), 1164–1170 (1998).

[3] Mirota, D., Uneri, A., Schafer, S., Nithiananthan, S., Reh, D., Ishii, M., Gallia, G., Taylor, R., Hager, G., and Siewerdsen, J., "Evaluation of a system for high-accuracy 3d image-based registration of endoscopic video to c-arm cone-beam ct for image-guided skull base surgery," *IEEE Transactions on Medical Imaging* **32**(7), 1215–1226 (2013).

[4] Otake, Y., Leonard, S., Reiter, A., Rajan, P., Siewerdsen, J., Ishi, M., Taylor, R., and Hager, G., "Rendering-based video-ct registration with physical constraints for image-guided endoscopic sinus surgery," in [*SPIE Medical Imaging*], (2015).

[5] Hartley, R. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, New York, NY USA (2003 (second edition)).

[6] Zhang, Z., "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11), 1330–1334 (2000).

[7] Okatani, T. and Deguchi, K., "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).