REGISTRATION OF NASOPHARYNGOSCOPIC VIDEO WITH TREATMENT PLANNING CT SCANS

Julian Rosenman, PhD MD[*†], Qingyu Zhao, MS[†], True Price, BS[†], Marc Niethammer, PhD[†], Ron Alterovitz, PhD [†], Jan-Michael Frahm, PhD[†], Bhishamjit Chera, MD[*], Stephen Pizer, PhD [*†]

*Department of Radiation Oncology
†Department of Computer Science
The University of North Carolina at Chapel Hill

Corresponding Author

Julian Rosenman, PhD MD
Department of Radiation Oncology
101 Manning Drive, CB 7512
Chapel Hill NC 27514

FAX

(919) 966-7681

EMAIL

rosenmju@med.unc.edu

Shortened running title

Registering endoscopic video with CT

Support

Supported in part by Grant 1-R01-CA158925-01A1

Conflict of Interest Notification

No any actual or potential conflicts of interest exist.

SUMMARY

Videos taken at nasopharyngoscopy may often contain information on mucosal tumor spread not seen on PET/CT because they record mucosal color and texture. We show how such a video can be reconstructed into a high resolution 3D textured surface which can then be registered to its corresponding radiotherapy treatment planning CT scan. This registration then allows for the direct transfer of information from the video to the radiation treatment planning system.

ABSTRACT

Purpose: We present a method that allows one to generate data from nasopharyngoscopy videos into a form suitable for use in radiation treatment planning, as well as for review of the data itself as a 3D object.

Methods and Materials: The accurate transfer of data from endoscopic video frames to the planning CT requires an explicit endoscopic video/CT registration. Our approach to accomplishing this goal is to first reconstruct an accurate 3D model of the anatomy from the 2D video data as seen on multiple video frames. That reconstruction, which we call an *endoscopogram*, can then be deformably registered to the CT scan.

To generate the endoscopogram we first compute a sparse 3D *point cloud* by tracking corresponding *feature points* from one video frame to another using standard structure-from-motion (SfM) techniques adapted to the pharyngeal region. We then use a second technology, shape-from-shading (SFS), to produce high-quality images from each video frame. Uncorrected, these SFS images contain dense information and in many places accurate *local* curvature, but they lack global spatial fidelity. Using our newly developed techniques, we can correct the SFS images using global information from the sparse SFS-generated point cloud as a constraint. We then register several of these corrected SFS images to each other to generate a 3D image that displays the entire surface area seen in the whole video.

Results: We have successfully computed a spatially accurate endoscopogram (shown below) and have registered it with the planning CT scan with an average accuracy of about 3-5 mms.

Conclusions: The technologies developed by our working group will allow accurate transfer of data from endoscopy images directly onto the planning CT scan as well as review of the video as a 3D object.

INTRODUCTION

Modern radiation therapy treatment planning relies on imaging to determine tumor location and spread. Although for several reasons CT scans are preferred as the base image for radiation treatment planning, information from MRI and PET scans can be added to the planning CT via registration. In fact, a rigid registration is usually all that is necessary because CT and MRI are acquired with the patient in a similar position and in the case of PET/CT these two scans are done in rapid sequence during the same imaging session.

In the past few years a new and important kind of medical imaging has become readily available although we do not usually think of it that way. These are videos taken during the time of endoscopy. For example, otolaryngologists and radiation oncologists regularly perform nasopharyngoscopy because they feel that direct visualization of the tumor provides information on tumor location (especially mucosal spread) not available on CT or even PET/CT. The radiation oncologist, in particular, uses that information to improve the accuracy of gross tumor volume (GTV) delineation on CT. However, this process of using the video data has to be done indirectly, through the physician's memory or notes, as there is currently no direct way to transfer the nasopharyngoscopic data to the planning CT.

This paper describes a new process whereby endoscopic video frames can be reconstructed into a 3D textured surface which we call an endoscopogram. The endoscopogram then can registered with the planning CT allowing selected data on it, such as gross tumor extent, to be added directly to the planning CT. In addition, the endoscopogram can be used as an efficient stand-alone review object, which could be easily compared to previous studies. The possible clinical benefits of this technology will be discussed as well as the technical challenges that still need to be fully overcome. To our knowledge this approach to improve radiation treatment planning has not been previous attempted.

METHODS AND MATERIALS

There are two major problems that must be solved before one can register an endoscopic video with the planning CT. The first is the reconstruction of an accurate 3D model from the 2D video data (the endoscopogram) and the second is the registration of the reconstructed endoscopogram with the CT.

Reconstruction of the endoscopogram: The problem of reconstructing a spatially accurate endoscopogram from the endoscopic video proved to require three separate processes, two of them new. The first process is that of determining the three-dimensional spatial position of as many "feature points" within the endoscopic video frames as possible, along with the camera position and orientation for each video frame. The method presupposes that, to first approximation, the scene remains still and the camera moves in a regular fashion, although the camera position and orientation are not known. Methods for computing this *point cloud* (set

of points for which the spatial position has been determined) under these conditions have been developed over many years [1-7], but they have typically not been used to reconstruct human anatomy. More commonly, such techniques are used to reconstruct "urban scenes" composed of buildings, roads and other objects that contain many straight lines and right angles as well as a fixed light source. Although the output of *structure from motion* (SfM) algorithms is typically a only a relatively *sparse* point cloud, under the assumption that the buildings and other structures are mostly comprised of flat or near-planar faces an accurate, high resolution 3D model can be reconstructed. That is not the case for endoscopy.

Human anatomy is noticeably lacking in straight lines and planar surfaces, and as a result, a high-resolution reconstruction of an endoscopogram from a sparse SfM-generated point cloud proved not to be possible. Moreover, since the anatomy is continuously deforming, we need to limit the structure from motion reconstruction to short frame sequences where the rigidity and fixed lighting assumptions hold approximately.

Another historically successful approach to reconstructing 3D models from 2D data, but one that does not require computation of a point cloud, is known as "shape-from-shading" (SFS). SFS is a technique that attempts to recover the 3D shape of a 2D image from the gradual variation of shading within the image. This is the reverse of what artists do to convey depth by an appropriate shading method. This idea, that of obtaining shape from shading, stems from work begun in 1952 when it was used to calculate the slopes and heights of mountains and craters on the moon from their shadows [8]. In the 1980s a series of papers formalized the method for images with a known lighting model, i.e., known light sources and surfaces of known reflectivity. Presently there are at least six major approaches to doing shape-from-shading from a single image [9]. In our experience, rendering the pharynx with SFS typically results in high quality 3D images with *locally* accurate curvature, but with *globally* inaccurate depth scaling, and significant connectivity problems, particularly in areas of occlusion, where part of the anatomy cannot be viewed because something is in the way.

Given that we now had a method to produce a globally accurate point cloud, but one that lacked information on local curvature, and an SFS imaging technique that produced that local curvature correctly by another means, we saw that the two approaches should be somehow combined. Unfortunately, the most obvious method to do this, that of registering the SFS images directly to the sparse SfM-generated point cloud, proved to be unsatisfactory as there appeared to be no good way to interpolate the surface between the SfM fiducials.

Fortunately, we were able to develop an iterative method to correct the global errors found in SFS-generated images. The approach begins with the assumption that there is *some* lighting model, varying across the surface that when used with current SFS algorithms would produce a surface with both locally and globally correct curvatures and depths. In this approach the global sparse depth map,

determined by the SfM point cloud, is used to compute an improved lighting model. In an iterative fashion, the new lighting model is then used to compute an improved depth map. Thus, though we refer to an SFS image as coming from a single video frame it is understood that the surface is corrected by an SfM point cloud derived from multiple temporally-local video frames. Technical details of this new method will be appear elsewhere [10]. The construction of spatially accurate shape-from-shading models is the second component of our approach.

The third process is to combine several SFS textured surfaces into a single textured surface so as to display the entire 3D region seen at endoscopy. We divide this step into 1) the formation of the overall geometry of the pharyngeal surface and 2) its texturing. The issue of proper texturing in the general case is still under study and will be described in a future publication.

Step 1 is necessary as no one video frame contains enough data to display the entire oropharyngeal surface, as occluding surfaces inevitably will be present. For example, one cannot see both the lingual and laryngeal sides of the epiglottis in a single 2D image. But because of inevitable patient motion between video frames that are temporally separated, a rigid registration between SFS surfaces is not satisfactory. It proved to be the case that a good way to deformably register these SFS images was a modification of the same method we used to register the SFS images with the CT itself, a surface-deformable registration. This approach, also new. will be described next.

Deformable registration of the planning CT with the endoscopogram:

We first evaluated the standard "head-in-hat" approach matching SfM points to the CT surface but found that method involving only locations wanting. In contrast, the deformable registration method that we have developed for registration of the endoscopogram with CT considers mechanical aspects of the deformation and locational and curvature aspects of the surfaces. The method is called *thin shell demons*, [11]. Motivated by the demons framework developed by Thirion [12], we regard one surface as an elastic thin shell with additional structural energy that can be attracted via virtual forces produced by the other surface. The result is a deformation process that tries to align similar geometric structures in a physically realistic way. The method satisfies our need that the attraction surface, from shape-from-shading, can have holes in the surfaces (due to occlusion) and still be effective. Experimentation has also shown that this thin shell demon approach gives satisfactory results when we use it to register one SFS surface with another as well.

Fusion of SFS surfaces: Our method for fusion involves registering selected SFS-generated surfaces together. At present we accomplish this task for each of a small number of pharyngeal regions, for example the epiglottal and laryngeal, and then we fuse the resulting surfaces for each region.

Putting it all together: With these software tools we are now capable of mapping points on the endoscopic video frames to points on the treatment planning CT scan, the goal of our project. To summarize, 1) We first reconstruct a spatially correct textured surface from selected video frames using SFS. 2) We then register the corrected SFS images with each other (fusion) so as to display the entire pharyngeal surface. We call this fused object an endoscopogram 3) Once we have completed construction of this endoscopogram we deformably register it to the treatment planning CT. 4) Then we are in position to identify any selected points (for example, the tumor) on the endoscopic video frames or on the endoscopogram, and we compute their corresponding position on the treatment planning CT. This step then allows us to complete our goal of bringing endoscopic video-generated data into the treatment planning process.

Testing the registration: Because the thin shell demon approach does not use point correspondences as a means of registration we can test the final registration result by manually selecting corresponding points on selected video frames and the CT scan use them to measure the overall accuracy of the global registration under the assumption that there is low human error in making these correspondences. To reduce this error as much as possible, a team of physicians and students worked together to select the corresponding points, using software that displayed the points, both those selected on video and those selected on CT, in both 2D and 3D formats.

RESULTS

Displays: For our initial studies we collected endoscopic videos and CT planning scans, under IRB guidelines, from six anonymized patients. Our nasopharyngoscope is an Olympus model that records full color images at a resolution of 720 x 480 interlaced. CT scans, from a late model Philips CT, have a resolution of 512 x 512 x 3 mm. No special tracking hardware was used to determine endoscope camera position or orientation.

Figure 1 shows a sparse point cloud calculated using structure-from-motion methods. The calculated likely camera trajectories during the procedure are shown in blue. Originally we had hoped to use this kind of image to register with the CT scan, but local curvature, needed for our deformable registration method cannot be accurately calculated from such an image.

Figures 2a and 2b: These textured 3D surfaces of the pharynx are displayed from a single, uncorrected shape-from-shading display. The quality of the image is high, but the spatial accuracy is low. Note that the anterior commissure of the vocal folds appears to be attached to the epiglottis in the image on the right. This is an example of the error that is made due to the presence of occluding surfaces.

Figures 3a, 3b and 3c: This is an endoscopogram, a fully 3D object, that is seen in three different orientations. On the top display one sees the vallecula and lingual

surface of the epiglottis. In the middle one sees the top of the epiglottis and the vocal folds and pyriform sinuses come into view. On the bottom one sees the anterior commissure and laryngeal surface of the epiglottis. The 3D display is fully interactive and thus one can review the entire pharyngeal surface very quickly.

Accuracy of endoscopogram/CT registration:
The accuracy of the registration for two patients, one with tumor and one without tumor was tested. Table 1, below, shows the results between the team's determination of corresponding points and those determined by the deformable registration. For example, for the anterior commissure the team chose the most likely candidate point on both the CT and the clearest video frame. That particular point was relatively easy to choose on both modalities, but the team found it difficult to agree on the exact locations of many of the other points, particular the tip of the epiglottis because the modalities of endoscopic video and CT are so different and, in that organ at least, there appeared to be a good deal of deformation when the patient was lying down (CT) and sitting up (endoscopy). All measurements below are in millimeters.

Table 1: Differences in millimeters between manual estimation of corresponding points on the endoscopogram and those on CT, determined by our deformable registration methods.

Patient	Tip of epiglottis	Vallecula	Anterior commissure	Right arytenoid	Left arytenoid	Right middle cord	Right pyriform sinus
Normal anatomy	6	3	3	7	n/a	2	n/a
With tumor	3	n/a	3	6	7	n/a	8

Besides the foregoing measurements based on manually chosen corresponding points we also measured the closest point distances between the surfaces of the endoscopogram and CT surface from the epiglottis down to the larynx. The RMS distance was 3 millimeters.

DISCUSSION

Importance of this work: The immediate objective of our work is that of enabling a physician to outline the tumor extent on the endoscopy video and have these contours accurately mapped onto the patient's planning CT scan. We do not envision that the importance of this process is that it will routinely result in small (millimeter) changes in GTV delineation so much as that on occasion it will cause substantial changes in GTV delineation. This situation will likely occur when CT and even PET/CT fail to show extensive mucosal spread of the tumor that is obvious on endoscopy. But this is unknown territory and later in our work we will begin clinical studies to test the above hypothesis.

High quality interactive endoscopograms such as shown in Figures 3 can also serve a valuable function for medical review. Reviewing an entire endoscopic video is time-consuming and thus not routinely done. However viewing an endoscopogram should allow the clinician to find and study areas of concern quickly and easily. As such, the endoscopogram may become the standard review image. In addition, the endoscopogram may be used for surgical planning or even to determine the best approach for a biopsy of a worrisome, persistent lesion. Finally, an endoscopogram can be used to detect (and document) small changes in the mucosal surface that occur over time, as seen on serial examinations.

Perhaps the most important consequence of this work will be the eventual routine incorporation of the data contained in all sorts of endoscopic images into the treatment planning process. Indeed, we have begun exploratory work on adapting these methods to the (perhaps) much harder task of colonic reconstruction from colonoscopy, but others such as esophagoscopy with ultrasound, bronchoscopy, and cystoscopy would also be reasonable to try.

Anticipated improvements: Our first attempt to measure the accuracy of the endoscopic/CT registration was beset by several difficulties. These included 1) the 3 mm thickness of the CT slices, 2) the inherent error of defining a single point in an extended object (such as the right arytenoid), and 3) the lack of accuracy when determining correspondences on steep surfaces, where small error become magnified, and 4) the difficulty in recognizing corresponding points in two such different display modalities. Despite the human error introduced by the above, it is still likely the case that some of the error is due to present inadequacies of both the reconstruction of the endoscopogram and its registration with the CT.

Potential areas of improvement include 1) Using the fact that surface deformations between video frames are slight to regularize the SFS reconstruction. 2) Learning the anisotropic mechanical properties for locations in the pharynx, and using these to assist in computing proper deformation. 3) Performing the registration of multiple SFS surfaces simultaneously as opposed to the current pair-wise method. 4) Tuning various parameters of the registration and reconstruction algorithms.

Other aspects of the process that we are working on are as follows: 1) Increasing the number of images that form the composition of the endoscopogram (perhaps to as many as 50); 2) combining the textures of the SFS surfaces into a single texture on the final endoscopogram and 3) extending this approach to other anatomic sites.

A tool is also under development to draw contours on the endoscopic frames or the endoscopogram and to transfer the resulting 3D contour to the planning CT. We anticipate using the tool to conduct a small clinical trial to test, in a prospective way, the potential value of this technology in radiation treatment planning of head and neck patients.

REFERENCES

[1] Nguyen MH, Wunsche B, Delmas P, **et al**. Modeling of 3D objects using unconstrained and un-calibrated images taken with a handheld camera. *Computer Vision, Imaging and Computer Graphics Theory and Applications Communications in Computer and Information Science* 2013;274:86-101.

[2] Faugeras O, Robert L, Laveau S, **et al**. 3-D reconstruction of urban scenes from image sequences. *Computer vision and image understanding* 1998; 69(3):292–309.

[3] Triggs B, McLauchlan PF, Hartley RI **et al.** Bundle adjustment — a modern synthesis. Vision Algorithms: Theory and Practice. Lecture Notes in *Computer Science* Volume 2002; 1883,(2000):298-372.

[4] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, pages. 1981;674–679.

[5] Tomasi C, Kanade T. Detection and tracking of point features. *Carnegie Mellon University Technical Report CMU-CS* 1991;91-132.

[6] Lowe, DG. Object recognition from local scale-invariant features". *Proceedings of the International Conference on Computer Vision 2*. 1999;1150–1157.

[7] Bay H, Ess A, Tuytelaars T, **et al.** SURF: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 2008;110(3):346–359.

[8] van Diggelen J. A photometric investigation of the slopes and the heights of the ranges of hills in the Maria of the Moon. *Bulletin of the Astronomical Institutes of the Netherlands*. 1952;11:283-289.

[9] Zhang R, Tsai PS, Cryer JE, **et al**. Shape-from-shading: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1999;21(8):690-706.

[10] Price T, Zhao Q, Rosenman J, **et al**. Shape from motion and shading in uncontrolled environments." CVPR 2016 Submission.

[11] Zhao Q, Price T, Niethammer M **et al**. Thin Shell Demons: A physics-motivated approach for surface registration. To appear.

[12] Thirion JP. Image matching as a diffusion process: an analogy with Maxwell's demons. *Med Image Anal*. 1998 Sep;2(3):243-60.

FIGURE LEGENDS

Figure 1: A sparse point cloud calculated using structure-from-motion methods. The calculated likely camera trajectories during the procedure are shown in blue.

Figures 2a and 2b: These textured 3D surfaces of the pharynx are displayed from a single, uncorrected shape-from-shading display. The quality of the image is high, but the spatial accuracy is low.

Figures 3a, 3b and 3c: A fully 3D textured surface we call an endoscopogram seen in three different orientations. Figure 3a is looking straight down at the larynx. In Figure 3b we have tipped the endoscopogram forward so as to view the vallecula. In Figure 3c we have tipped the endoscopogram backwards to view the laryngeal side of the epiglottis. Because the 3D display is fully interactive one can review the entire pharyngeal surface very quickly.

FIGURES

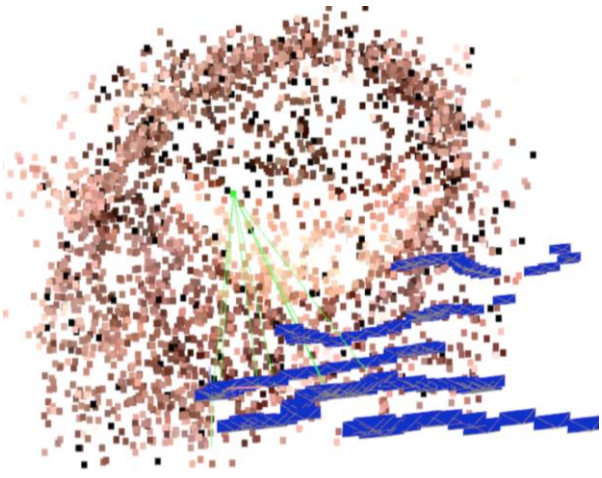
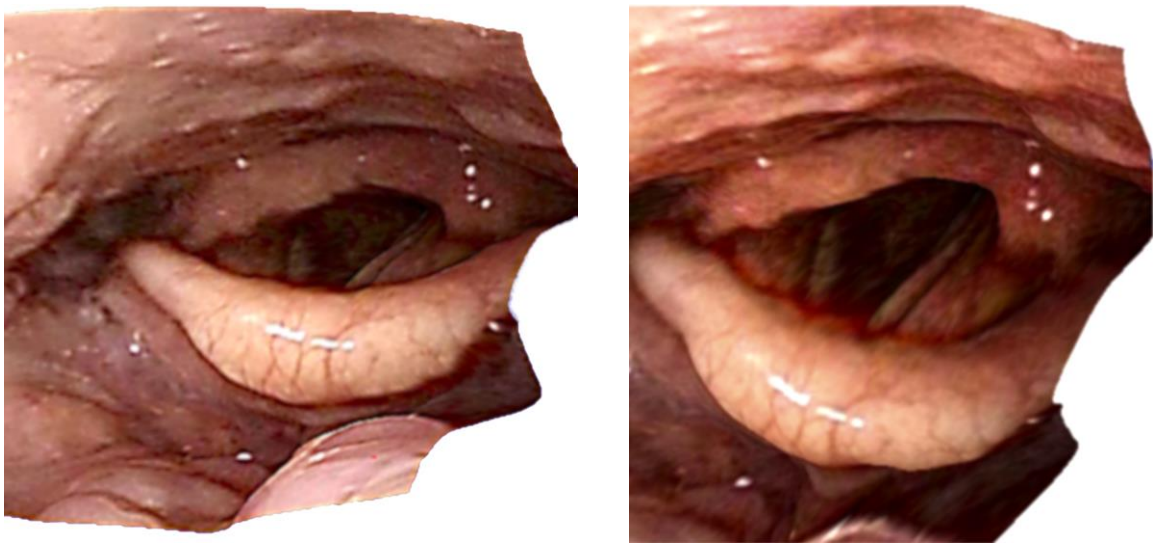


Figure 1



Figures 2a and 2b