Toward Automated Model Building from Video in Computer-Assisted Diagnoses in Colonoscopy

Dan Koppel[†], Chao-I Chen[†], Yuan-Fang Wang[†], Hua Lee[‡]
[†]Department of Computer Science
[‡]Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA

Jia Gu, Allen Poirson, Rolf Wolters
STI Medical Systems
733 Bishop Street, Suite 3100
Honolulu, HI

ABSTRACT

A 3D colon model is an essential component of a computer-aided diagnosis (CAD) system in colonoscopy to assist surgeons in visualization, and surgical planning and training. This research is thus aimed at developing the ability to construct a 3D colon model from endoscopic videos (or images). This paper summarizes our ongoing research in automated model building in colonoscopy. We have developed the mathematical formulations and algorithms for modeling static, localized 3D anatomic structures within a colon that can be rendered from multiple novel view points for close scrutiny and precise dimensioning. This ability is useful for the scenario when a surgeon notices some abnormal tissue growth and wants a close inspection and precise dimensioning. Our modeling system uses only video images and follows a well-established computer-vision paradigm for image-based modeling. We extract prominent features from images and establish their correspondences across multiple images by continuous tracking and discrete matching. We then use these feature correspondences to infer the camera's movement. The camera motion parameters allow us to rectify images into a standard stereo configuration and calculate pixel movements (disparity) in these images. The inferred disparity is then used to recover 3D surface depth. The inferred 3D depth, together with texture information recorded in images, allow us to construct a 3D model with both structure and appearance information that can be rendered from multiple novel view points.

Keywords: Calibration, enhanced reality, image-guided therapy, modeling, visualization

1. INTRODUCTION

The problem of modeling the structure and behavior of a colon for computer-assisted colonoscopy is quite challenging. The complexity depends on a variety of factors; the most important of which are the scale and sophistication of the description, the characteristics of the structure, and the amount of available data. For example, a computer description might be sought only for localized anatomical features, such as polyps, tumors, or abnormal cancerous tissue growth. When a small anatomical feature is analyzed, it may be possible to navigate the scope around the feature with negligible disturbance to its structure. Hence, the structure may be considered static, which allows a 3D computer model to be constructed from video images alone. On the other hand, if the description of a whole colon is sought, one has to address significant tissue deformation resulted from navigating the scope through multiple folds and turns of a colon. A 3D model is much more difficult to obtain given that the scope movement likely will disturb the soft colon structure and results in non-rigid deformation. In this case, an image mosaic approach to construct only a 2D appearance model from video data might be more attainable—with 3D structure data supplied from other sensing modalities such as MRI and CT. In this paper, we report our preliminary results on modeling static, localized anatomic structures in a colon (e.g., a polyp) using video images alone.

Further author information: (Send correspondence to Yuan-Fang Wang)
Yuan-Fang Wang: E-mail: yfwang@cs.ucsb.edu, Telephone: 1 805 893 3866

2. BACKGROUND

Image (or video)-based rendering and modeling have received much attention in the computer graphics and computer vision communities lately. There are basically two approaches: The rendering approach focuses on stitching, or registering, multiple images together into a larger and more complete panorama without explicitly recovering the 3D structures of the scene objects.[1–12] Hence, the goal is more in rendering than modeling. On the other hand, the modeling approach focuses on recovering the 3D scene structure.[13–27] By combining the inferred 3D structure with the recorded surface texture (appearance), it is possible to derive a model with correct 3D structure and photo-realistic appearance.

In more detail, a faithful 3D model must depict both an object's *structure* and *appearance* with high fidelity and accuracy. The structural attributes answer the "where" question, as in where an object should appear in an image, and the appearance attributes answer the "how" question, as in how an object should look like. A modeling algorithm must judiciously decide how the structure and appearance data are gathered, recoded, and used, as there is a trade-off between algorithmic complexity and model sophistication.

As just mentioned, the modeling approach is "*structure preferential*" where the 3D scene structure is explicitly recovered.[13–27] Structural information can be gathered using a hardware or software means. For example, a laser range finder may be employed to scan and record a single (or a few) dense 3D map of the environment, or more commonly, sophisticated 3D shape inference algorithms are used to deduce the 3D scene structure from video images of the environment. The rendering approach is "*appearance preferential*," where a large collection of images are acquired and stitched together into a 2D environment map.[1–12] Such a panoramic map captures the appearance of the environment without explicitly recovering the 3D depth.

The trade-off is in algorithmic complexity and model sophistication. A 3D model allows images of the environment to be rendered from arbitrary novel camera poses, not limited to depicting certain sub-regions within a fixed panorama. The disadvantage of the modeling approaches is in needing special hardware or sophisticated software to recover 3D information; a process that can be time consuming and prone to error.

While appearance (texture) information is explicitly encoded in a video image, structure (depth) information is not. However, depth perception is prevalent in the animal kingdom, particularly through stereopsis. The basic principle of stereopsis is well understood. The slightly different positions each eye perceives of a 3D object provide depth cue (disparity). A computational stereo algorithm proceeds by identifying the correspondences of image features in a pair of images (e.g., by matching features of a similar color, intensity, and texture), calculating the difference in the features' locations in these images (disparity), and using the disparity values to infer the surface depth. The time-consuming and error-prone step is in matching image features, which involves a search. The complexity of the search is drastically reduced if the cameras are placed in certain convenient configurations. For example, if the camera pair are positioned side-by-side so that the films are coplanar and the cameras are pointing in the same direction, disparity of a pixel (called the retinal or horizontal disparity in this standard stereo configuration) can be estimated by performing only a one-dimensional search in the other image along the camera's horizontal scanline.[28, 29]

When a single camera, instead of a pair of cameras, is used for image acquisition, a similar effect, called the motion parallax,[28, 29] provides useful depth cue—if the perspective of the camera is allowed to vary. Intuitively speaking, two views of a single camera from different perspectives can be considered a (sequential) stereo pair. Hence, the motion parallax, which measures how much an image feature has shifted after executing the camera motion, carries similar depth cue as the disparity. Useful motion parallax information is recorded using a moving camera if the imaged object is not infinitely far away and the movement of the camera is not a pure rotation about its optical center.[28–30]

While parallax and disparity are conceptually the same, the analysis of parallax is significantly harder than that of disparity. The difficulty is mainly due to two reasons: (1) The placement of the two cameras in stereopsis is fixed and known in advance, while the movement of the single camera in parallax is most likely *not* known (e.g., using a hand-held camcorder for architectural modeling). As the relative pose of the stereo cameras, or the movement of a single camera between two snapshots, is required in depth inference, the camera's motion needs to be inferred first in the parallax analysis. Inferring a camera's motion from images turns out to be quite challenging and a large body of literature has been reported,[28–30] and (2) In a mobile robotic application,

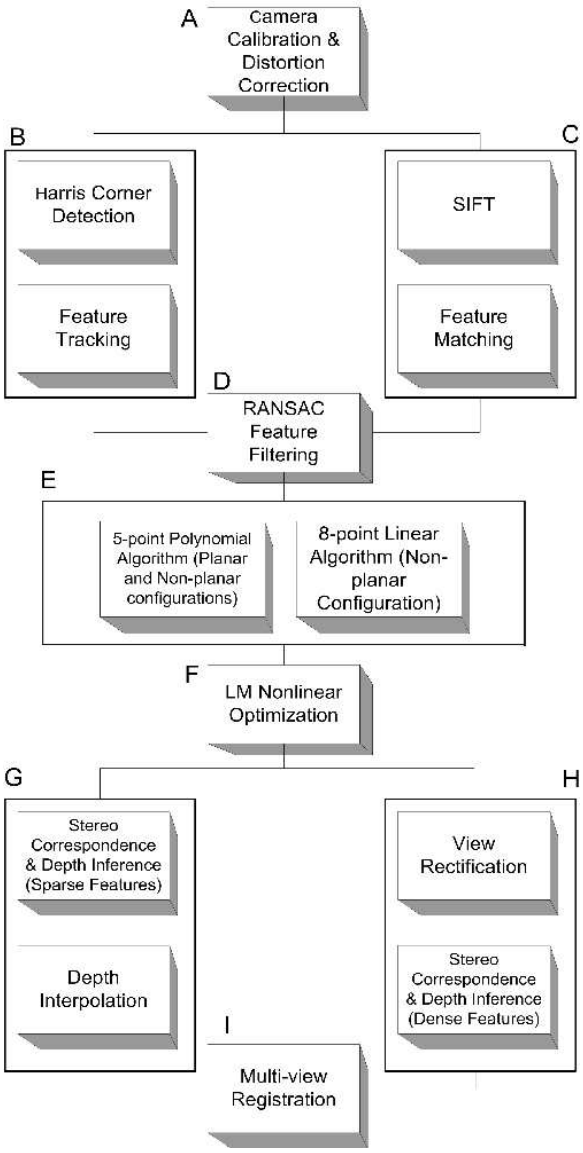

Figure 1. System flowchart for structure and appearance modeling.

the movement of the platform is often constrained by the environmental layout (e.g., the mobile platform must follow the hallways and corridors in a building). This implies that adjacent views are likely to be acquired in some awkward configurations. For example, if the mobile platform is navigating through a long hallway, adjacent views most likely are in a front-and-back configuration instead of a side-by-side one. Images must be rectified and image pixels rearranged somehow before the standard stereopsis analysis algorithms can be applied.

3. METHODS

Here, we summarize our work on modeling the static appearance and structure of body anatomy using the video of real endoscopic operations. This is to construct a 3D description of visible body anatomy with high visual realism from the operation video. The 3D model allows the observed anatomy to be rendered from arbitrary novel viewpoints and enables precise measurement of the size and location of anatomical anomalies, such as polyps and tumors. The flow chart of our algorithm is depicted in Fig. 1. Different modules in Fig. 1 and their functions are described in more detail below.

• *Camera calibration and distortion correction* (Box A): In video-endoscopy, difficulty in maneuvering the scope in a highly constricted body cavity results in large blind spots. Hence, endoscopic procedures frequently employ cameras using lenses of a short focus length (e.g., a fisheye lens) to provide a large view volume. Such lenses do cause significant distortion that degrades the accuracy of the constructed 3D models. Calibration is an offline process to estimate the intrinsic camera parameters and to correct image distortion, if any.[28,29]

• *Feature selection, tracking, and matching* (Boxes B and C): This step identifies prominent and semi-invariant features and matches these features across multiple images to establish their correspondences. We use two complementary paradigms: continuous tracking (box B) and discrete matching (box C).

With a high video frame rate, there is often trifling change in the appearance and position of image features in adjacent frames. We therefore detect prominent features (using the Harris corner detector[31]) and track their locations in images through a localized search operation. While a large number of trackers are available, we have opted to use an FFT-based tracker[32] that is accurate and achieves real-time performance for reasonably complex scenes.

If only a few isolated snapshots of the anatomy are acquired, changes in a feature's pose and appearance in these snapshots can be significant to render tracking infeasible. Instead, we compute SIFT features[33] which are insensitive to scale, location, orientation, and color change in images. We match these SIFT features in two images, regardless of their locations, to establish the feature correspondences.

• *Robust camera motion inference* (Boxes D, E, and F): This step uses the matched image features in two views to infer the camera's movement in between the views. The core process is either a 5-point polynomial algorithm or an 8-point linear algorithm.[29,30,34,35] The 5-point algorithm handles both planar and non-planar 3D configurations, and hence, is more general than the 8-point counterpart that fails if the 3D scene is planar. However, we did observe that 8-point algorithm provides more accurate results for front-and-back camera motion that is common in colonoscopy (this is confirmed in[34]).

The names of the inference algorithms refer to the minimum numbers of pairs of matched image features in two views that are needed for deducing the camera's motion parameters. In reality, we track and match significantly more features than just five or eight. Furthermore, tracking/matching results are necessarily imprecise due to noise and image quantization. Catastrophic failure in tracking (loss of trajectory) and matching (erroneous pairing assignments) do happen occasionally. To improve the robustness in camera motion inference, we use a nonlinear selection strategy called RANSAC (Random Sample Consensus).[36]

RANSAC operates on a hypothesis-and-verification paradigm. It randomly selects a minimum set of features (5 or 8) to compute the camera motion parameters (hypothesis), and then uses the remaining features to validate the results (verification). That is, if the camera motion parameters are accurately estimated based on the selected set of image features, they should put the two cameras in the correct stereo configuration where the majority of the remaining image features can be reasonably matched geometrically (*sans* those suffering catastrophic failure in tracking and matching). This process is repeated a number of times and the best hypothesis is retained.

Finally, nonlinear optimization (Box F) is used to give a final "polish" to the best result from RANSAC and the linear algorithms. We use the Levenberg-Marquart (LM) algorithm[29] which is essentially a combination of the Gauss-Newton method and gradient descent. While LM is a nonlinear iterative optimization procedure, convergence is fast because a good initial guess has been obtained (Boxes D and E) to guide the search.

• *Stereo rectification, matching, and depth inference* (Boxes G and H): This step is to infer 3D surface depth and construct a 3D model that captures both structure and appearance information. We consider two different approaches. In one approach (Box G), only the depths of the tracked/matched image features (from Boxes B and C) are explicitly computed to form a sparse depth map. Depths of the intermediate pixels are estimated through bi-linear interpolation[37] from those of the tracked/matched features. This approach is computationally efficient and works reasonably well if the anatomic structure is smooth.

A more accurate 3D model can be constructed by computing pixel disparity and inferring 3D depth at *each and every* pixel in the images (Box H). As mentioned, to efficiently and reliably performing the stereo analysis, the image pair should be in a standard side-by-side configuration. If not, the two images should be rectified first to rearrange the image pixels in such a way that the corresponding pixels in the two images lie on the same

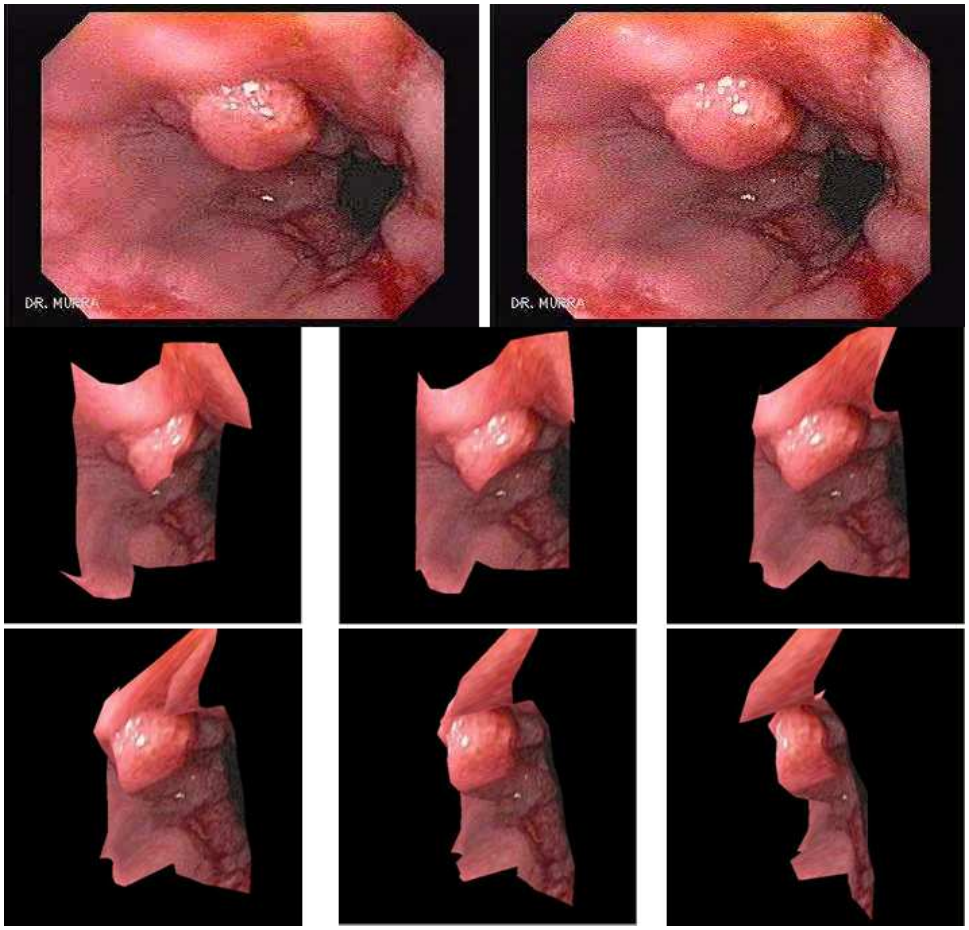

Figure 2. 3D reconstruction results using the sparse landmark approach.

image scan lines.[38] We then apply a stereo matching algorithm based on dynamic programming,[39] which takes into consideration pixel-, neighborhood-, and globally-based similarity criteria in matching.

• *Multi-view registration* (Box I): The final step is for registering partial 3D models constructed from multiple 2-view analyses into a more complete 3D model. We treat each partial 2-view model as a cloud of 3D points, and these point clouds are related by rigid-body transforms in space. We solve the registration problem through iterative refinement of the rigid-body registration parameters to match 3D point clouds with one another.[40] The process is efficient as we already have a good initial guess to the inter-frame camera movement from the previous step (Boxes D, E, and F).

4. EXPERIMENTAL RESULTS

We show below sample results using images from real colon exams. Figs. 2 and 3 show 3D models constructed using the sparse landmark approach (Box G). The first row shows the two images that are used in the model construction and the other rows show novel views synthesized from the model.

Figs. 4 and 5 depict results using dense stereo reconstruction and multi-view registration (Boxes H and I). Fig. 4 illustrates a 2-view modeling result. Fig. 4(a) summarizes stereo rectification, matching, and depth inference solutions. The first row in Fig. 4(a) depicts the stereo pair before rectification while the second row shows the pair after rectification. The rectification process rearranges the images by putting corresponding pixels on the same image scan lines in two images. The third row in Fig. 4(a) shows the stereo matching results. Stereo

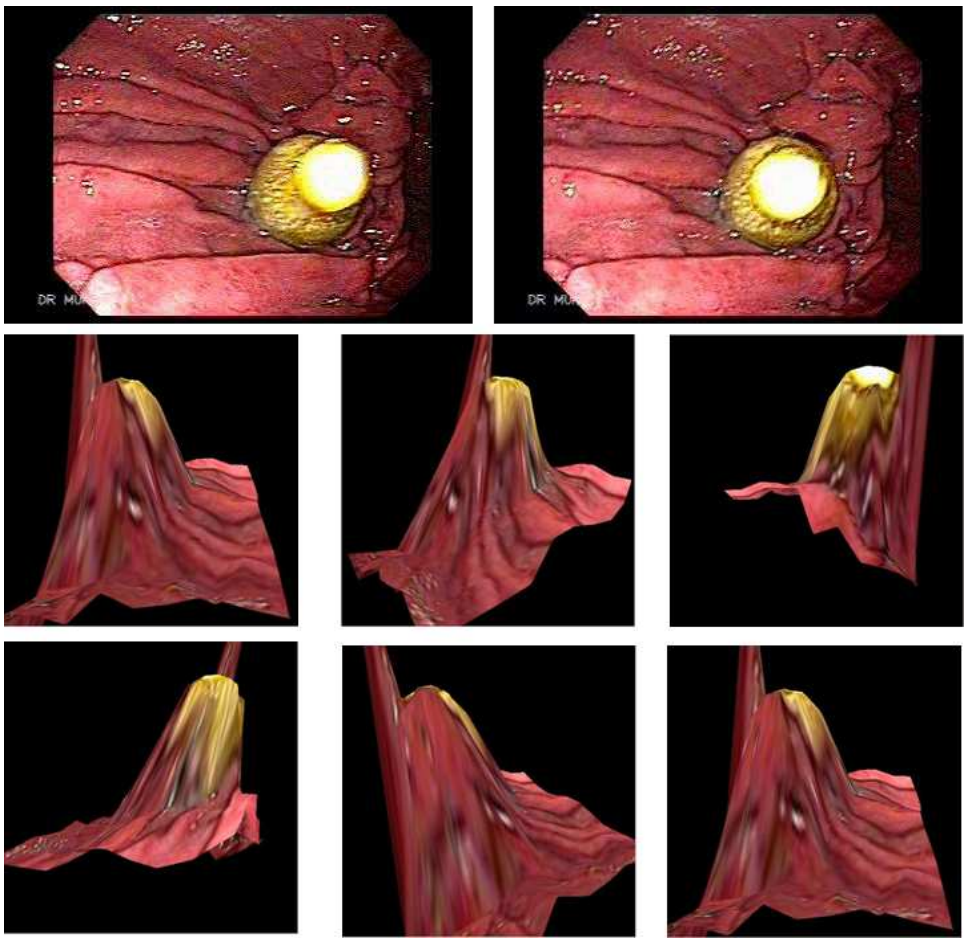

Figure 3. Another example of 3D reconstruction results using the sparse landmark approach.

disparity is displayed in gray scales on the left while the depth profile inferred from disparity is depicted on the right. Fig. 4(b) portrays sample novel views of the 3D computer model.

Fig. 5 shows a sample 3-view result. Two pairs of images (left column) are used in the 2-view analysis to construct two 3D profiles (middle column) that are then merged and smoothed (right column) in Fig. 5(a). Sample novel views rendered based on the computerized model are depicted in Fig. 5(b).

5. CONCLUDING REMARKS

This paper summarizes our on-going research in automated model building in colonoscopy. A 3D colon model is an essential component of a computer-aided diagnosis (CAD) system in colonoscopy. Such a model can be of many uses, for example, to assist surgeons in visualization, pre-op surgical planning, and surgical training. The ability to construct a 3D colon model from endoscopic videos (or images) is thus critical. Our future plan is to tackle long-range model construction with simultaneous camera movement and organ deformation.

REFERENCES

1. E. Adelson and J. Bergen, "The Plenoptic Function and the Elements of Early Vision," in *Computation Models of Visual Processing*, MIT Press, Cambridge, 1991.
2. S. E. Chen and L. Williams, "View Interpolation for Image Synthesis," in *ACM SIGGRAPH Conf. Proc.*, pp. 279–288, 1993.

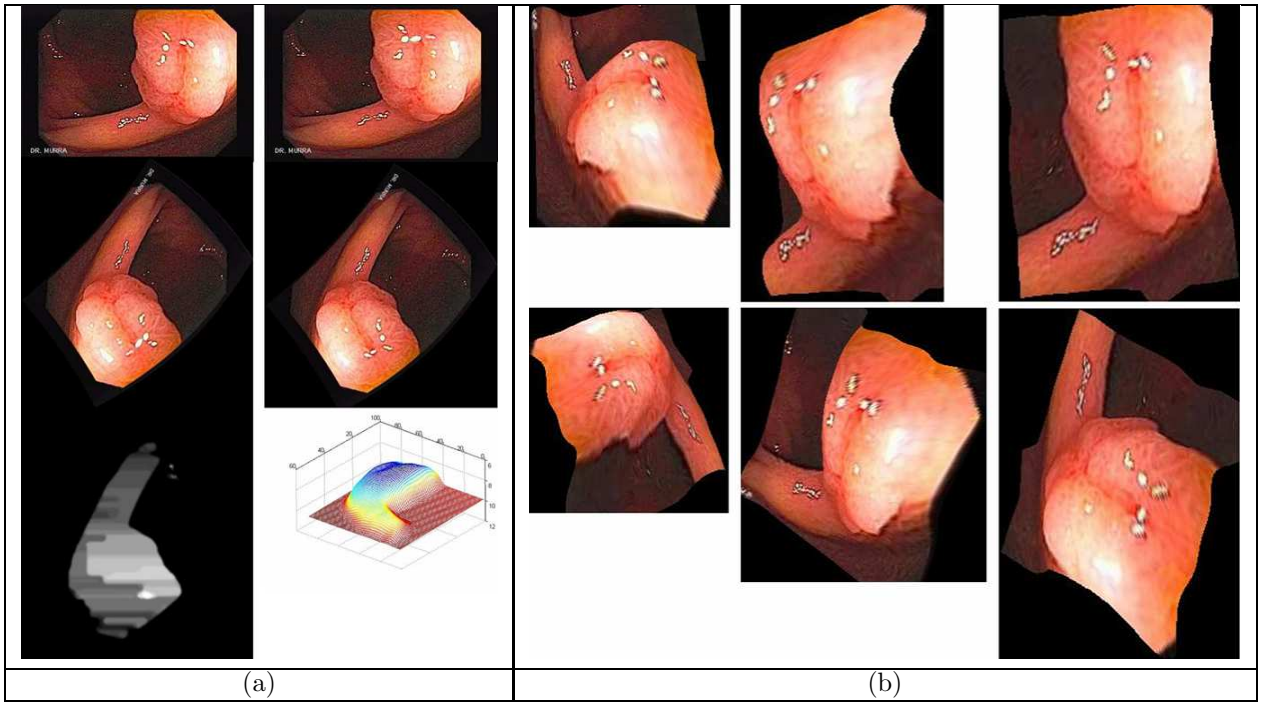

Figure 4. (a) Sample results on stereo rectification, matching, and depth inference, and (b) novel view rendering using models constructed in (a).

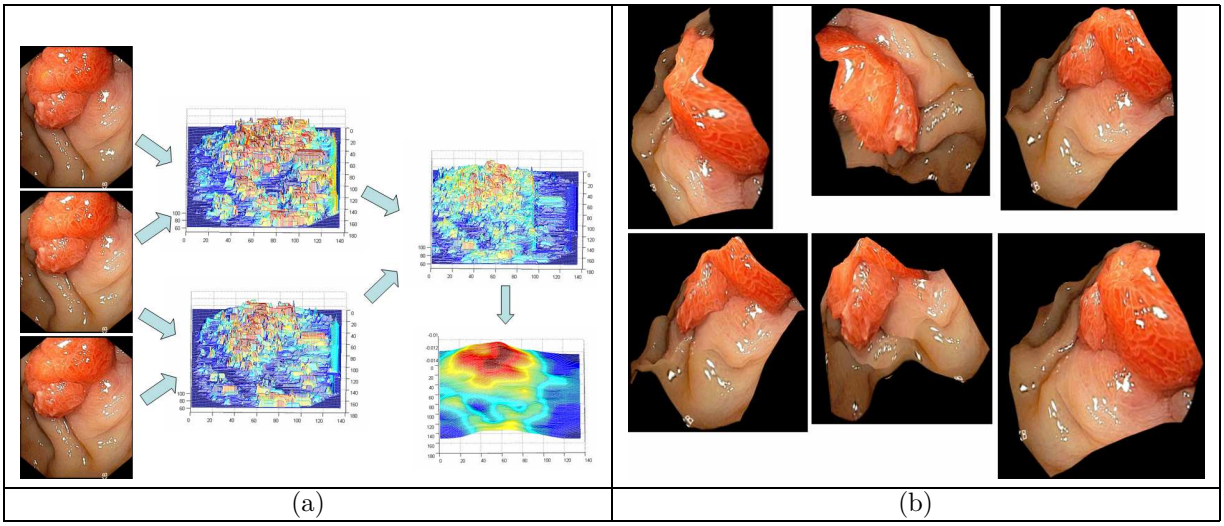

Figure 5. (a) Stereo matching, depth inference, and partial model registration results, and (b) novel view rendering using models constructed in (a).

3. S. E. Chen, "QuickTime VR: An Image-Based Approach to Virtual Environment Navigation," in *ACM SIGGRAPH Conf. Proc.*, pp. 1–10, 1995.
4. S. Gortler, L. we He, and M. F. Cohen, "Rendering from Layered Depth Images," Tech. Rep. MSTR-TR-97-09, Microsoft Research, 1997.
5. S. J. Cortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," in *ACM SIGGRAPH Conf. Proc.*, pp. 43–54, 1996.

6. M. Levoy and P. Hanrahan, "Light Field Rendering," in *ACM SIGGRAPH Conf. Proc.*, pp. 31–42, 1996.
7. J. X. Chai, X. Tong, S. C. Chan, and H. Y. Shum, "Plenoptic Sampling," in *ACM SIGGRAPH Conf. Proc.*, pp. 307–318, 2000.
8. L. McMillian and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," in *ACM SIGGRAPH Conf. Proc.*, pp. 39–46, 1995.
9. S. M. Seitz and C. Dyer, "View Morphing," in *ACM SIGGRAPH Conf. Proc.*, pp. 21–30, 1996.
10. H.-Y. Shum and S. B. Kang, "A Review of Image-based Rendering Techniques," in *IEEE/SPIE Visual Communications and Image Processing (VCIP)*, pp. 2–13, 2000.
11. H. Y. Shum and L. W. He, "Rendering with Concentric Mosaics," in *ACM SIGGRAPH Conf. Proc.*, pp. 299–306, 2000.
12. R. Szelisi and H. Y. Shum, "Creating Full View Panoramic Image Mosaics and Environmental Maps," in *ACM SIGGRAPH Conf. Proc.*, pp. 251–258, 1997.
13. J. D. Bonet and P. Viola, "Poxels: Probabilistic Voelized Volume Reconstruction," *Int. J. Comput. Vision* **35**, pp. 418–425, 1999.
14. P. Debevec, C. Taylor, and J. Malik, "Modeling and Rendering Architecture from Photograph: A Hybrid Geometry- and Image-Based Approach," in *ACM SIGGRAPH Conf. Proc.*, pp. 11–20, 1996.
15. A. Dick, P. Torr, and R. Cipolla, "Modelling and Interpretation of Architecture from Several Images," *Int. J. Comput. Vision* **60**, pp. 111–134, 2004.
16. C. Dyer, *Volumetric Scene Reconstruction from Multiple Views*, pp. 469–489. Foundation of Image Understanding, Kluwer, 2001.
17. K. Kutulakos and S. Seitz, "A Theory of Shape by Space Carving," *Int. J. Comput. Vision* **38**, pp. 199–218, 2000.
18. D. Morris and T. Kanade, "Image Consistent Surface Triangulation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, pp. 332–338, 2000.
19. P. Narayanan, P. Rander, and T. Kanade, "Constructing Virtual Worlds using Dense Stereo," in *Proc. Int. Conf. Comput. Vision*, pp. 3–10, 1998.
20. M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual Modeling with a Hand-held Camera," *Int. J. Comput. Vision* **59**, pp. 207–232, 2004.
21. C. J. Taylor, "Surface Reconstruction from Feature Based Stereo," *Int. J. Comput. Vision* , pp. 184–190, 2003.
22. S. Seitz and C. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring," *Int. J. Comput. Vision* **35**, pp. 151–173, 1999.
23. S. Seitz, B. Curless, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, 2006.
24. S. Sinha and M. Pollefeys, "Multi-View Reconstruction using Photo-Consistency and Exact Silhouette Constraints: A Maximum-Flow Formulation," in *Proc. Int. Conf. Comput. Vision*, pp. 349–356, 2005.
25. C. Tomasi and T. Kanade, "Shape and Motion from Image Streams under Orthography—a Factorization Method," *Int. J. Comput. Vision* **9**(2), pp. 137–154, 1992.
26. D. Scharstein and R. Seliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int. J. Comput. Vision* **47**, pp. 7–42, 2002.
27. N. Snavely, S. Seitz, and R.Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," in *ACM SIGGRAPH Conf. Proc.*, pp. 332–338, 2006.
28. G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluwer Academic Publishers, The Netherlands, 1996.
29. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Camridge, MA, 2003.
30. O. Faugeras, *Three-Dimensional Computer Vision*, MIT Press, Cambridge, MA, 1993.
31. C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *Fouth Alvey Vision Conference*, pp. 147–151, 1988.

32. D. Koppel, Y. F. Wang, and H. Lee, "Image-Based Rendering and Modeling in Video-Endoscopy," in *Proc. of IEEE Int. Symp. on Biomedical Imaging*, pp. 272–279, (Arlington, VA), April 2004.
33. D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision* **60**, pp. 91–100, 2004.
34. D. Nister, "An Efficient Solution to the Five-Point Relative Pose Problem," *IEEE Trans. Pattern Analy. Machine Intell.* , pp. 756–770, 2004.
35. R. I. Hartley, "In Defense of the Eight-Point Algorithm," *IEEE Trans. Pattern Analy. Machine Intell.* **19**, 1997.
36. M. Fischler and R. Bolles, "RANdom Sample Consensus: A Paradigm for Modeling Fitting with Application to Image Analysis and Autoamted Cartography," *Communications of ACM* **24**, pp. 381–395, 1981.
37. G. Farin, *Curves and Surfaces for Computer Aided Geometric Design*, Academic Press, San Diego, CA, 1988.
38. M. Pollefeys, R. Koch, and L. V. Gool, "A Simple and Efficient Rectification Method for General Motion," in *Proc. Int. Conf. Comput. Vision*, 1999.
39. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, Second Edition*, MIT Press, Cambridge, MA, 1990.
40. P. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," *IEEE Trans. Pattern Analy. Machine Intell.* **14**, pp. 239–256, 1992.