IMAGE-BASED RENDERING AND MODELING IN VIDEO-ENDOSCOPY*

Dan Koppel and Yuan-Fang Wang

Department of Computer Science
University of California
Santa Barbara, CA 93106

Hua Lee

Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106

ABSTRACT

We present algorithms for image-based rendering and modeling in video-endoscopy. Our algorithms are aimed at alleviating two common viewing problems in video-endoscopy: that the scope view is often distorted and oriented wrongly (a dis-orientation problem), and that the scope view is acquired from a viewpoint that deviates from the surgeon's (a dis-association problem). Our solutions strive to alleviate these problems to arrive at a more "open-surgery-like" and "surgeon-centered" display. We present experimental resutls based on real endoscopic video to illustrate our image enhancement procedures.

1. INTRODUCTION

Our research is aimed at enhancing the visual feedback to the surgeon in endoscopy to shorten the operation time, improve patient safety, and achieve cost savings in health care. Video-endoscopy, a mode of minimally-invasive surgery, has proven to be significantly less invasive to the patient. However, it creates a much more complex operation environment that requires the surgeon to operate through a video interface. Visual feedback control and image interpretation can be challenging and non-intuitive. We identify two significant visualization problems in endoscopy: that the scope view is often distorted and oriented wrongly (a dis-orientation problem), and that the scope view is acquired from a viewpoint that deviates from the surgeon's (a dis-association problem). In this paper, we present our solutions in alleviating these problems and some preliminary results.

2. METHODS AND RESULTS

Here, we present our current work on image rectification and image-based rendering (IBR) in endoscopy. Image rectification is needed to alleviate the dis-orientation problem by maintaining the "head-up" display at all times, regardless of the physical maneuver (panning and rotation) of the endoscopy. IBR is used for stitching together the operation video to construct a 3D description of the visible portion of the operation cavity with high visual realism through image-based texture mapping. This 3D description allows the visible portion of the operation cavity to be rendered from a viewpoint that is closer to the surgeon's perspective to alleviate the dis-association problem.

2.1. Image Rectification

The image Rectification algorithm comprises three stages:
1.) In the first stage, a few image features are selected and tracked in endoscopic video. The features are areas of the image that have a high number of edges or corners with a high intensity contrast. The correspondences of image features are established using an affine model.
2.) The second stage takes the 2D coordinates of the corresponding features in successive frames and estimates the camera motion parameters, using the 8-point algorithm [1, 2]. Here the coordinates of the centers of the tracked 2D features are taken as inputs to the 8-point algorithm. This stage recovers the perceptual depth of the tracked features and the camera transformation.
3.) Finally, in the third stage, we infer the direction of an abstract "head-up" vector in the camera's current reference frame, based on the recovered camera transformation parameters. The head-up vector is the one that is located directly in front of the camera and is pointing in the "up" direction in the *surgeon's* fame of reference. Knowing the vector's 3D coordinates in the *camera*'s current frame allows us to project it onto the image plane. The deviation of this projection from the image's y-axis tells us by how much the image should be rotated to make the arrow appear again as "up" on the screen. Execution of this rotation then rectifies the camera frame, as well as the entire image, to confirm the surgeon's frame of reference.

While the rectification algorithm is rooted in traditional computer vision, *our contribution is in significantly improving the efficiency and robustness of the traditional techniques to suit this new application domain.* In particular, we reformulated the 2D tracking problem using Fourier analysis and were able to achieve over 100-fold speed increase over the conventional spatial hierarchical correlation-based techniques [1, 2]. This speed up is very significant, as the bottleneck in image rectification is in 2D tracking. As the 3D structure inference step takes negligible time comparing to the 2D tracking step, and the actual rectification of images can be efficiently performed using the graphics processor in modern-day PCs. Furthermore, this improvement is achieved without any special hardware (e.g., DSP) acceleration. With hardware acceleration, even greater speedup is possible for potentially real time or near real time tracking. One last thing to note is that the tracking method is general for all surface types and does not require any special object model to guide the search.

For 3D analysis, our formulation corrects an oversight in the conventional 8-point algorithm that makes it susceptible to numerical error when the tracked points assume an approximately planar configuration. In video-endoscopy operations, the camera often times moves very close to an organ or the abdominal wall (for a closer inspection by the surgeon). Hence, the configuration of the tracked features easily reduce to a (nearly) planar one. Without our correction, the recovered camera parameters will be susceptible to large numeric errors so as to render rectification unreliable.

* The research was support in part by Karl-Storz Imaging, Inc. and the State of California Micro Program.

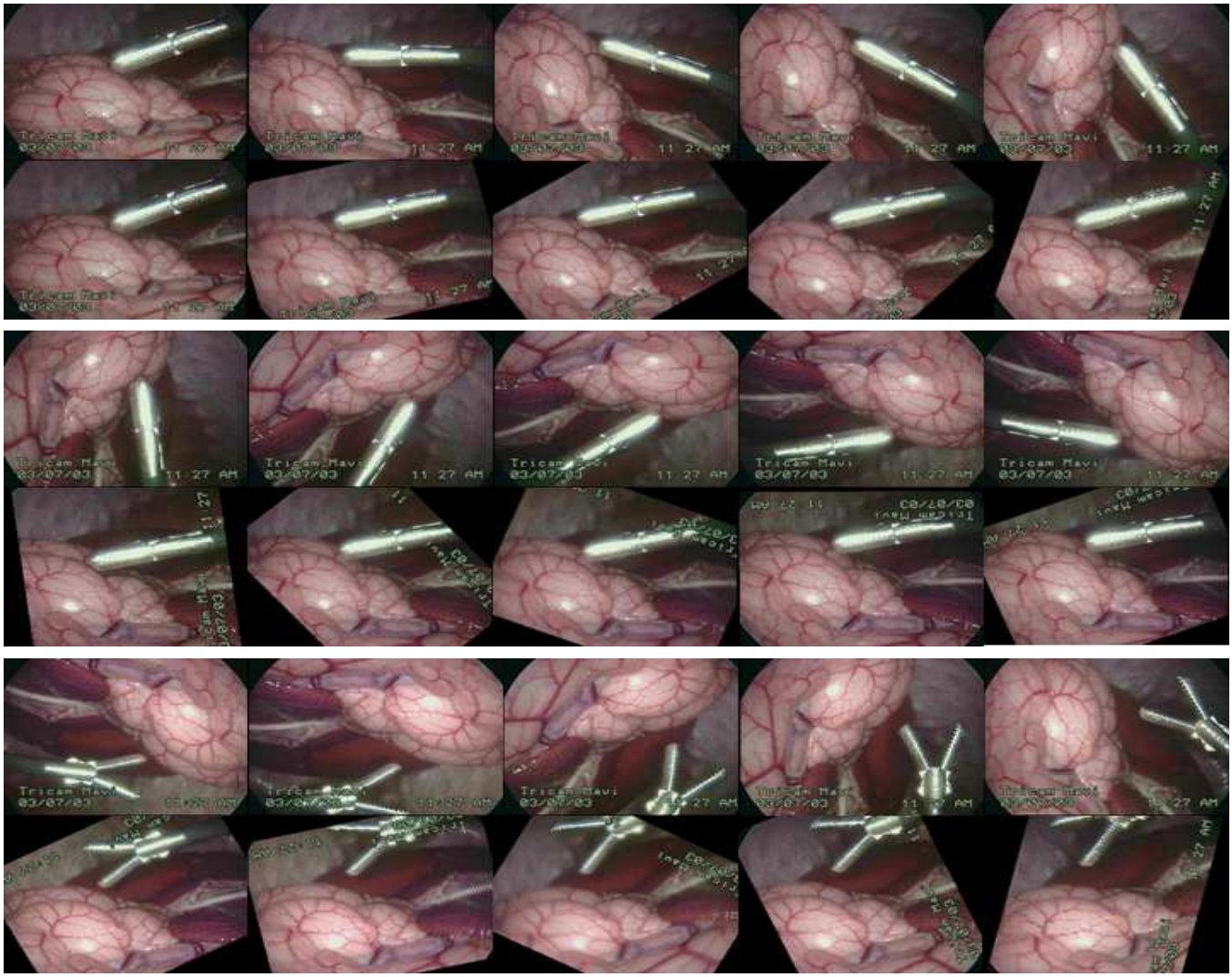


Fig. 1. Original and rectified sequences (top and bottom rows, respectively) from real endoscopy surgery

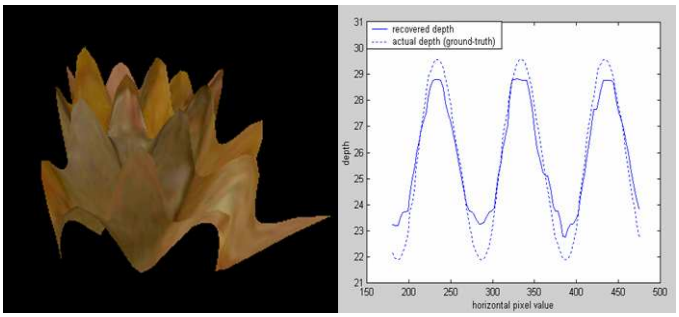
Furthermore, we employ redundancy and robust error norm to significantly improve the accuracy and minimize the possibility of loss of track.

In Fig. 1, video obtained from real endoscopy surgery inside an abdominal cavity was used for rectification. Here the camera executed general movement in which all degrees of motion freedom were exercised. Additionally, a free-moving instrument was present in the field of view which further complicated the rectification process. Fig. 1 shows sample frames from an original and a rectified sequence. The original and rectified images are shown in pairs of two rows: The top rows show the original sequence without rectification and the bottom rows show the rectified results. As can be seen that the views in the bottom rows stay very stable even with large rotation of view and instrument movement.

2.2. Image-Based Rendering

To alleviate the dis-association problem it would be beneficial that photo-realistic views of the operation cavity be rendered from a viewpoint that is closer to the surgeon's perspective. Generally speaking, faithful rendering of 3D models must depict both an object's *geometry* and *appearance* with high fidelity and accuracy. As far as image rendition is concerned, the geometric 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 geometry and appearance data are gathered, recoded, and used, as there is a trade-off, in terms of storage and computation, between recording an object's geometry and appearance.

On one extreme (which we call "*structure preferential*"), a laser range finder may be used to record a single (or a few) dense 3D maps and video images of the environment. On the other extreme (which we call "*appearance preferential*"), a large collection of images can be acquired and stitched together into a 2D environmental map. [3, 4, 5, 6, 7, 8, 9]. Such a map captures the appearance of the environment without any explicit 3D structural infor-



(a) (b)

Fig. 2. (a) Synthetic data (with known ground-truth) used to validate algorithm, and (b) comparison of recovered depth and ground-truth for a slice of the 2D surface

mation. Novel images are then re-sampled and interpolated from the environmental map. The tradeoff is that, in the structure preferential approaches, 3D depth is explicitly recorded, which allows images of the environment rendered relatively effortlessly from novel camera poses (through re-projection and re-sampling). The disadvantage of the structure preferential approaches is in needing special hardware or software to recover 3D information from video and time consuming 3D sampling processes.

Here we show that a middle-of-the-road (or best-of-both-worlds) approach can be advantageous compared to the extremes. An observation we make is that 3D worlds of interest are seldom random, and a high degree of regularity and redundancy do exist in structures. Structural regularity shows up in images as homogeneous regions (in terms of shading, color, and texture) with smooth, well-defined boundaries separating them. Instead of recovering and recording structural information pixel-by-pixel using, say, a laser range finder or other specialized hardware, *we can exploit scene regularity and redundancy to recover 3D structural information in a computationally sufficient and efficient manner.* Structural regularity allows segmenting images into regions, of which 3D structures can be inferred as a whole with a minimum amount of effort to insure efficiency.

Comparing to the structure preferential approaches, the advantage of our approach lies in that no specialized hardware (such as a laser range finder) and slow 3D sampling processes are required. Comparing to the appearance preference approaches, we infer and record just adequate, or sufficient, 3D information to allow efficient geometry and appearance computation from novel views. Efficient 3D inference requires robust and effective segmentation and feature tracking techniques. As mentioned, our tracking techniques achieved near real-time performance when implemented in software. Both segmentation and tracking algorithms are amenable to hardware acceleration. These algorithms thus allow us to recover an adequate amount of 3D structures in a computationally effective way using a small number of images and without restriction on the imaging configuration. The algorithm for image-based rendering comprises three stages:

1.) *Segmentation* The difficulty of image segmentation is due mainly to the "busy-ness" of image contents and corruption by image noise. Hence, a segmentation algorithm that can key in on major image structures (edges) without being distracted by minor, non-essential brightness perturbation and noise will be very valuable. Our approach is to use diffusion to eliminate nonessential region

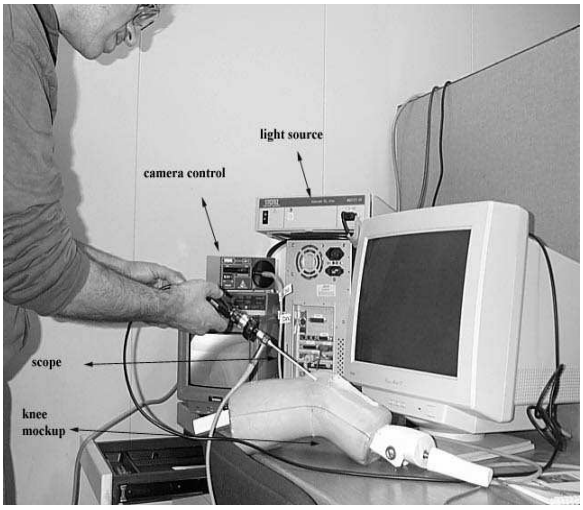


Fig. 3. Experiment platform of a knee mockup, a camera and control, a scope, and a light source.

details while preserving region boundaries [10]. The choice of a diffusion-based framework is made because it is highly (or even trivially) parallelizable and amenable to hardware acceleration.

2.) *Shape inference and interpolation* After images are segmented, the 3D shapes of the segmented regions are recovered and interpolated. The interpolation algorithm comprises three stages. The first two stages are for feature extraction and tracking and 3D depth inference. These were described above in Sec. 2.1. Thirdly, based on the recovered 3D feature positions, we interpolate the depths of all other points in the region based on a low-order shape equation.

3.) *Mosaic building* Once the images are segmented with 3D structures recovered for each segmented region, building a mosaic becomes a relatively straightforward process. This is because (i) for inferring 3D structures, we already select and track distinct features in each region. Such features serve as prominent landmarks for registering regions from different video frames, and (ii) the 3D inference process recovers both the perceptual depths of tracked features and the movement of the camera. Hence, the geometric transform between different camera frames are readily available that allows 3D structures recovered in video to be related to one another in a common reference frame.

We validated our IBR procedure by testing it on synthetic data with known ground-truth values first. A real image (obtained from an endoscopic video sequence) serves as the texture mapped onto a geometric model. A second image of the synthetic model is generated from the first one, consistent with a specified depth map and camera movement. The geometric model used here had an "egg-carton-like" structure (where the depth equals the product of two sinusoids - one horizontal and the other vertical - see Fig. 2.a). Using the two images as input, the algorithm was able to deduce a depth map that could be compared to the ground-truth used to generate the second image. The results are shown in Fig 2.b. Since the depth maps require three dimensions to display, we show only a slice of each map. More specifically, we choose a horizontal slice (with vertical pixel coordinate fixed at 230). The figure shows a close correspondence between the ground-truth value and the re-

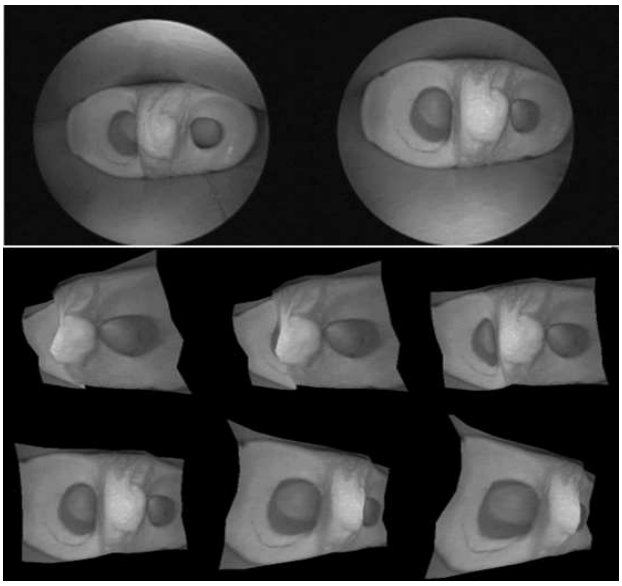


Fig. 4. Top: images used in computing depth map from the inside of a knee mockup. Bottom: several novel views inferred by the algorithm and rendered with perspective projection. (Note that the algorithm's input images differ by only a small angle.)

covered depth.

The IBR algorithm was then tested using both real endoscopy images and images from inside of a knee mockup as shown in Fig. 3. The video was generated by a Karl-Storz camera (Telecam 20212130U), hooked up by S-video to an ATI All-In-Wonder capture card in a PC. Fig. 4 shows one example of the IBR algorithm. Real video (from inside the knee mock-up provided by Karl Storz Imaging, Fig. 3) was used in Fig. 4, where the top row shows the two images processed by the algorithm. Based on these two images, we were able to construct a 3D model of the visible part of the knee cavity. On the bottom row, several novel views are displayed (using perspective projection). After tracking key features and inferring their depths, views of the operation cavity can be rendered from a new angle.

In Fig. 5, we tested the algorithm on video obtained from real endoscopic procedures. We constructed depth maps that allowed renderings of tissue from new angles. Tissue with protrusions exhibited the occlusions characteristic of non-planar geometry, as rotated novel views were calculated and displayed. A sample shot of this tissue from a mesh hernioplasty is shown in Fig. 5. As the virtual rotation is performed, occlusion of the left-most tissue by the grey structure is apparent. These results illustrate how an accurate depth estimation allows rendering from novel vantage points.

3. CONCLUSION

This paper summarizes our current research on viewing enhancement for video-endoscopy and present our preliminary results. Our future work is to further enhance the IBR techniques to merge and register object description constructed using IBR with that from the Visual Human Dataset for unrestricted, assisted viewing.

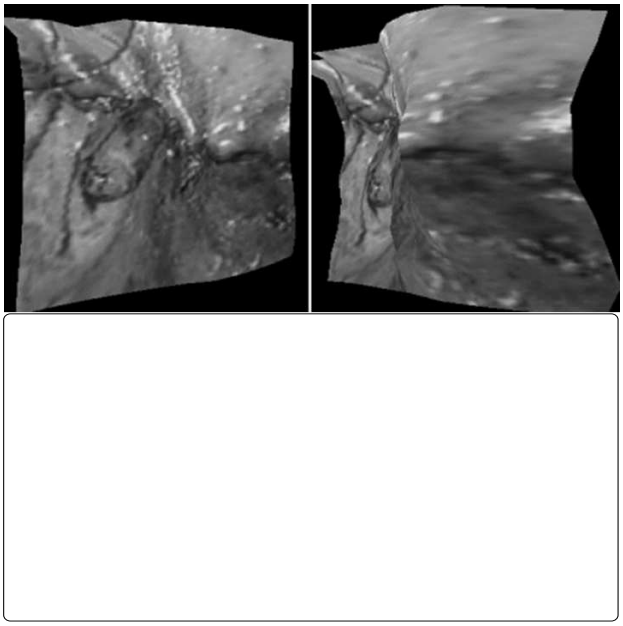


Fig. 5. Views of tissue from a mesh hernioplasty and a rotationally re-rendered view

4. REFERENCES

[1] O. Faugeras, *Three-Dimensional Computer Vision*, MIT Press, Cambridge, MA, 1993.

[2] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluwer Academic Publishers, The Netherlands, 1996.

[3] S. E. Chen, "QuickTime VR: An Image-Based Approach to Virtual Environment Navigation," in *SIGGRAPH Conf. Proc.*, 1995, pp. 1–10.

[4] S. J. Cortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," in *SIGGRAPH Conf. Proc.*, 1996, pp. 43–54.

[5] M. Levoy and P. Hanrahan, "Light Field Rendering," in *SIGGRAPH Conf. Proc.*, 1996, pp. 31–42.

[6] J. X. Chai, X. Tong, S. C. Chan, and H. Y. Shum, "Plenoptic Sampling," in *SIGGRAPH Conf. Proc.*, 2000, pp. 307–318.

[7] L. McMillian and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," in *SIGGRAPH Conf. Proc.*, 1995, pp. 39–46.

[8] S. M. Seitz and C.R. Dyer, "View Morphing," in *SIGGRAPH Conf. Proc.*, 1996, pp. 21–30.

[9] R. Szelisi and H. Y. Shum, "Creating Full View Panormic Image Mosaics and Environmental Maps," in *SIGGRAPH Conf. Proc.*, 1997, pp. 251–258.

[10] P. Liang and Y. F. Wang, "Local Scale Controlled Anisotropic Diffusion with Local Noise Estimate for Image Smoothing and Edge Detection," in *Proceedings of International Conference on Computer Vision*, Bombay, India, 1998, pp. 193–200.