35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Markerless Tracking for Augmented Reality for Image-Guided Endoscopic Retrograde Cholangiopancreatography

Thinh T. Nguyen, Hoeryong Jung, and Doo Yong Lee, *Senior Member, IEEE*

***Abstract*—This paper proposes a markerless tracking method with adaptive pose estimation for augmenting 3D organ models on top of the endoscopic image for Endoscopic Retrograde Cholangiopancreatography (ERCP). While many applications of augmented reality (AR) to surgeries need special markers to track the camera's position and orientation in the live video, our method employs the feature detection techniques to track the endoscopic camera. One of the most difficult problems when applying feature-based method to AR for ERCP is the lack of texture & highly specular reflection surface of duodenum in the endoscopic images, which does not provide a stable number of keypoints to track in the endoscopic video sequence. By introducing an adaptive weight function in the combination of reference-current frame tracking with previous-current frame tracking, we enhance the tracking performance remarkably. The proposed method is evaluated using an endoscopic video of a real ERCP and 3D duodenum model reconstructed from CT data of the patient. The result shows real-time performance and robustness of the method.**

I. Introduction

ERCP is an image-guided procedure used to treat the problems in bile ducts and pancreatic ducts [1] which has been widely used for a long time. In an ERCP procedure, an endoscope with a small camera attached on its tip is navigated through the patient's mouth passing the stomach into the duodenum. Then the doctor needs to locate the major papilla, and inserts a catheter into the patient's bile duct or pancreatic duct. A dye is injected into the common bile duct and pancreatic duct to visualize the two ducts in the X-ray image. However due to the limitation of the endoscopic camera view, the doctor often finds it difficult to locate the major papilla and insert the catheter into this major papilla.

One way to improve the vision of the doctor is to provide 3D view of hidden organs on top of the endoscopic image by using augmented reality techniques. Most of successful methods for augmented reality surgery employed fiducial markers for tracking the organ which can be considered as the most stable and reliable approach as in [2], [3], [4]. Using fiducial marker is easy and accurate approach, however sometimes it may be expensive or impossible to put any markers on the objects like in ERCP surgery. Few research related to registering 3D model onto endoscopic image can be found in [5], [6], [7]. For example, Yeny Yim el at., [6] has proposed an approach that find the optimal viewpoint of virtual camera by maximize the mutual information between 2D endoscope image and 2D rendered CT image from virtual camera. The use of mutual information can be a good candidate for initialize the optimal viewpoint of camera but it is quite computationally expensive and it has not been tested to find the real endoscopic camera view point. Another approach is using feature-detection techniques which provide an excellent tool to visual tracking without having to use any external markers. However, most of studies have been applied to track objects which have rich texture as in [8], this approach to augmented reality surgery where images often have poor texture should be explored more.

The present paper presents a markerless camera tracking method for registering 3D model on top of endoscopic image in ERCP. Our method utilized state-of-the-art feature detection techniques to track the endoscopic camera without using any external markers. We also introduce an adaptive pose estimation which handles the lack of textures of endoscopic images of duodenum. The preliminary result shows a robust, real-time tracking performance with fusion image.

II. Methods

The method developed in this study was based on model-based tracking approach and is shown in Figure 1. The

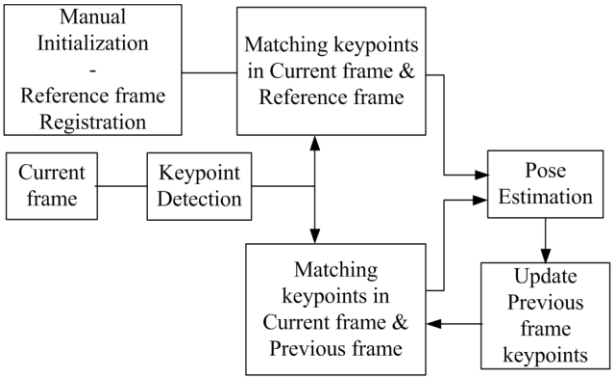


Figure 1. Overview of camera tracking procedure.

endoscopic camera pose was estimated for each frame in the video sequence by tracking keypoints in the endoscopic video sequence provided that the 3D locations of these features are known from 3D reconstructed model of duodenum.

To provide a reference of these keypoints and their 3D corresponding location, before the camera tracking could be triggered, a manual initialization was done by simply selecting some corresponding points between a virtual image of duodenum and a chosen endoscopic image. The keypoints in the chosen endoscopic image (reference frame) was tracked in

* Research supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 20110029043).

Thinh T. Nguyen is with the Korea Advanced Institute of Science and Technology, Daejeon, Korea(e-mail: n_2t@kaist.ac.kr).

Hoeryung Jung is with the Korea Advanced Institute of Science and Technology, Daejeon, Korea (e-mail: junghl80@kaist.ac.kr)

Doo Yong Lee is with the Korea Advanced Institute of Science and Technology, Daejeon, Korea (e-mail: leedy@kaist.ac.kr).

current frame to estimate the camera pose. To enhance the stability we also tracked the keypoints in previous frame; the 3D locations of the keypoints in previous frame were updated each time the new camera pose was calculated. Two sets of matched keypoints in current frame tracked from previous frame and reference frame were used to estimate the current camera pose.

A. Keypoints detection

The choice of keypoint detection and descriptor was considered with care because the poor texture of duodenum surface. We adopted SURF[9] for keypoint detection and FREAK[10] for keypoint description. For the sake of computation a GPU version of SURF was used to detect keypoints in each frame. In our experiments, SURF was the most suitable to use with FREAK descriptor. The advantage of using FREAK binary-keypoint descriptor was that it does not only generate robust feature but it also extremely faster in matching these binary-keypoints compared with traditional descriptors such as SIFT[11] and SURF. Another problem of endoscopic image is the highly specular surface of duodenum which creates many specular regions on the image of duodenum, which leads to many unexpected keypoints. One way to reduce the effect of this highly specular effect is to remove keypoints detected in that regions by using a thresholding mask as in Fig 2.

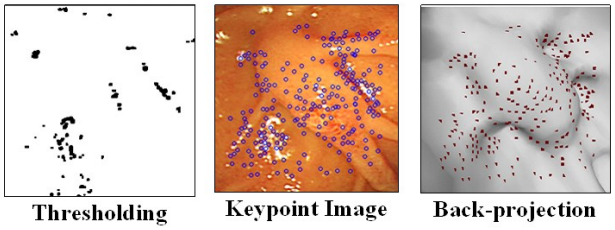


Figure 2. Keypoint detection and back-projection

B. Current frame and previous frame tracking

The matched keypoints between current frame and previous frame are also added to calculate camera pose together with the matched keypoints between current frame and reference frame for smoother estimation. To use the previous frame keypoints that we generated approximate 3D locations of keypoints in each frame after the camera pose was updated as in [8]. Figure 2 shows keypoints detected in endoscopic camera and 3D locations of the keypoints generated by OpenGL back-projection on virtual image. However, instead of using Normalized Cross Correlation as in KLT tracker [12], we took the advantage of binary keypoint to match keypoints in two consecutive frames explicitly which produced more stable matches.

C. Pose Estimation.

The endoscopic camera pose of current frame is estimated using two set of 3D/2D points obtained from matches between current frame and each of 2 frames: reference frame and previous frame. Suppose that we have got the previous camera pose at frame *k-1* which is parameterized as a 6-element vector as in (1).

$$p^{(k-1)}=[rx,ry,rz,tx,ty,tz] \quad (1)$$

By tracking the keypoints of reference and previous frame in the current frame *k*, we estimate the current camera pose as follows:

$$p^{(k)}=\text{argmin}\sum_{i}^{N}\rho\left(\text{proj}(x_i,y_i,z_i,p)-(u_i,v_i)\right)+\sum_{j}^{M}\rho\left(\text{proj}(x_j,y_j,z_j,p)-(u_j,v_j)\right). \quad (2)$$

where N is the number of correspondences between first frame and current frame, while M is the number correspondences between the current frame and reference frame. The 3D locations of keypoints in the reference frame are (x_i,y_i,z_i) while the corresponding projected locations in the current frame are (u_i,v_i). The projection of each point (x_i,y_i,z_i) onto imaging plane using camera parameter p is $\text{proj}(x_i,y_i,z_i,p)$. Similarly, (x_j,y_j,z_j) are 3D points in previous frame and (u_j,v_j) are their corresponding image points in current frame. We adopted Tukey M-estimator function ρ in [13] for a robust pose estimation. In order to solve this robust estimation a Levenberg-Marquat non-linear optimization was utilized. The problem of Tukey estimator could be that it creates many local minimums. Therefore to have a good result with Tukey estimation we applied RANSAC algorithm[14] to eliminate outliers and provide a good initial guess for Tukey Estimation.

D. Pose estimation with adaptive weight function.

Although using RANSAC and M-estimator could reduce the effect of outliers, misestimating is unavoidable due to error accumulation and wrong matches. Because the 3D locations keypoints generated in previous frame are extracted from an estimated pose, if the previous pose is not correct the error will be added to the estimate pose of next frame. Another problem is that the number of correct matches between current frame and reference frame is really unstable caused by the highly specular reflection of the surface of duodenum which can be seen in Figure 3.

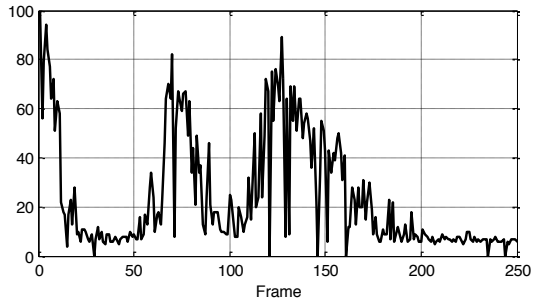


Figure 3. Number of good matches between current frame and reference frame.

While the number of matches between current and previous frames are really large because the two frames are almost similar. These differences between the two input datasets could lead to the drift and jitter of pose estimator. In this study we have proposed a better approach to deal with these problems by limiting the number of matches between 2 consecutive frames and modifying the weight associated with matches between current frame and reference frame in the optimization function in (2) as follows:

$$p^{(k)}=\text{argmin}\ \lambda\sum_{i}^{N}\rho\left(\text{proj}(x_i,y_i,z_i,p)-(u_i,v_i)\right) + \sum_{j}^{Mmax}\rho\left(\text{proj}(x_j,y_j,z_j,p)-(u_j,v_j)\right) \quad (3)$$

In this equation, Mmax is the number of best matches between current frame and previous frame, and λ is a function of the number of good matches between current frame and reference frame N as follows:

$$\lambda(N)=\begin{cases}\lambda_0 & \text{for } N<Nmin\\ \lambda_0+\left(\frac{N-Nmin}{Nmax}\right)^2 & \text{for } Nmin<N<Nmax\\ \lambda_0+\left(\frac{Nmax-Nmin}{Nmax}\right)^2 & \text{for } N>Nmax\end{cases} \quad (4)$$

The weight function was chosen so as to have high value when then number of good matches is high and low when the number of good matches are low because reliably of these matches is correlated with the number of good matches, which were obtained by a coarse pose estimation with RANSAC scheme using previous pose as the initial.

III. Results

The method was evaluated on one data set provided by a hospital including a set of abdominal CT scan images and a recorded ERCP video of the same patient. Only model of duodenum is needed for tracking evaluation purpose. Before the ERCP procedure, abdominal CT scan images were taken. Water had been plumped into the duodenum to expand the duodenum as it was in the real operation. We used Mimics software to reconstruct the 3D model of duodenum. The acquired CT scan images has slice increment 1.5mm and pixel size 0.679mm the duodenum model used in this experiment has bounding box of about 71x65x114(mm3). Endoscopic camera was calibrated and virtual camera was set up [10] to fit with this camera model [11]. Depending on the graphics API library, different setups can be applied. The implementation was tested in a PC with Intel® Core™ i7 CPU 860 @2.80Hz, RAM: 4.00 GB, Windows 7, Nvidia GTX 580 graphics card. The graphical user interface provides 4 view ports to user including virtual image window, real image window, fusion image window and a 3rd person view window. To initialize camera pose, the user need to choose reference frame and select corresponding points between real image and virtual image. Then the camera tracking can be triggered immediately. The fusion image window shows an image where 3D organ models are overlapped on endoscopic image. And 3rd person window shows the camera position and orientation with respected to duodenum model. To evaluate, we compared the tracking result with the true value which was created by manually registering for each frame like in initialization step.

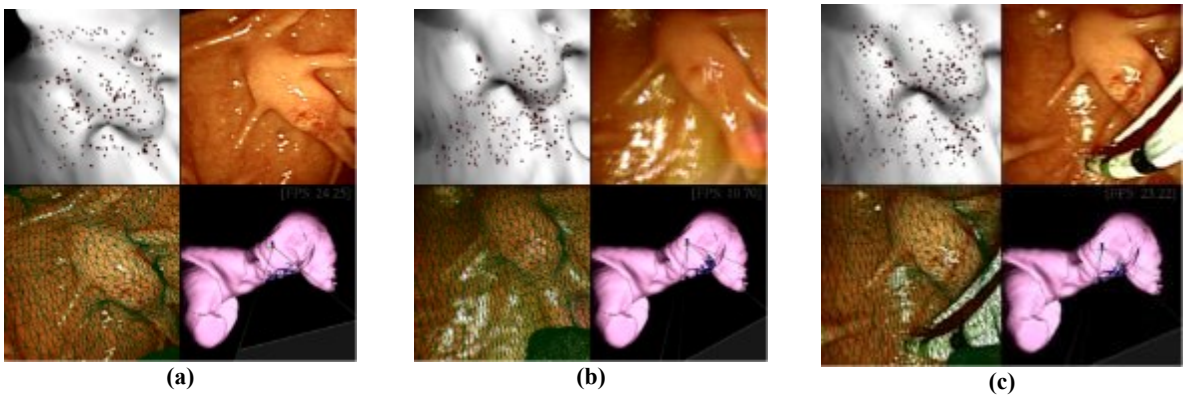
The visual tracking result in Figure 4 shows that even when there is some occlusion as some water was sprayed into duodenum or the catheter is inserted, the method still shows a robust and accurate result. The average computation time for camera pose estimation for one frame was about 37ms equivalent to an average frame rate of about 27 frames per second which means it is applicable to run real-time. Figure 5 show different tracking results to illustrate the robustness and accuracy of the method. We found that the estimated trajectory using modified weight function had smaller error compared with result by using original weight function.

IV. Conclusion and Future Work

Prior works have focused on AR surgeries in which fiducial markers are used to enable tracking procedure. However, these studies have not focused on surgeries where it seems impossible to put any markers on patients' body such as ERCP. In this study we have indicated the difficulties of ERCP and requirements of an AR system for ERCP procedure and developed a markerless camera tracking method utilizing FREAK binary keypoint for real-time computation for AR to ERCP. The problem caused by a large number of mismatches due to highly specular surface of the duodenum is handled by a new adaptive weight function for pose estimation. However, some limitations are worth noting. The method requires a manual initialization which should be addressed in the future research.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 20110029043).



(a) (b) (c)

Figure 4. Pose estimation results. Top left: virtual image, keypoints are marked as red dots; Top-right: current endoscopic image. Bottom left: Current endoscopic image fused with virtual model. Bottom-right: Third person view of virtual camera and duodenum model. Frame (b) is when water is sprayed on the duodenum. Frame (c) is when there is some occlusion because of the catheter.

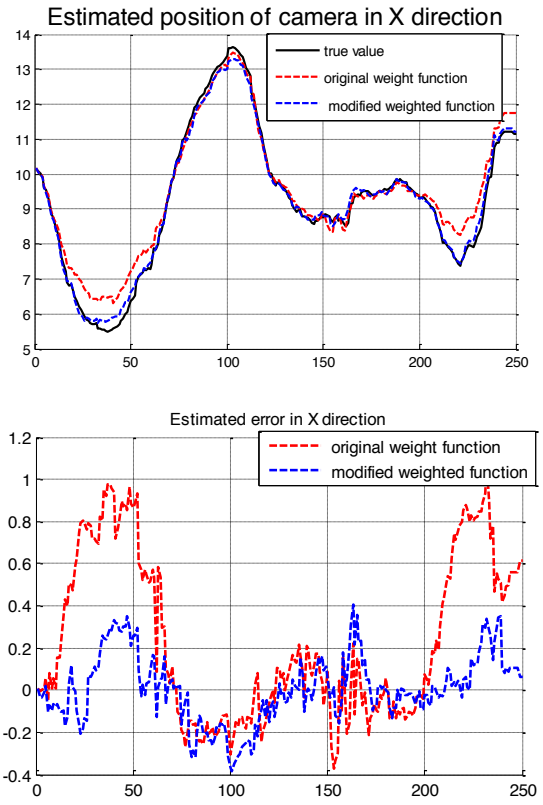


Figure 5. Estimated position of camera in x direction (mm) in 250 frames.

REFERENCES

[1] A.D.A.M,"ERCP," *http://www.nlm.nih.gov/medlineplus/ency/imagepages/19213.htm*, 2009.

[2] L. Lemieux and R. Jagoe, "Effect of fiducial marker localiza-tion on stereotactic target coordinate calculation in CT slices and radiographs," *Phys. Med. Biol,* vol. 39, pp. 1915–1928, 1994.

[3] P. F. Hemler, S. Napel, T. S. Sumanaweera, R. Pichumani, P. A. van den Elsen, D. Martin, *et al.*, "Registration error quantification of a surface-based multimodality image fusion system," *Med Phys,* vol. 22, pp. 1049-56, Jul 1995.

[4] A. C. D. Vandermeulen, J. Michiels, H. Bosmans, P. Suetens, G. Marchal, G. Timmens, P. van den Elsen, M. Viergever, H.-H. Ehricke, D. Hentschel, R. Graumann, "Multi-modality image registration within COVIRA," in *Medical Imaging - Analysis of Multimodality 2D/3D Images*. vol. 19, M. H. K. L. Beolchi, Ed., ed, 1995.

[5] X. W. Wang, Q. Zhang, Q. O. Han, R. G. Yang, M. Carswell, B. Seales, *et al.*, "Endoscopic Video Texture Mapping on Pre-Built 3-D Anatomical Objects Without Camera Tracking," *Ieee Transactions on Medical Imaging,* vol. 29, pp. 1213-1223, Jun 2010.

[6] Y. Yim, M. Wakid, C. Kirmizibayrak, S. Bielamowicz, and J. K. Hahn, "Registration of 3D CT Data to 2D Endoscopic Image using a Gradient Mutual Information based Viewpoint Matching for Image-Guided Medialization Laryngoplasty," *JCSE,* vol. 4, pp. 368-387, 2010.

[7] J. Jomier, E. Bullitt, M. Van Horn, C. Pathak, and S. R. Aylward, "3D/2D model-to-image registration applied to TIPS surgery," *Med Image Comput Comput Assist Interv,* vol. 9, pp. 662-9, 01 2006.

[8] Y. Park, V. Lepetit, and W. Woo, "Extended Keyframe Detection with Stable Tracking for Multiple 3D Object Tracking," *Ieee Transactions on Visualization and Computer Graphics,* vol. 17, pp. 1728-1735, Nov 2011.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding,* vol. 110, pp. 346-359, Jun 2008.

[10] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast Retina Keypoint," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, Providence, USA, 2012.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, Nov 2004.

[12] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," 1981, pp. 674-679.

[13] Z. Y. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and Vision Computing,* vol. 15, pp. 59-76, Jan 1997.

[14] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Second ed.: Cambridge University Press, ISBN: 0521540518, 2004.