"A Window on Tissue" - Using Facial Orientation to Control Endoscopic Views of Tissue Depth

Juan P. Wachs, *Member IEEE,* Kausheek Vujjeni, Eric T. Matson, *Member IEEE,* Stephen Adams

***Abstract*— An endoscope is an invaluable tool to interpret conditions within a body. Flexible endoscopes are controlled by a set of rotational knobs requiring a doctor's hands to guide and locate the view. This research uses a combination of a camera, facial recognition techniques and software to create a hands-free gesture recognition application for use by a physician to probe the internals of a human body. The physician will utilize the head movements to move the endoscopic camera freeing their hands to perform a procedure or other functions.**

I. Introduction

Surgeons can use the endoscope to traverse internal cavities of the patient's body; simply manipulating the controls and instruments passed through a channel in the endoscope allows them to study certain regions and extract samples.

Two problems exist with the current manual control of flexible endoscope: the control is (a) not intuitive, and (b) requires the use of both hands. For the pan and tilt commands, the surgeon needs to rotate two knobs in the endoscope handle. This manipulation provides a poor mapping of the relationships between the controls and their effects. Knob rotation may match population stereotypical behavior for rotation in the pan axis, but not on the tilt axis. While the rotation in the endoscope's knobs are in the same direction, the pan and tilt angles in the camera are orthogonal. The interface can be improved by making the direction of control movement correspond to the desired camera movement.

To successfully use the endoscope, the surgeon must use one hand to hold the handle and operate the suction, air, and water buttons. The other hand is used to rotate the controls knobs on the side of the handles. This control layout design constrains the surgeon's hands to camera navigation and visualization, and additional activities requiring manipulations are not possible. Laser fibers for laser surgery, biopsy cutting baskets, and various other instruments and catheters can be inserted through the biopsy channel of the endoscope and utilized for many procedures. This requires an additional set of hands.

In this paper we propose the use of head movements to control an endoscope. This allows robust navigation of the endoscope's camera in 3 degrees of freedom (DOF) while releasing both hands for additional surgical instrumentation manipulation. The initial prototype is validated in a virtual environment.

II. Related work

Recent advances in robotic technologies and human-computer interfaces allow sophisticated control and recognition applications to offer alternatives to the standard, human-controlled endoscope. Arguably the most popular system is the Da Vinci surgical system [1]. This system is the first operative surgical robotic commercial system capable of performing advanced surgical techniques such as suturing, cutting and clamping. The main obstacles with Da-Vinci are the large learning times (8-10 patients),the high price (approximately $1 million), and it is designed for rigid endoscopes only.

One possible solution is the use of *gestures*. The main advantages of gestures are that they are intuitive, fast and highly expressive. Gesture includes facial expressions, hand and body movements and postures. We will focus on facial expressions (including gaze patterns) and the 3D orientation of the face, since the goal is to allow the surgeon to use his hands to manipulate additional instruments.

Recently, mouth expressions were used for automatic laparoscope control of a 3DOF robot [2]. A real-time hands-free gaze control was implemented using an eye tracking system [3]. In [4], a system for freehand manipulation of an exoscope and an endoscope through a head-mounted unit was validated in a virtual environment. *Face-Mouse* is a touch-less accurate rigid laparoscope control [5]; a robotic laparoscope positioning system for solo surgery based on a real-time, face-tracking technique has also been developed. They used only 3 DOF control, assuming fronto-parallel and distance-constant interaction. The same approach was adopted in Freehand and EndoAssist (Prosurgics) [6], where the tracking was achieved through wearing markers in the surgeon's cap (encumbered interfaces).

Manuscript received April 23, 2010. This project was funded with support from the Indiana Clinical and Translational Sciences Institute Grant # RR025761 from the National Institutes of Health, National Center for Research Resources, Clinical and Translational Sciences Award.

J. P. Wachs is with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (phone: 765-496-7380; fax: 765-496-1299; e-mail: jpwachs@ purdue.edu).

K. Vujjeni is with Electrical Engineering Department, Purdue University, West Lafayette, IN 47907 USA (kvujjini@purdue.edu).

E. T. Matson is with the College of Technology, Purdue University, West Lafayette, IN 47907 USA (ematson@purdue.edu).

E. Adams is with the Purdue Large Animal Veterinary School, Purdue University, West Lafayette, IN 47907 USA (adamss@purdue.edu).

Our system is similar to FaceMouse, Freehand and EndoAssist in the sense that the surgical instrument is controlled only using touch-less head movements, however, it controls a flexible endoscope, instead of a laparoscope, without attaching active markers to the head; thus reducing possible infections caused by non-sterile devices.

III. METHODOLOGY

In this section we outline the approach to tracking the surgeon's head displacement frame by frame and its projection in a 3D model. The algorithm described here is based on the implementation of Baggio and Solyga [7], and it is a general approach for 3D object tracking.

A. *Calibration and Initialization*

The intrinsic parameters of an RGB camera (focal length and the geometric distortion introduced by the lenses) used to track the head movements are obtained using a calibration grid before the system is used for the first time.

First, the system processes every image capture from the camera searching for the user's head. To determine if the user is standing in front of the camera, the Viola and Jones detector [8] is applied to every frame captured by the camera. Once the user's head is detected continuously through a number of frames, interesting points in the user's face are identified for tracking. These points are found using strong corners in the image using a corner detector [9]. Let q_t denote the 2D points' position in the image plane at time *t* and let the corresponding 3D position of those points be Q_t.

$$q_t = \{q_t^0 ... q_t^n\}$$
$$Q_t = \{Q_t^0 ... Q_t^n\} \quad (1)$$
$$q_t^i = \{x_t^i, y_t^i\} \text{ and } Q_t^i = \{x''_t, y''_t, z''_t\}$$

The points in the image plane and in the 3D model are related by a 3 x 4 transformation matrix *M=[R| T]*, where *R* is the orientation of the object (a 3d rotation matrix) and *T* is its position (a 3d translation vector) with respect to the camera. Let *A* denote the matrix including the intrinsic parameters. Once these parameters are determined the tracking system can be used. Then, the points are related thus:

$$q_t^i = A \times [R_t \mid T_t] Q_t^i \quad (2)$$

The problem is that the corner detector gives only the points in the image view q_t and the remaining information is unavailable. To solve this problem, we assume the following:

a. The face is in frontoparallel position with respect to the camera
b. A 3D model of the face (not the head) can be approximated as a vertical cylindrical model.

Given an initial image (reference template) of the head with the face facing the camera, a cylindrical head model is created as the corresponding head pose. This model will be used through the entire tracking process. Since the face is in front of the camera, the 3D coordinates of the head model are: *x'=x, y'=y* and *z'* is found using the circle equation, see Fig. 1.

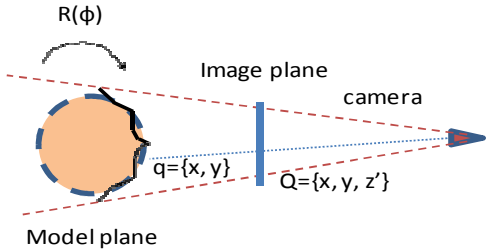

Fig. 1. The cross section of a head as viewed from the top. The dashed line is the approximated model with the solid line is a piecewise representation. In front of it, the image plane with a selected point on the image and its corresponding point on the cylindrical model.

To find the rotation and translation matrices we use the POSIT algorithm [10] ("POS algorithm with iteration") which assumes that the exact dimensions of the 3D object are known. To compute the pose *M*, we must provide the 2D locations and corresponding 3D locations of at least four non-coplanar points on the surface of that object (the face). Since Q_t and q_t were already found using the cylindrical model, the pose can be computed using the POSIT algorithm.

B. *Tracking*

According to the surgeon's head movement, the pose matrix will continuously change as long as the tracking quality is not too degraded by noise and cumulative errors. Computing *M* only requires us to keep track of the 2D locations, since the 3D model remains unchanged. The point correspondences between the 2D and 3D views must be correct, otherwise the pose will not be accurate. To track the interesting points in the image, we use a popular sparse tracking technique called Lucas-Kanade (LK) optical flow [11]. We use an implementation that is based on image pyramids, which allows the tracking of head motions more rapid than those the standard LK tracker can detect. A subset of points in q_t and q_{t-1} are matched together so the correlation between intensity patches around the points is maximized.

Naturally some points will be discarded in the process since no match was found in the previous frame, and some of the matches will be incorrect due to noise, occlusions or fast movements. To overcome this problem, a robust estimator iterative algorithm, RANSAC is used to discard outliers [12]. The input to RANSAC is a set of observations (q_t and Q_1 points), a parameterized model that is fitted to the observations (Eq. 3), and additional calibration parameters.

$$q_{t-1}^i = A \times [R_{t-1} \mid T_{t-1}] Q_1^i \quad (3)$$

This method for tracking is simple, and yet can sustain an accurate tracking, as long as the pose does not change too much (up to 60 degrees in any orientation), and a set of points between the image and the model are matched with

high accuracy. Only four points matched between frames are enough to calculate the pose using POSIT.

C. Endoscopy Control

The recognition module recognizes the surgeon's face movements based on the differential increments in the position and pose of the 3D head model, and thus controls the translational and rotational position of the endoscope camera. We limit translation to the z axis (forward & backward corresponding to push & pull of the endoscope) since movement to the sides is not possible. The orientation of the endoscope's camera is controlled by the pitch and yaw angles (which usually are mapped with two knobs on the endoscope's handle). Roll is discarded since it does not have an analog function in standard endoscopes, see Fig. 2.

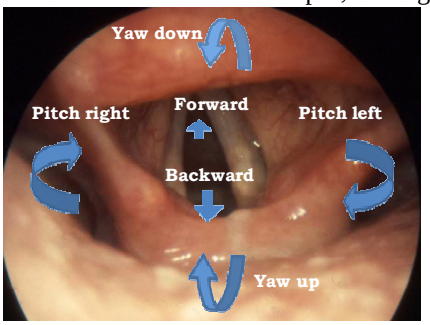

Fig. 2. Position and orientation commands used to control the endoscope's camera; 6 actions are required for a standard endoscopic procedure.

Actions are sent to the endoscope control engine when any movement (or combination of movements) is large enough to pass a threshold. Since the movements are represented by angles and distances, for each variable a different threshold is selected. The camera's movement is continuous and linear as long as the head's movement is above the specific threshold. To stop the movement, the user should move back to the original position, so the angles and distances are below the threshold. The function w is expressed with eq. (4):

$$For\ i=1,2,3: w_i = \begin{cases} \Delta k_1\, sign(a_i) & if\ |a_i| > \tau_i \\ 0 & if\ 0 \le |a_i| \le \tau_i \\ \Delta k_2 sign(b_i) & if\ |b_i| > \eta_i \\ 0 & if\ 0 \le |b_i| \le \eta_i \end{cases} \quad (4)$$

Where, a_i is the rotation in each of the three Euler angles ($1 \le i \le 3$) representing pitch, roll and yaw; b_i is the translation in each of the three axes ($1 \le i \le 3$); and w_i is the amount of movement sent to the endoscope control in the same orientation as a_i or the same direction as b_i; Δk_1 and Δk_2 are constant indicating the increments added to the orientation and direction, respectively; τ_i and η_i are the upper and lower threshold, respectively.

Using the system of equations in (4) allow us to control the endoscope movements using small increments in each direction, regardless the absolute angle or position of the head. The only important fact is whether the face's orientation or position is over the specified threshold, and as long as this inequality is true, the control's angle and movement is incremented with a fixed rate, see Fig. 3.

To adapt the system to a robot control, actuators or just a virtual reality model, only the increments Δk_1 and Δk_2 need to be re-calibrated according to the type of controlled mechanism. To demonstrate the feasibility of the system, a tasks in a virtual reality environment will be used in the current paper.

IV. EXPERIMENTS

In order to compare the robustness and precision of our system compared to traditional interfaces, an experiment was designed; an endoscopic surgery is simulated using a 3D model of the larynx. The task used in this experiment simulated the process of a larynx biopsy, which is a common surgical procedure to determine the existence of laryngeal cancer. The task assigned to the users of the system is to reach the sphere (suspicious tissue) through navigation.

The user is instructed to reach the sphere using two control modalities: head movements and the computer keyboard. The keyboard is analogous to the endoscope knobs in the sense that the camera movements do not follow a cognitive model stereotype.

We conducted with 4 users running 10 trials each, for both the keyboard control and the facial control, where from each trial we measured time completion tasks and we recorded the sequence of commands evoked to complete the task. Each user received a short training session consisting of three trials for the head and two trials for the keyboard.

Fig. 4 shows the time comparison for a person using the keyboard versus a person using the vision system, in seconds. While each shows the vision system takes longer, subjects A and B are beginners and users C and D are intermediate users, who completed a previous training

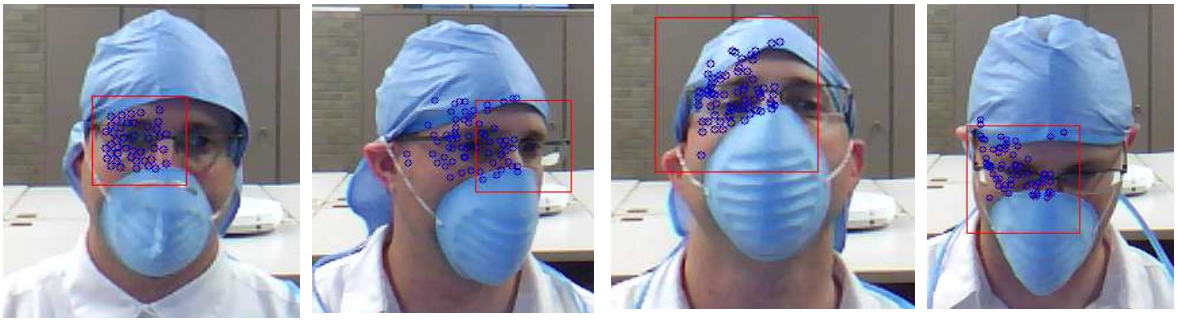

Fig. 3. Recovering the position and orientation of the surgeon's face through landmarks tracking on the face. The landmarks are placed initially in the right eye region, and continuously tracked through the endoscope operation.

experience. While the time for the keyboard usage is relatively similar for all users, there is a significant difference between beginning and intermediate users for the vision system.

Based on the time completion of the tasks, the learning curve was calculated shown in Fig. 5. The learning rate (LR) is $Y_n = Y_1 n^{-b}$. Where Y_n is the predicted completion time, in seconds, for the n^{th} trial, and Y_1 is the time for the first trial, and b is: log r/log 2, where r is the learning rate. The learning rate for the keyboard based system and the head movements based systems were 82.95% and 78.63%, respectively. Since lower learning rates mean faster learning we conclude that the head movement control based system is easier to learn compared to the manual control system.

V. CONCLUSION

This research uses a combination of a camera, facial recognition techniques and software to create a hands-free gesture recognition application utilizing a physician's head movements to move the endoscopic camera. While the use of this system will take typically more time than a hand-based endoscope, the freeing of the hands outweighs the slight increase in time.

Comparing learning times, the head movement control based system resulted in faster learning times compared to the manual control for all users. Further, this system has several advantages. The physician never has to look away from the patient's anatomy, as the head movement required to position the camera is minimal.

Future work includes involving a larger group of users, including surgeons from the Purdue School for Veterinary Medicine, extending usability tests, and improving the robustness of the vision recognition algorithm. Another feature to be added is a mechanism to allow the user to "disconnect" from the tracking system, so he can move his head without always moving the endoscope.

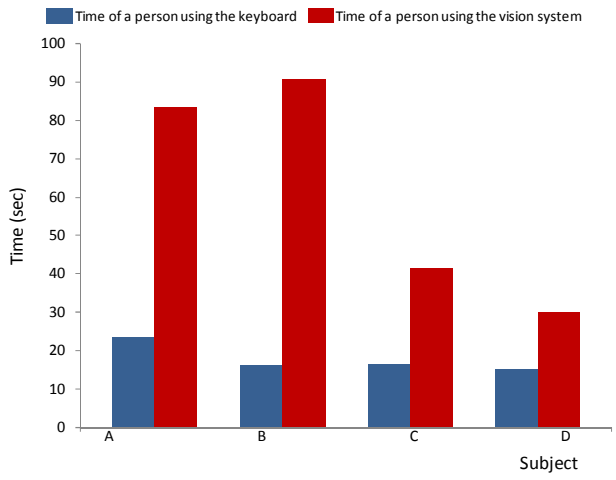

Fig. 4. Time comparison between keyboard and facial control

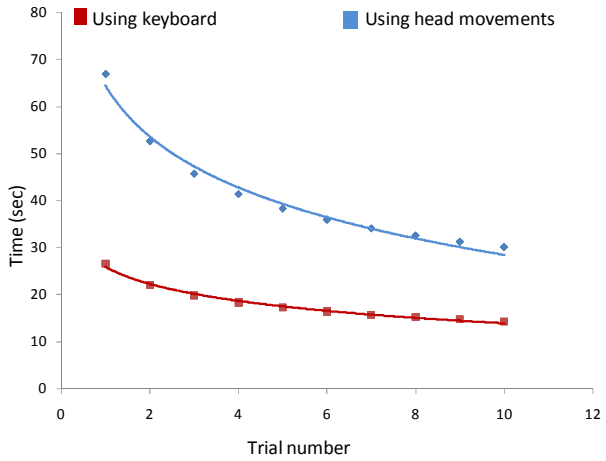

Fig. 5. Learning curves for the manual control and head movement based control systems.

REFERENCES

[1] http://www.davincisurgery.com/

[2] Gomez, J.-B.; Ceballos, A.; Prieto, F.; Redarce, T.; Mouth gesture and voice command based robot command interface. IEEE International Conference on Robotics and Automation, 2009. ICRA '09. Pp. 333 – 338

[3] Cheng-Li (Geoffrey) Tien. 2009. Building Interactive Eyegaze Menus for Surgery. Master Thesis. Simon Fraser University, Canada.

[4] Serefoglou S, Laurer W., Perneczky A, Lutze T, Radermacher K. Combined endo- and exoscopic semi-robotic manipulator system for image guided operations. Med Image Comput Comput Assist Interv. 2006;9(Pt 1):511-8.

[5] Nishikawa, A., Hosoi, T., Koara, K., Negoro, D., Hikita, A., Asano, S., Kakutani, H., Miyazaki, F., Sekimoto, M., Yasui, M., Miyake, Y., Takiguchi, S., Monden, M.: Face mouse: A novel human-machine interface for controlling the position of a laparoscope. IEEE Trans. on Robotics and Automation 19(5), 825–841 (2003)

[6] Prosurgics (2009). Available at http://www.prosurgics.com.

[7] Baggio, D. Solyga, P. Enhanced human computer interface through webcam image processing library. Google Summer of Code. http://code.google.com/u/danielbaggio/

[8] P. Viola and M. Jones. Robust Real-time Object Detection. International Journal of Computer Vision, 2001.

[9] J. Shi and C. Tomasi, "Good features to track," 9th IEEE Conference on Computer Vision and Pattern Recognition, June 1994.

[10] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," Proceedings of the European Conference on Computer Vision (pp. 335–343), 1992.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proceedings of the 1981 DARPA Imaging Understanding Workshop (pp. 121–130), 1981.

[12] M. A. Fischler and R. C. Bolles, "Random sample concensus: A paradigm for model fi tting with applications to image analysis and automated cartography," Communications of the Association for Computing Machinery 24 (1981): 381–395.