Episode Classification for the Analysis of Tissue/Instrument Interaction with Multiple Visual Cues

Benny P.L. Lo, Ara Darzi, and Guang-Zhong Yang

Royal Society/Wolfson Medical Image Computing Laboratory
Imperial College London,
London, United Kingdom
{benny.lo,a.darzi,g.z.yang}@imperial.ac.uk

Abstract. The assessment of surgical skills for Minimally Invasive Surgery (MIS) has traditionally been conducted with visual observation and objective scoring. This paper presents a practical framework for the detection of instrument/tissue interaction from MIS video sequences by incorporating multiple visual cues. The proposed technique investigates the characteristics of four major events involved in MIS procedures including idle, retraction, cauterisation and suturing. Constant instrument tracking is maintained and multiple visual cues related to shape, deformation, changes in light reflection and other low level images featured are combined in a Bayesian framework to achieve an overall frame-by-frame classification accuracy of 77% and episode classification accuracy of 85%.

1 Introduction

Endoscopy, including bronchoscopy and laparoscopy, is the most common procedure in Minimal Invasive Surgery (MIS). It is carried out through natural body openings or small artificial incisions. It achieves its clinical goals with minimal inconvenience to patients, reduced patient trauma, shortened hospitalisation, and improved diagnostic accuracy and therapeutic outcome. Despite the major advantages these techniques attract, they require a high degree of manual dexterity from the operator as the complexity of the instrument controls, restricted vision and mobility, difficult hand-eye co-ordination, and the lack of tactile perception are major obstacles. To alleviate these problems, MIS specific training is indispensable for the safe practice of these procedures. Thus far, the development of Virtual Reality (VR) simulators has been a major focus of research in surgical technology as they allow for comprehensive training of MIS specific surgical tasks [1,2]. Predominantly, these simulators have attempted to accurately model the mechanical properties of the tissue and its interaction with instruments [3,4]. Apart from providing a general framework for surgical training, VR simulators also allow a quantitative assessment of basic surgical skills as the motion of the instruments and tissue, as well as their interactions, are known [5,6]. Such information, however, is not available during real procedures. Despite the fact that simulators have advanced significantly in recent years, they are still not realistic enough to be taken as the only source of training, nor for the acquisition and assessment of certain advanced surgical skills. To facilitate objective assessment of surgical

R.E. Ellis and T.M. Peters (Eds.): MICCAI 2003, LNCS 2878, pp. 230–237, 2003.

skills in real procedures, much research has been focused on the design and development of special MIS tools equipped with force and torque sensors for measuring the kinematics of the instruments [7,8]. By analyzing the force and torque applied to the tools, mathematical models can be applied to classify surgical movements during the operation such that quantitative information can be derived [7,8]. These systems, however, do not consider instrument/tissue interaction and it is necessary to cross validate with *in situ* captured video sequences to achieve a comprehensive visual assessment of the procedure.

One prerequisite of quantitative skills assessment involving tissue/instrument interaction is the segmentation of surgical episodes and the identification of tissue deformation in response to instrument movements. In this paper, we propose a novel approach to MIS video episode segmentation and motion analysis based on multiple visual cues. The proposed technique investigates the characteristics of four major events in MIS: idle, retraction, cauterisation and suturing, whist maintaining constant instrument tracking. Multiple visual cues related to shape, deformation, changes in light reflection and other low level images featured are combined in a Bayesian framework to achieve a high classification accuracy.

2 Method

Prior to the classification of tissue/instrument interactions, colour segmentation is first applied to the endoscopic video to segment MIS tools from the background tissue. This is then followed by tracking of the MIS tools to derive their associated motion characteristics. To measure instrument induced tissue deformation, optical flow and shape-from-shading based techniques are used. In addition, four other low-level visual cues are incorporated to augment the classification accuracy within a Bayesian framework.

2.1 Tissue/Instrument Segmentation

For tissue/instyrument segmentation, we exploited their intrinsic colour difference manifested within the video sequence. A Bayesian classifier, which models the colour distribution as unimodal Gaussian distributions in the hue-saturation space, was used. Specifically, the likelihood of an pixel $\boldsymbol{x}$, where $\boldsymbol{x}$=(*hue, saturation*), belonging to class w_i [w_0=*instrument and* w_1=*tissue*] is defined as follows:

$$p(x \mid w_i) = \frac{1}{2\pi\sqrt{|C_i|}} \exp\left[-\frac{1}{2}\left(x - \bar{x}_i\right)^T C_i^{-1} (x - \bar{x}_i) \right] \tag{1}$$

$$Tools(x,t) = \begin{cases} 1 & p(x \mid w_0) > p(x \mid w_1) \\ 0 & otherwise \end{cases} \qquad Tissue(x,y) = \begin{cases} 1 & p(x \mid w_0) < p(x \mid w_1) \\ 0 & otherwise \end{cases}$$

where C_i represents the covariance matrix and $\bar{x}_i$ the mean vector. To train the classifier, two manually segmented images were used to derive the distribution of the colour components.

2.2 Instrument Tracking

In order to analyse the temporal behaviour of the instrument tip, a polygonal model was used to track the movement of the MIS tools. To ensure temporal consistency and minimising tracking errors, the CONDENSATION algorithm was applied to predict the locations of the instruments [9,10]. The conditional probability $p(z_t|x_t=s_t)$ of the observation vector z_t of the instrument model at time t, given that the feature vector x_t is equal to the state vector s_t, is defined as follows:

$$p(z_t \mid x_t = s_t) = \frac{1}{A}\sum \delta\big(Instrument(x, y), polygon(x, y, s_t)\big) \tag{2}$$

where s_t denotes the state vector, which consists of the parameters of the polygonal instrument model, and $polygon(x, y, s_t) = 1$, if (x, y) lies within the model defined by s_t and is 0 otherwise. In the above equation, A is the area of the polygon. By applying the CONDENSATION algorithm, instruments can be tracked in the highly cluttered endoscopic scenes.

2.3 Tissue Deformation

As no information regarding the deformability and other mechanical properties of the tissue is available in real MIS operations, tissue deformation induced by instrument interaction can only be inferred through its appearance in video sequences. Thus far, most MIS operations use a single camera setup. We used optical flow [11] as a means of estimating tissue deformation. The updating scheme requires a smoothness constraint of adjacent motion vectors, which is guaranteed by the application of tissue/instrument segmentation in previous steps. The following updating equations were used in this study:

$$u = \bar{u} - \frac{\frac{\partial f}{\partial x}\left(\frac{\partial f}{\partial x}\bar{u} + \frac{\partial f}{\partial y}\bar{v} + \frac{\partial f}{\partial t}\right)}{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + 3\lambda}, \qquad v = \bar{v} - \frac{\frac{\partial f}{\partial y}\left(\frac{\partial f}{\partial x}\bar{u} + \frac{\partial f}{\partial y}\bar{v} + \frac{\partial f}{\partial t}\right)}{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + 3\lambda} \tag{3}$$

where $\bar{u}$ and $\bar{v}$ are the means of motion vector u and v respectively, and λ is the Lagrange multiplier. For analysing instrument induced tissue deformation, it is important to separate local deformation from global tissue movements caused by respiration, digestion and pulsation of the arteries. To this end, the variance of the flow field was chosen as the main indicator of localised tissue deformation in response to instrument interaction. It is well known that optical flow based technique works well when there is sharp change in tissue shape or in the presence of rich surface texture. For gradual deformation of the smooth soft tissue, we used a shape-from-shading technique that exploited the unique geometrical constraints between the endoscopic camera and the light source. In 3D space, they are located immediately next to each other and always move in synchrony. By assuming the surface to be Lambertian with constant albedo, the depth of the tissue in relation to the camera at each time frame can be estimated with Taylor series expansion, as proposed by Tsai and Shah [11]:

$$Z_t^n(x,y) = Z_t^{n-1}(x,y) + \frac{\dfrac{-f(Z_t^{n-1}(x,y))}{df(Z_t^{n-1}(x,y))}}{dZ(x,y)} \quad where \quad Z_0^0(x,y) = 0 \tag{4}$$

$$\frac{df}{dZ_t^n} = \left(\frac{(p+q)(pp_s + qq_s + 1)}{\sqrt{\left(p^2+q^2+1\right)^3}\sqrt{p_s^2+q_s^2+1}} - \frac{p_s+q_s}{\sqrt{p^2+q^2+1}\sqrt{p_s^2+q_s^2+1}} \right)$$

$$p = \frac{\partial Z}{\partial x} \quad q = \frac{\partial Z}{\partial y} \quad p_s = \frac{\cos\tau\sin\sigma}{\cos\sigma} \quad q_s = \frac{\sin\tau\sin\sigma}{\cos\sigma}$$

In Equation (4), $Z_t^n(x,y)$ represents the depth value of pixel (x, y) at time t after n iteration, and τ is the tilt of the illuminant and σ is the slant of the illuminant. Consequently, the deformation of the tissue can be estimated by integrating the depth change over image frames.

2.4 Changes in Specular Highlights

Although the movement of the instruments and tissue provide most of the information about their interaction, the complexity of the tissue morphology often leads to erroneous classification. To enhance the accuracy of the classification results, other visual cues were also used. During the process of cauterisation, the instruments barely move, as surgeons often take extra caution while cauterising tissues and blood vessels. This can lead to significant difficulty in using movement as a cue for identifying tissue/instrument interaction. It has been found that the intensity level of the tissue being cauterised often increases significantly, a measurement of the changes or movements of the specular highlight is used to aid the identification of cauterisation, *i.e.*,

$$Specular(t) = \frac{1}{N_I}\Sigma\left|S(x,y,t) - S(x,y,t-1)\right| \quad where \quad S(x,y,t) = \begin{cases} 1 & I(x,y) > 200 \\ 0 & otherwise \end{cases} \tag{5}$$

where N_I is the size of the image, and $I(x, y)$ represents the intensity level of a pixel at (x, y).

2.5 The Presence of Suture and Suturing Movements

During a suturing procedure, the tools usually do not have direct contact with the tissue. As such, the detection of suturing is difficult without further visual cues. Unlike other events, suturing requires the use of suture, which has a very distinctive colour. As such, colour segmentation followed by line aggregation was used. As sutures are long and thin, the ratio between the length of line segments and their areas is used to measure the likelihood of the presence of sutures, *i.e.*,

$$P_1(Suture \mid t) = \frac{1}{N_s}\sum_i \frac{length(line_i)}{area(line_i)} \tag{6}$$

where $line_i$ is the i^{th} line segments found in the scene, and N_s is the total number of line segments. In addition, another likelihood measure for the presence of the suture was defined as projecting the line segments onto an image buffer and then comparing this with the colour filtered image by using the following formula,

$$P_2(Suture \mid t) = \frac{1}{N_p}\sum_y \sum_x \delta(projected(x, y), filtered(x, y)) \tag{7}$$

In the above equation, N_p is the image size, *projected(x, y)* is the image buffer where the lines are projected onto, and *filtered(x, y)* is the colour segmented image. Furthermore, information concerning suture movements was also used, where

$$SutureMovement(t) = \frac{1}{N_s}\sum_i moved\left(line_i(t), line_i(t-1)\right) \quad where \tag{8}$$

$$moved\left(line_i(t), line_i(t-1)\right) = \begin{cases} 1 & p\left(line_i(t) = line_i(t-1)\right) < 0.8 \\ 0 & otherwise \end{cases}$$

$$p\left(line_i(t) = line_i(t-1)\right) = \frac{1}{N_l}\sum_j^{N_l} \left| line_{ij}(t) - line_{ij}(t-1) \right|$$

N_l represents the number of line segments previously identified at *t-1*.

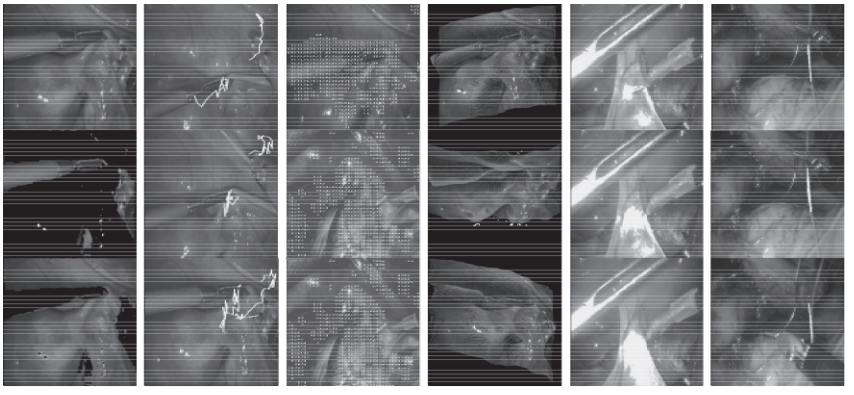
2.6 Classification

A naïve Bayesian network has been employed to fuse the different visual cues and classify different event episodes. From the available sequences, the four major types of events were used. It is worth mentioning here that the idle state not only includes situations when the instruments are stationary, but also represents instances when instruments undergo 3D motion but with no interaction with the tissue. In accordance with the four events defined, the Bayesian network was designed to have one root node with 4 states, each representing one of the event type, and 8 children, where each child node represents a visual cue and each visual cue is quantised into 5 states. The network was constructed by learning its parameters from a training data set which consists of 121 samples.

3 Results

The proposed technique has been applied to five different sequences, in which 3 of them are recorded from robotic surgeries and the others are recorded from laparoscopic operations. Fig 1 illustrates representative image frames after applying different feature extractors proposed in this paper. Figs 1(a) and 1(b) demonstrate the result of tissue/instrument segmentation and subsequent motion tracking for a laparoscopic video sequence, with trajectory highlighting the movement history of the centroids of

the instrument model. Fig 1(c) illustrates the optical flow fields estimated from the video images whereas Fig 1(d) provides sample 3D views of the reconstructed 3D surface of the soft tissue by using the laparoscopic shape-from-shading algorithm. Figs 1(e) and 1(f) are derived from the detection of specular highlights for identifying cauterisation and suture line segments, respectively.



(a) (b) (c) (d) (e) (f)

Fig. 1. (a) Laparoscopic image before colour segmentation (top) and after instrument (middle) and tissue (bottom) segmentation. (b) Results after instrument tracking where the paths taken by the left and right instruments are highlighted with white lines. (c)-(d) The derived optical flow fields and 3D surface reconstructed from the laparoscopic shape-from-shading algorithm. (e)-(f) Changes in specular highlights (white) during cauterisation and line segment (white) detection for the identification of sutures.

Table 1. Classification results for the testing video sequences consisting of 1762 video frames in total which involve the four different instrument maneuvours.

	Events	Idle	Retraction	Cauterisation	Suturing
Frames	*Accuracy*	68.9%	88.3%	57.5%	87.7%
	No of frames	761	495	141	365
	Overall				**77.3%**
Episodes	*Accuracy*	71.4%	100%	60%	100%
	No of episodes	28	19	5	16
	Overall				85,3%

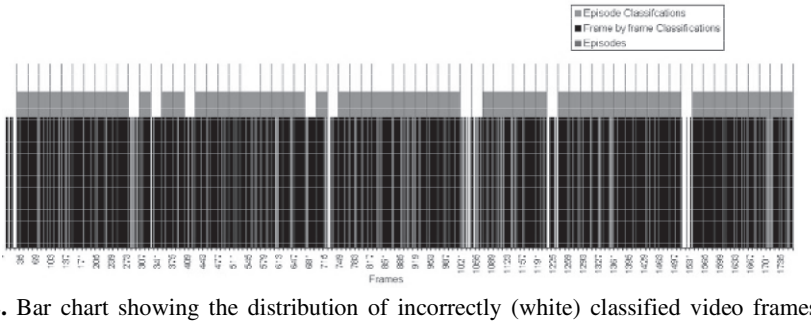


Fig. 2. Bar chart showing the distribution of incorrectly (white) classified video frames and episodes throughout the testing video sequences consisting of 1762 image frames, the blue lines indicate the boundary between episodes.

In order to assess the overall performance of the proposed framework for multiple visual cue integration, 5 video sequences consisting of 1762 video frames in total were used and the accuracy associated with different types of motion are listed in Table 1. In terms of frame-by-frame classification, the accuracy of the proposed technique is about 77%. Since surgical movements are often continuous and normally take at least a few hundred milliseconds to complete, we have also evaluated the accuracy of the proposed technique in segmenting video episodes by incorporating the temporal information. As shown in Table 1, the associated accuracy reaches an overall value of 85.3%, with individual episode accuracy ranging from 60% to 100%. To provide an overview of how the algorithm performs over time, Fig 2 provides a bar chart showing the distribution of the mis-classified frames/episodes throughout the entire video sequence.

4 Discussion and Conclusion

This paper provides a unified framework for integrating different visual cues in video sequence segmentation for MIS procedures. Thus far, limited research has been conducted in applying video sequence processing for MIS procedures, and the majority of research is concerned with enhancing VR simulators rather than studying the dynamics involved in real operations. The development of accurate video segmentation and tissue/instrument tracking has clear advantages in that the system does not involve *ad hoc* tracking hardware which can be problematic in real life operations. The results shown in this paper demonstrate that the analysis of tissue/instrument differentiation in different tissue/instrument interactions can be achieved with a reasonably high accuracy. Nevertheless, the results also indicate a relatively low accuracy of the system in differentiating cauterisation and idle instrument movements (60%-71%). This is mainly caused by the lack of depth perception with monocular Laparoscopic systems. The use of specular highlight alone is not sufficient to differentiate between the two, and further visual cues must be incorporated for the performance of the system to be improved. With the steady improvements in endoscopic CCD/CMOS sensors, binocular systems are becoming increasingly available in routine endoscopic procedures. In this case, stereo 3D reconstruction can be used for improved depth reconstruction and deformation tracking. Although in this paper results concerning instrument motion characteristics of individual trainees are not presented, they can be readily derived from the motion tracked data from the proposed processing framework.

References

1. Muller-Wittig W, Bockholt U, Arcos JLL and Vossl G. Enhanced training environment for minimally invasive surgery. Proceedings of the Tenth IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2001), 269–272, 2001.
2. Brown I, Mayooran Z, Seligman C, Healy DL, Guglielmetti M, Reston M, and Sean Hart. Engineering design of a virtual reality simulator for gynaecological endoscopy. The seventh Australian and New Zealand Intelligent Information Systems Conference, 77–80, 2001.

3. Frank AO, Twombly IA, Barth TJ and Smith JD Finite element methods for real-time haptic feedback of soft-tissue models in virtual reality simulators. Proceedings of the IEEE Virtual Reality 2001, 257–263, 2001.
4. Brown J, Sorkin S, Bruyns C, Latombe JC, Montogomery K and Stephanides M. Real-time simulation of deformable objects: tools and application. Proceedings of the fourteenth conference on Computer Animation 2001, 228–258, 2001.
5. Gallagher AG and Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Surgical Endoscopy, Springer-Verlag York, 16: 1746–1752, 2002.
6. Shah J and Darzi A. Simulation and skills assessment. Proceedings of the International Workshop on Medical Imaging and Augmented Reality 2001, 5–9, 2001.
7. Rosen J, Brown JD, Chang L, Barreca M, Sinanan M and Hannaford B. The BlueDRAGON - a system for measuring the kinematics and dynamics of minimally invasive surgical tools in-vivo. Proceedings of the ICRA '02 IEEE International Conference on Robotics and Automation 2002, 2:1876–1881, 2002.
8. Ang WT, Riviere CN and Khosla PK. An Active Hand-held Instrument for Enhanced Microsurgical Accuracy. Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention 2000, 11–14, Oct 2000.
9. Isard M and Blake A. CONDENSATION - conditional density propagation for visual tracking. International Journal of Computer Vision, 29(1), 5–28, 1998.
10. Black MJ and Jepson AD. Recognizing Temporal Trajectories using the Condensation Algorithm. Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 16–21, 1998.
11. Tsai PS and Shah M. Shape From Shading Using Linear Approximation. Image and Vision Computing Journal, 1994.