Automated Editing of Medical Training Video via Content Analysis

Kevon Andrews[c] Daniel Ring[a] Anil Kokaram[a] Fadel Al Sabah[b] T. Clive Lee[b] Cathy Radix[c]

[a]Electronic Engineering Department, Trinity College, University of Dublin, College Green, Dublin 2, Ireland

[b]Department of Anatomy, Royal College of Surgeons Ireland, St. Stephen's Green, Dublin 2, Ireland

[c] Department of Electronics and Computer Engineering, University of the Wset Indies, St Augustine, Trinidad and Tobago.

ABSTRACT

Physicians in the early part of their training inevitably undertake a course in Anatomy. Unfortunately, the amount of training medical students have with real bodies has decreased. This is further exacerbated with the increasing gap between numbers of medical students and resources available. Medical faculties worldwide are increasingly turning to video training sessions as a complement to practical sessions. This paper presents a number of automated content access and enhancement tools which have been designed to alleviate the difficulty of editing these sessions. The system is being deployed at the Royal College of Surgeons in Ireland.

Keywords: Content based retrieval, Motion Estimation, Image Segmentation, Video Object Tracking and Enhancement, Bayesian Inference

1. INTRODUCTION

The early stages of any course in Medicine includes a substantial module in Human Anatomy. Courses are typically structured to include a number of dissection sessions on human cadavers. However, as the study of medicine grows ever more specialist, the amount of time devoted to the practical aspects of dissection has decreased. The use of video/graphics technology to improve the quality of contact time for students is therefore being explored by a number of institutions. At Stanford, for instance, [`http://www.carykornfeld.com/3dtv.html`] the use of 3D live TV for projecting demonstration dissections allows a larger number of students to view the process than is possible around a crowded dissection table. Using 3D TV also allows the video projected to show the correct spatial relationships between anatomical components, an important aspect for clinical practice. At the Royal College of Surgeons in Ireland (RCSI), demonstration video recorded of a complete human dissection performed by a qualified doctor is used for students to preview their task before coming into the Anatomy Room. It is also a valuable aid to revision and improves the learning ability of the students within the allotted time. Other institutions use image or video aids in this way, but many use synthetic video or still images.[1–3]

Unfortunately, to make the video useful, professional editing of the recorded session is required. This takes time and skill that the Medical lecturers do not necessarily have. In addition, in playback of the material in a lecture theatre or on a home PC, important anatomical landmarks are not well lit or contrasted. This is because the exposed tissue cannot necessarily be arranged for optimal viewing in each situation. There is a need therefore for an assistive tool that facilitates editing of such training sessions by those not skilled in video editing.

This paper presents a number of tools constituting such a system. The idea is to design content analysis algorithms that parse the video at a high level of understanding so as to detect important episodes. The parsed episodes are then processed to highlight important features in the field of view, yielding an augmented expression of the exposed anatomy. The net effect is to create a 10 min summary of an up to 8 hour video recording of a dissection session.

Further author information: (Send correspondence to Daniel Ring)
E-mail: ringdk@tcd.ie

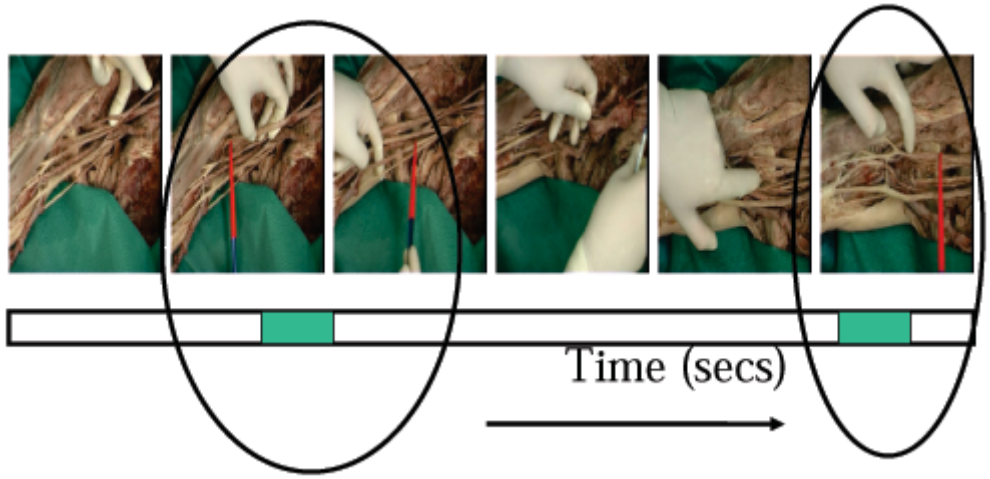

Figure 1. An overview of the content in 5 mins of footage (out of several hours recorded). The highlighted sections on the timeline show relevant content in which an organ or feature is being highlighted. In those images, the pointer is clearly visble and also moves relatively slowly or deliberately. The other pictures contain content in which the Doctor is merely preparing the feature for discussion.

1.1. Overview

During the dissection process, the Doctor performing the operation pauses at important stages and often uses a pointer to indicate the region of interest. This is shown in Figure 1. Locating the pointer or instrument for dissection in the field of view is the most powerful feature for parsing. In practice the pointer is covered with coloured tape which facilitates detection in the field of view. Tracking the pointer then allows the area of interest to be isolated and enhanced/augmented as necessary. Pointer tracking also gives an index of interest within each episode since the Doctor usually pauses the motion while speaking about an important anatomical feature. The next sections outline the algorithms for pointer segmentation, orientation measurement and tracking. This is then used as a feature for parsing and content augmentation. Various different simple augmentation schemes are discussed leading on to interesting representations for summarizing this footage.

2. INSTRUMENT LOCATION

Figure 1 shows a sequence in which the pointer is used by a doctor to delineate some feature, after preparing the cadaver in a useful manner. The images are taken from different episodes of the session. This sequence is typical of the observed video, and examples can be seen at `www.sigmedia.tv/anatomy`. Detecting and segmenting the instrument in the field of view is the first step in content analysis.

2.1. Colour Segmentation

The goal is to estimate a label image, $L()$, in which $L(\mathbf{x}) = 1$ represents a pixel at location $\mathbf{x} = [i, j]$ detected as belonging to the instrument. $L(\mathbf{x}) = 0$ represents a pixel in the background. Colour is used as the key feature for segmentation. The colour of the instrument is modelled as a Gaussian distribution roughly in the red region of the spectrum. Consider that an initial guess for a configuration of $L_0()$ is available. It is then required to estimate the label field at each site, given the state of the sites around. A solution can be generated by manipulating $p(l(\mathbf{x})|L(\mathbf{x}), I, \theta)$. This is the pdf of the unknown label $l(.)$ at site x given the observed image $\mathbf{I}(\mathbf{x}) = [I^h(\mathbf{x}),\ I^s(\mathbf{x})]$ at that site, the parameters of the gaussian colour model θ and the known labels around the current site. Note that just two colour planes are used (Hue I^h and Saturation I^s) in order to give the process robustness to varying lighting conditions. This is a simple approach to colour segmentation. Proceeding in a Bayesian fashion, the required distruibution can be factored as below

$$p(l|L, I, \theta) \propto p(I|\theta, l)p(l|L) \tag{1}$$

To proceed, sensible choices must be made for the likelihood $p(I|...)$ and the prior $p(l|L)$.

2.1.1. The Likelihood

A sample of pixels of the instrument in the view was used to model the colour distribution of the instrument. The HSV colourspace was used. The colour distribution is modelled with a 2D Gaussian, as the luminance component is ignored. This facilitates detection in varying lighting conditions.

$$p(\mathbf{I}(\mathbf{x})|L(\mathbf{x})) \propto \begin{cases} \exp - \left(\frac{(I^h(\mathbf{x})-m_h)^2}{2\sigma_h^2} \frac{(I^s(\mathbf{x})-m_s)^2}{2\sigma_s^2} \right) & \text{for } L(\mathbf{x}) = 1 \text{ i.e. the instrument} \\ \exp - (\alpha) & \text{for } L(\mathbf{x}) = 0 \text{ i.e. the background} \end{cases} \tag{2}$$

The mean and variance of the two colour planes used are defined as m_h, σ_h^2 and m_s, σ_s^2 respectively.

2.1.2. The Prior

Pixels constituting the instrument will cluster together since the instrument itself is convex : i.e. not a random distribution of spatial particles. To encourage the label field to behave in this way implies that it must be smooth. Using a Gibbs energy function for the prior yields the following expression

$$p(L(\mathbf{x})|\mathcal{N}^L(\mathbf{x})) \propto \exp(-\Lambda \sum_{n=1}^{8} \lambda_n (L(\mathbf{x}) \neq L(\mathbf{q}_n)) \tag{3}$$

where $\mathbf{q}_n$ indexes the 8 nearest neighbour labels around the current site, and $\mathcal{N}^L$ is that collection of 8 nearest neighbour labels. Hence the most likely label field is that in which the label at the current site is the same as all the neighbours. λ_n is set to $1/\sqrt{2}$ for diagonal neighbours and 1 for the other 4 neighbours. Λ controls the overall smoothing strength and values of 2 to 5 are appropriate.

2.1.3. Solution

The optimal neighbourhood configuration is chosen by iterating over each site in the image, selecting that label which maximises the local p.d.f. $p(L(\cdot)|\mathbf{I}(\cdot), \mathcal{N}^L(\cdot))$. This is the ICM algorithm as proposed in.[4] Maximising the probability implies minimizing the log likelihood. With this simplification it is possible to propose the process of segmentation as an energy minimisation process that proceeds as below.

1. Calculate the log-likelihood E_L at every pixel site as

$$E_L(\mathbf{x}) = \frac{(I^h(\mathbf{x}) - m_h)^2}{2\sigma_h^2} + \frac{(I^s(\mathbf{x}) - m_s)^2}{2\sigma_s^2} \tag{4}$$

2. Initialise the label field by thresholding the log-likelihood (for instance) over the whole image. In other words

$$L^0(\mathbf{x}) = \begin{cases} 1 & \text{if } E_L(\mathbf{x}) > T \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

3. For every pixel in the image $\mathbf{x}$ calculate the two energies as below

$$E_0(\mathbf{x}) = \alpha + \Lambda \sum_{n=1}^{8} \lambda_n (0 \neq L(\mathbf{q}_n)) \tag{6}$$

$$E_1(\mathbf{x}) = E_L(\mathbf{x}) + \Lambda \sum_{n=1}^{8} \lambda_n (1 \neq L(\mathbf{q}_n)) \tag{7}$$

4. Select the optimal estimate of $L(\cdot)$ at that site as $L(\cdot) = 0$ if $E_0 \leq E_1$ and $L(\cdot) = 1$ otherwise.

5. If the new label configuration is different from the last, repeat the iteration over the whole image starting at step 3.

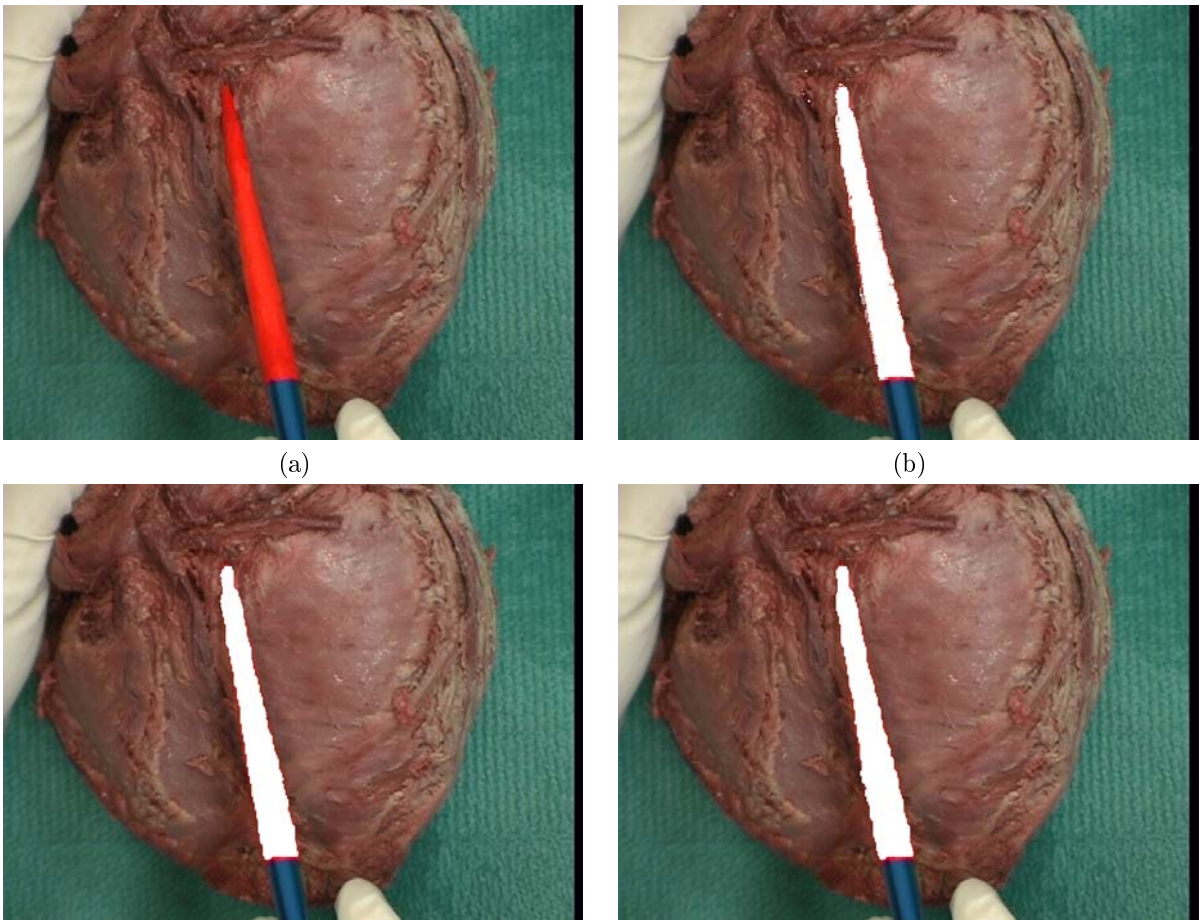

Figure 2. (a) Original image (b) Thresholded Log-Likelihood (c) Segmented image using ICM, after 5 iterations at finest level, with $\alpha = 2.58$, $\Lambda = 5$ (d) Segmented image using multiresolution ICM, after 2,2,1, iterations at the 3 levels (from coarse to fine); $\alpha = 2.58$, $\Lambda = 5$. The red instrument is correctly delineated in all views.

In practice a checkerboard scan is a better site visitation strategy. Note that a value for α can be chosen using a confidence test argument. Considering just the likelihoods, it can be seen that if $E_L < \alpha$, $L(\cdot) = 1$ is chosen, i.e. a pixel on the instrument is detected. Given that the colour model chosen is Gaussian, then α is similar to a confidence test on a gaussian. Hence for a 99% confidence that the pixel chosen is indeed the instrument colour (red usually), $\alpha = 2.58^2/2$.

A multiresolution version of this algorithm is employed to increase speed and convergence properties. The multiresolution MRF process used was as presented in.[5] The idea is to associate $m \times m$ labels with each other to define macropels. Thus, at level 3, 4×4 pixels define a single label, while at level 2 2×2 pixels define a single label. In the implementation used here, the iterative process starts at level 3 with refinement taking place at the other finer levels. Iterations at a level are terminated when no further change occurs at a particular level.

2.1.4. Typical Performance

Figures 2, 3 and 4 show the performance of this process for typical images. The process correctly segments the instrument in all views and the multiresolution approach generally leads to smoother segmentations. The same parameters were used in both views $\alpha = 2.58$, $\lambda = 5$ with equal success despite the varying material. The use of colour for segmentation in this case is clearly correct.

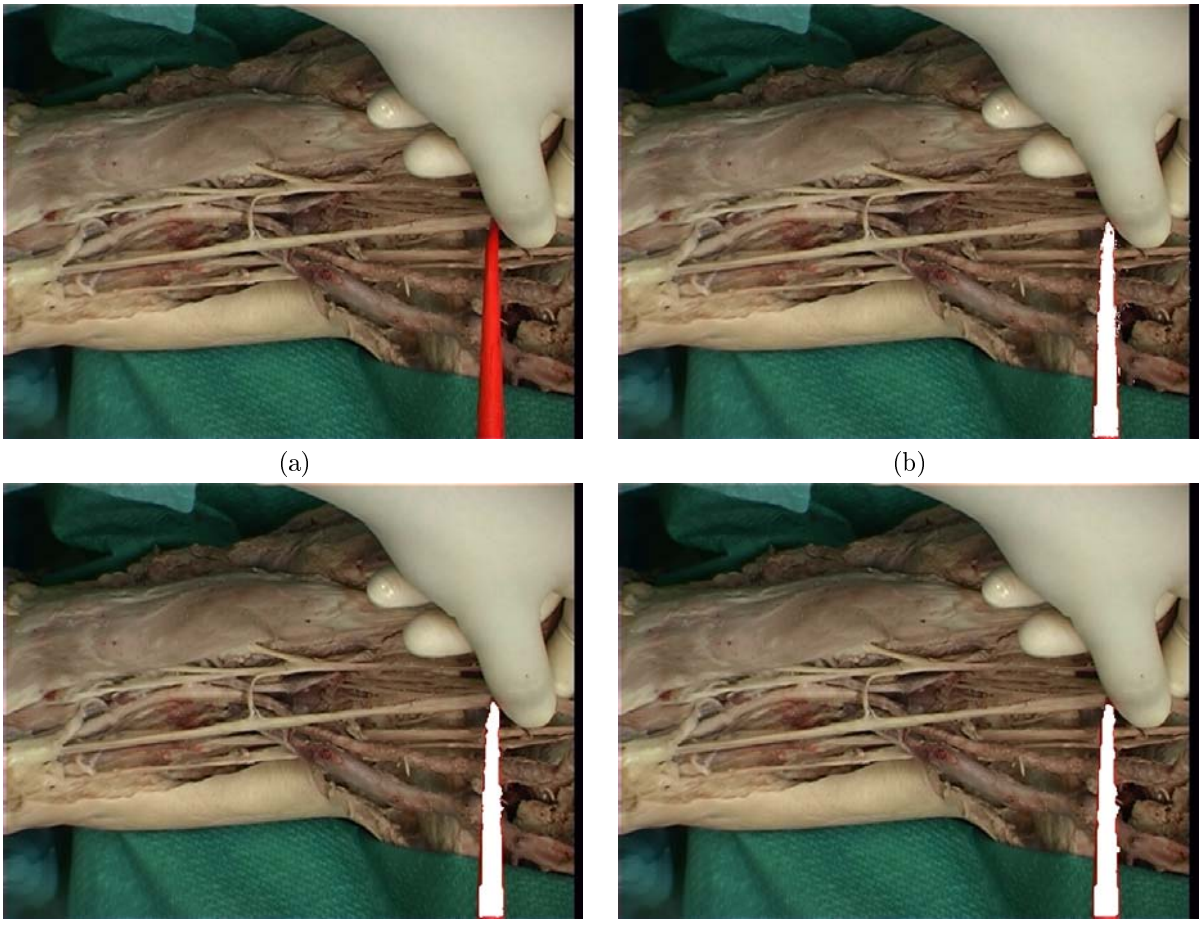

Figure 3. (a) Original image (b) Thresholded Log-Likelihood (c) Segmented image using ICM, after 5 iterations at finest level, with $\alpha = 2.58$, $\Lambda = 5$ (d) Segmented image using multiresolution ICM, after 2,2,1, iterations at the 3 levels (from coarse to fine); $\alpha = 2.58$, $\Lambda = 5$. The red instrument is correctly delineated in all views.

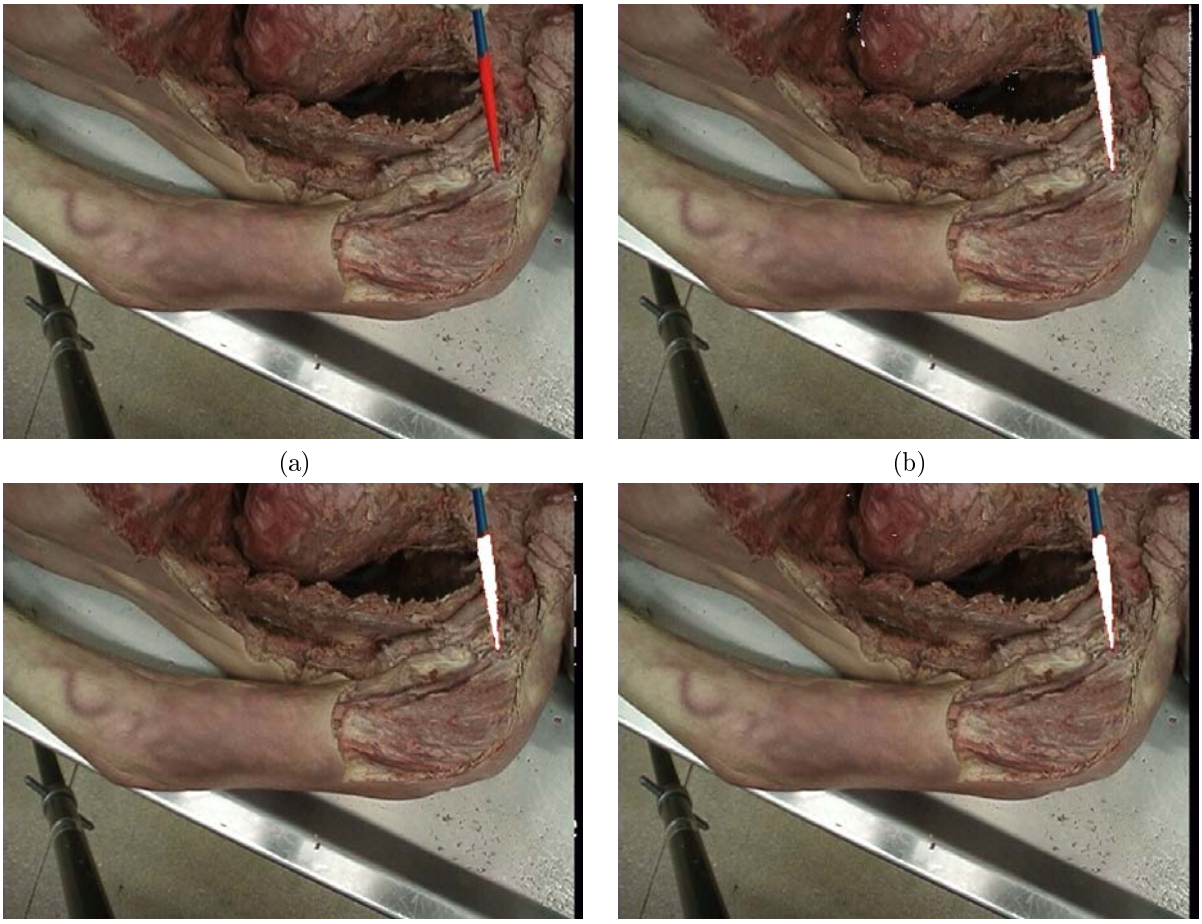

Figure 4. a) Original image (b) Thresholded Log-Likelihood (c) Segmented image using ICM, after 5 iterations at finest level, with $\alpha = 2.58$, $\Lambda = 5$ (d) Segmented image using multiresolution ICM, after 2,2,1, iterations at the 3 levels (from coarse to fine); $\alpha = 2.58$, $\Lambda = 5$. The red instrument is correctly delineated in all views.

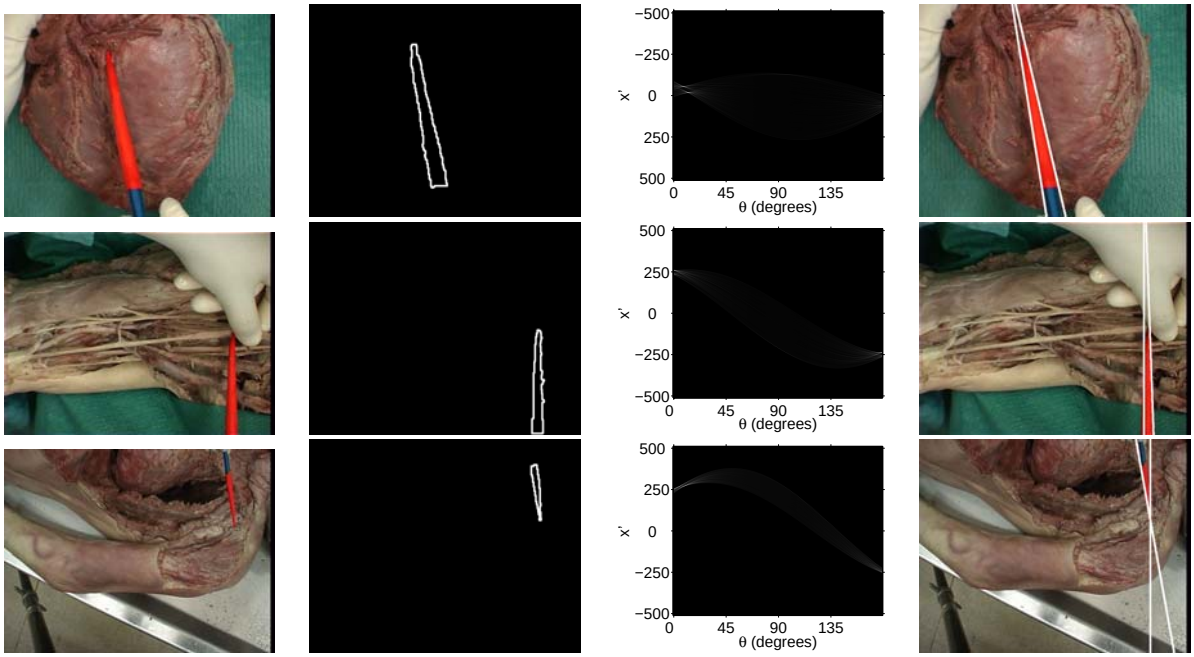

Figure 5. Instrument delineation on differing content. Left to Right: Original image, Edges of the segmented instrument, The Hough Transform of the segmented edges, and the estimated instrument boundaries superimposed on the image.

2.2. Delination

Estimating the location that the instrument is indicating requires an estimate of the location of the two sides of the instrument. The intersection of the lines $\mathbf{p}_h$ at the sides roughly indicates that location. Line estimation is performed using the Hough Transform on the binary label field $L(\mathbf{x})$. Due to the orientation of the pointer there are frames in which these lineas may be parallel and hence no intersection can be calculated. The location of the tip of the instrument in the previous frame is then taken as a rough estimate for this intersection point.

To estimate the exact tip of the instrument, a bounding box that just contains the segmented shape is calculated. One of the points on the boundary of the segmented shape will be the tip of the pointer. These points can be further restricted as those which lie on the bounding box boundary itself. The point in this set that is closest (in Euclidean sense) to the point $\mathbf{p}_h$ is chosen as the pointer tip. This process is illustrated in Figures 5.

3. SHAPE AND MOTION FOR PARSING

The goal of parsing the surgical footage automatically is to alleviate most of the time needed of the Medical professional during the post-production stage. Certain shots that are important to the viewer can be delimited by distinct motion patterns of the pointing instrument. For example, when the instrument is stationary, it is generally because the conducting doctor wants to emphasize the given structure. When there is a gentle motion with a constant velocity, the doctor is generally highlighting a specific region. If fast or erratic motion is observed, it is generally due to the doctor moving to another area. Finally, if the instrument is not in the shot, the shot is generally not important.

The goal is therefore to detect how much the instrument is currently moving. The orientation and shape of the pointer in the view is summarised by the shape of the main lobes in the Hough space used for instrument delineation. The geometric moment of the data in Hough space provides a concise expression of this information and is calculated as follows.

$$M(p) = \sum_{\mathbf{x}=[h,\ k]} (h - c_h)^p (k - c_k)^p \tag{8}$$

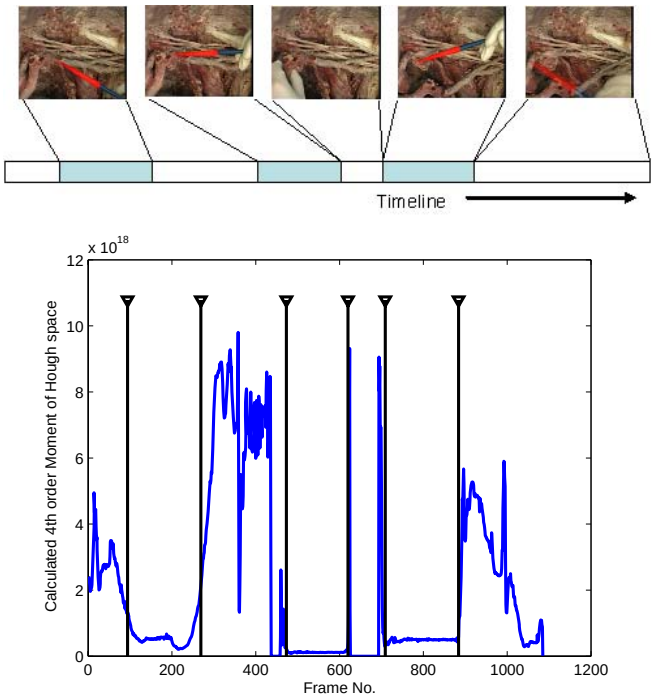

Figure 6. *Example of Motion Parsing using 4th order Moment in Hough space. The top sequence shows the content analysed in this 45 second portion of video. The potions where the instrument is in view are highlighted in the timeline. The plot beneath shows the 4th order moment of Hough space versus frame. The episodes where there is little motion are shown clearly as plateus of this signal. Thresholding the differential of the signal yields episode detection as shown by the black vertical lines.*

where $[c_h, c_k]$ is the central index in Hough space, and $M(p)$ is the pth order moment. In this 2D case, the horizontal and vertical axes refer to the ρ and θ parameters of the Hough space. Figure 6 shows the evolution of the 4th order geometric moment ($M(4,4)$) of the Hough Transform for each frame; hence motion of the instrument causes changes in this feature. Regions of interest in time are therefore delineated by low activity in this moment, i.e. low differential in time.

Due to the aggregation of an entire space to a single digit, this method is relatively robust. Two relatively similar frames will exhibit a moment difference proportional to the difference between the frames. By taking *high* order moments ($p > 2$), the risk of two frames of differing content having similar calculated moments is lowered. A useful property of moment calculation is that the calculated moment takes the value of zero upon the absence of the segmented instrument, indicating to the parsing system that this current shot is unimportant.

One of the main advantages in using moments of the Hough space for motion detection instead of a subsequent-frame difference technique is the emphasis it places on rotational movement as well as spatial movement. It can be seen that the gentle motion of illustrating a structure exhibits a relatively low moment difference compared to the sudden motion of the instrument leaving the frame or re-orienting itself. This property enables simple thresholding to achieve a relatively high success rate. The results of this can be seen in Figure 6. For that example of 45 secs a threshold of 1 on the differential yields semantically relevant shot boundaries to within 10 frames of their correct* location.

4. VIDEO CONTENT AUGMENTATION

Having indexed important points in the video using the pointer motion it is then possible to augment the content. The idea here is to highlight important semantic features in the field of view and so create more compelling content, and certainly more understandable training video. In a sense this idea is precisely that being exploited by comic book artists (see Figure 7) when they draw lines around objects to indicate motion for instance. The work by Massey and Bender[7] is also relevant here, although their purpose was to investigate how

*estimated visually

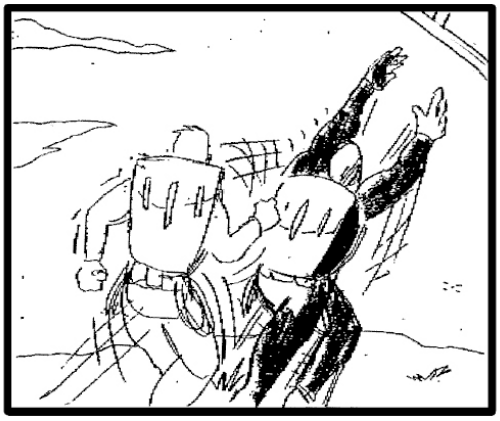

Figure 7. Onion Skinning *to create the effect of motion in a still.*[6]

(a) (b) (c) (d)

Figure 8. The motion of the instrument is superimposed on the image, hence highlighting the feature of interest.

motion could be portrayed in a still. Content augmentation has also been used to good effect by Denman et al[8] in sports.

Consider for the moment that a semantically relevant event in the video stream has been detected at a particular frame. In the dissection videos, the *start event* is the instant that the instrument can be considered to be in the same position for a few frames. After the motion of the instrument, another consecutive event of the instrument being in the same position can be considered as the *finish event*. Because the semantics of the detected event are known a-priori it is possible to exploit the fact that only motion of the instrument, between the identified *start event* and *finish event*, are relevant.

Given that the doctor performing the demonstration would use the pointer to highlight the conceptual shape of organs (for instance), a track of the pointer in the field of view could be made visible and post-processed in such as way as to highlight that contour on the image itself. Figure 8 shows this directly arising from the pointer track. Of course it would be necessary to smooth out the track since the Doctor cannot keep the tip steady enough given the scale of the zoom used. This is a releatively minor matter and can be achieved with a Gaussian FIR for instance. The amount of smoothing would depend on the effect required. A more interesting post-process is to highlight a band of material around the pointer This is particularly useful for highlighting linear structures embedded in the clutter of other anatomical features. Figure 9 shows a number of examples in which the a region around the estimated pointer track is enhanced (using colour histogram equalisation) and the area outside this region turned to grayscale. This has the effect of directing attention and also highlighting the feature of interest. This is important for demonstrations in lecture theatres in particular.

5. USAGE

Over 25 hours of video have been recorded. The automated process described here is being incorporated into a toolkit that can be used directly for automated post-processing. The pointer delineation is very robust and accurate and has been used successfully to parse and highlight 5 mins of summarised content. Demonstration material is available at `www.sigmedia.tv/anatomy`.

(a) (b) (c) (d)

Figure 9. Highlighting by contrast enhancement in a band around the path of the pointer.

6. ACKNOWLEDGMENTS

We thank the RCSI for funding this project.

7. FINAL COMMENTS

This work has presented a number of tools for indexing medical training video as well as augmenting the content in a useful way. The work is applicable to surgery in general. They key aspect is to locate and delineate the dissection instrument or pointer. This is facilitated by colouring the instrument uniquely so that it is well contrasted with the rest of the anatomical material. Indexing important points of interest in the video can be achieved by detecting when the pointer is relatively stationary. Content augmentation consists of superimposing tracks of the pointer over the image, hence indicating shapes of organs etc, as well as enhancing regions of interest around the pointer hence directing attention more effectively. The work presented here is in place in the RCSI and the first training videos are being used for the training of students in the new academic year 2005/06. The work continues along the direction of *cartoonising* anatomical features after indexing so that students may more easily appreciate their connection with conceptual drawings in standard medical texts.

REFERENCES

1. L. U. C. S. S. of Medicine, "www.meddean.luc.edu/lumen/meded/grossanatomy/dissector/."
2. T. V. Human, "http://www.uchsc.edu/sm/chs/open.html."
3. I. M. Curriculum, "http://imc.meded.com/integrated/demos/hademo/index.htm."
4. J. Besag, "On the statistical analysis of dirty pictures.," *Journal of the Royal Statistical Society B* **48**, pp. 259–302, 1986.
5. F. Heitz, P. Prez, and P. Bouthemy, "Multiscale minimization of global energy functions in some visual recovery problems," *CVGIP : Image Understanding* **59**, pp. 125–134, January 1994.
6. R. Dony, J. Mateer, and J. Robinson, "Automated reverse storyboarding," in *IEE 1st European Conference on Visual Media Production*, pp. 193–202, March 2004.
7. M. Massey and W. Bender, "Salient stills: Process and practice," *IBM Systems Journal* **35**(3,4), 1996.
8. H. Denman, N. Rea, and A. Kokaram, "Content-based analysis for video from snooker broadcasts," *Journal of Computer Vision and Image Understanding, Special Issue on Video Retrieval and Summarization* **92**, pp. 141–306, November/December 2003.