Correlating Histology and Spectroscopy to Differentiate Pathologies of the Colon

Duane Carey[1]
d.carey@medical-research-centre.com
Gavin Rhys Lloyd[1]
Neil Shepherd[1,2]
Nick Stone[1]
Catherine Kendall[1]
Toby Breckon[3]

[1] Biophotonics Research Unit, Gloucestershire Hospitals NHS Foundation Trust, Leadon House, Gloucester, UK

[2] Pathology Department, Gloucestershire Hospitals NHS Foundation Trust, Cheltenham, UK

[3] School of Engineering, Cranfield University, UK

Abstract

The techniques and procedures associated with histology are, in most cases, suitable for the diagnosis of colonic carcinomas. However, in cases such as epithelial misplacement the morphology of a stained tissue sample is homologous to that of cancer. This can lead to patients being misdiagnosed and undergoing unnecessary surgery.

To prevent this surgery we suggest that the epithelium of tissue samples be examined using infrared (IR) spectroscopy. In this study, IR maps of tissue sections were registered to standard histology images so that epithelial specific spectra could be collected. The differences between these spectra were explored by using Principal Component Analysis (PCA). This paper provides a novel protocol detailing how histology specific spectra can be collected. The potential usefulness of these spectra is demonstrated through the separation of epithelial misplacement cases and colonic carcinomas within PCA space.

1 Introduction

A pathologist will diagnose disease states by examining Haematoxylin and Eosin (H&E) stained tissue sections under a microscope. Staining enables the structural morphology of a tissue section to be highlighted and evaluated. Although this method of diagnosis is generally very accurate, some benign conditions can still be misdiagnosed as cancer. Epithelial Misplacement, EM, is an example of a benign pathology which often gets confused with Polyp Cancers (PC) because its morphology is homologous to that of invasive cancer.

EM polyps resemble PCs because of the environment in which they exist. EM polyps are associated with the sigmoid colon and in this region polyps are easily damaged. This is because the sigmoid colon is constantly fluxing and this movement will compress any polyps found there. The damage affects the polyps structure and forces epithelium from the exterior of the polyp into its interior. This forced movement means that when a tissue sample is sectioned for pathological assessment the sections will contain epithelial islands (indicated by a white arrow on Figure 2A). These islands are a sign indicative of invasive cancer and their presence causes pathologists to suggest that the region surrounding these benign polyps be removed from the colon [1].

However, for EM this surgery is unnecessary as only in invasive Polyp Cancers (PC) would these islands have malignant potential. In PC the islands are formed from epithelial cells which have moved under the influence of their own genetics. Therefore, the biochemistry of PC islands will be different to that of the islands within cases of EM. Infrared spectroscopy (IR) can be used to characterise these biochemical differences and facilitates the discrimination of cases of EM from PC [2], [3], [4].

2 Methods

In this study anonymous, retrospective, paraffin embedded tissue blocks were selected by a consultant histopathologist. From these blocks three contiguous 5µm thick tissue sections were cut and placed onto two normal histology slides and one IR reflective low *e* slide. One of the normal slides was stained with MNF116, a cytokeratin antibody, and the other with H&E. Images of the standard histology slides were made from the 1.25x objective lens of a Leica camera microscope. The IR images were measured using the linear array detector of the Perkin Elmer Spotlight 400 IR Spectrometer. The following protocol was then used to collect epithelial specific spectra for input into a classification model. All post collection image processing was conducted using Matlab R2007A (Mathworks, Natick USA) on a standard desktop personal computer equipped with an I7 Quad core 3.60GHz processor and 8Gb of RAM.

2.1 Histology Image generation

The standard histology images are acquired as a series of overlapping images in a grid like fashion. The overlapping images for each tissue section must first be stitched together before this sample can be further analysed. To enable this, Scale Invariant Feature Transform (SIFT) descriptors were used to find points of correspondence between the overlapping images. The SIFT algorithm finds these points from the dominant edge features of an image. It does this by constructing a difference of Gaussian pyramid for a pair of overlapping images. A subsequent search for maxima and minima within this pyramid means that keypoints [5] can be found. The minimum Euclidean distance between gradient and orientation feature vectors made for these keypoints enables points of correspondence between the overlapping images to be defined (Figure 1).

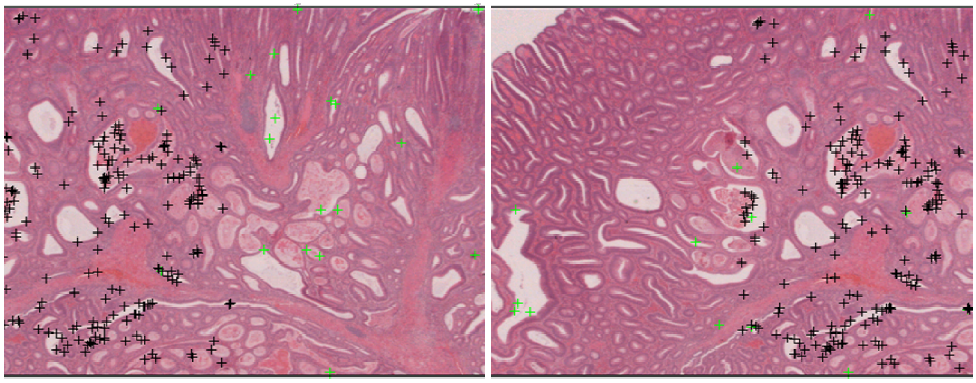

Figure 1. The matching of points of correspondence between overlapping images. The black points are the SIFT points retained after the calculation of the NCC. The green points are the SIFT points rejected.

However, the SIFT algorithm identifies many false positives but these can be removed through calculating the Normalised Correlation Coefficient (NCC) [6]. This was initiated

in this instance by defining 100 by 100 pixel segments around one of the overlapping images matched SIFT points. These image segments were used with the image it overlapped in the calculation of the NCC. As the images are of the same scene these segments should be highly correlated with regions of the overlapping image. Therefore, only SIFT points which produced a NCC above 0.95 were retained (black points in Figure 1) and allowed for the overlapping images to be stitched together. A translation was found between two images when the Euclidean distance between two consecutive SIFT points was the same [6]. This process was repeated until all of the overlapping images of a tissue section were stitched together (Figure 2A&B).

2.2 Chemical image acquisition and quality control

The IR images were acquired with a spectral resolution of 6 cm^{-1} and a spatial resolution of 25μm. Since spectra will be used within classification models it is very important that they are of an appreciable quality. To ensure this an area was summed between 1500cm^{-1} and 1600cm^{-1} wavenumbers, a biologically important area [3], [4], for all of the spectra contained within an image. Anything which was found to be two standard deviations above or below the mean of this sum was then excluded from further analysis.

2.3 Image registration

Registration of images in this work is difficult because they are acquired from different modalities. Therefore, methods which rely on the maximisation of similarity metrics are not well suited. In this instance, registration was achieved by maximising the overlap of binary masks. These binary masks were made from analysing each image of a tissue sample with PCA. The PCA scores were thresholded using a supervised *t*-test, which is equivalent to Otsu's method [7], that had some *a priori* information about the intensities of the images background pixels. This information was obtained by manually selecting a region of background pixels from a principal component that explained the variation between the background and foreground of an image. To create a binary mask using this background information a threshold is iteratively increased between the minimum and maximum of the selected principal component. At each threshold a *t*-value is produced that compares pixels below this threshold against the pre-selected background pixels. The threshold which causes the *t*-value to exceed the 95% confidence limit is used to produce a binary mask. This mask is used in the minimisation of Equation 1 so that an optimal linear transformation, T, can be found. T is composed of a rotation, translation and scaling factor and was optimised using the simplex search method [8].

$$\mathrm{E} = \sum_1^{\mathrm{N}} \sum_1^{\mathrm{M}} ((\mathrm{I2} - \mathrm{T}(\mathrm{I1})))^2 \qquad (1)$$

Here *E* is the total image registration error, *I1* represents the histology image, *I2* represents the IR chemical map and *(N, M)* represents the images spatial dimensions. All of the multimodal images were zero-padded to the same dimension.

However, these linear transformations are not sufficient because of the stretching which occurs to a tissue sample when it is being prepared. Therefore nonlinear registration is needed and this was achieved by fitting a cubic B-spline grid to an image [9]. The nonlinear transformation found was the result of moving the vertices of this grid to minimise Equation 2:-

$$E = w_i E_{img} + (w_d E_{div} + w_r E_{rot}) + w_c E_{cons} \quad (2)$$

Where E is the objective function to be minimised, E_{img} is a measure of the dissimilarity between the images I_2 and $T(I_1)$, $(w_d E_{div} + w_r E_{rot})$ is a regularisation term

based on the divergence and curl of the B-splines and E_{cons} is a term relating to the consistency of the registration and prevents stretching. The w_i, w_d, w_r and w_c weights allow an element of control over the deformation of the spline grid [9].

2.4 Image segmentation

The nonlinear transformations applied to the multi modal images enable very accurate feature correspondences to be achieved. This alignment allows for an image containing strong features to be used in the probing of other images where these regions of interest are less apparent. Here epithelial islands are only apparent within the standard histology images and thus to gather information on the spectroscopic characteristics of these islands the warped standard histology images were segmented. The segmentation was carried out by registering MNF116 antibody stained images and H&E images together. The MNF116 antibody is specific for a tissue sections epithelium (Figure 2 B) but it also stains other regions of a tissue sample (e.g. blood vessels). However, H&E images strongly stain blood vessels red and so the registration of the different stained images together enables the use of the Consensus Principal Component Algorithm (CPCA) algorithm [10] in preserving only the epithelium within an image. Simple manual thresholding of the CPCA super-scores allowed a binary mask to be produced and enables epithelial specific spectra to be collected and used with exploratory statistical methods.

2.5 Discriminant Analysis

In this instance, discrimination was achieved by using NIPALS PCA [11]. It facilitated the exploration of the spectroscopic variation that existed between the different pathology groups. All images of a tissue sample were vector normalised, mean centred and in the case of the IR images baseline corrected. This prevents any anomalies from affecting the results of the PCA.

The determination of significant variation by ANalysis Of VAriation (ANOVA) with 95% confidence limits assures that the different pathology groups can be separated. ANOVA uses the ratio of the intra and inter group variances against a critical value determined from the F distribution in the selection of discriminatory components. In this instance the three most discriminatory components were used so that the variation between the different pathology groups could be visualised.

3 Results

The use of histology specific spectra requires that the final image transformations be as accurate as possible. The end results are presented in Figure 2 C&D. As these results are only preliminary the root mean square error of the image differences, along with manual inspection, was taken as a measure of accuracy. The fact that these images came from different modalities means that the binary representation of the tissue foreground within these images will be slightly different. The main difference which is found is how the mucin pools are represented. This difference can be explained by the varying paraffin content that exists between the different sections produced for a tissue sample. This causes one image to have more holes within it than the other (indicated by the blue arrows on Figure 2 C&D). Even though these defects are apparent we can still determine that the epithelial islands have been accurately registered. This is because in the interior of the

difference image the only significant change apparent is around the edges of the images epithelial islands (indicated by red arrows on Figure 2 C&D).

From these registered images epithelium specific spectra was collected via image segmentation. In this case the output of the CPCA algorithm (Figure 2E) which acted on images presented in Figure 2 A&B was used to generate epithelial specific spectra.

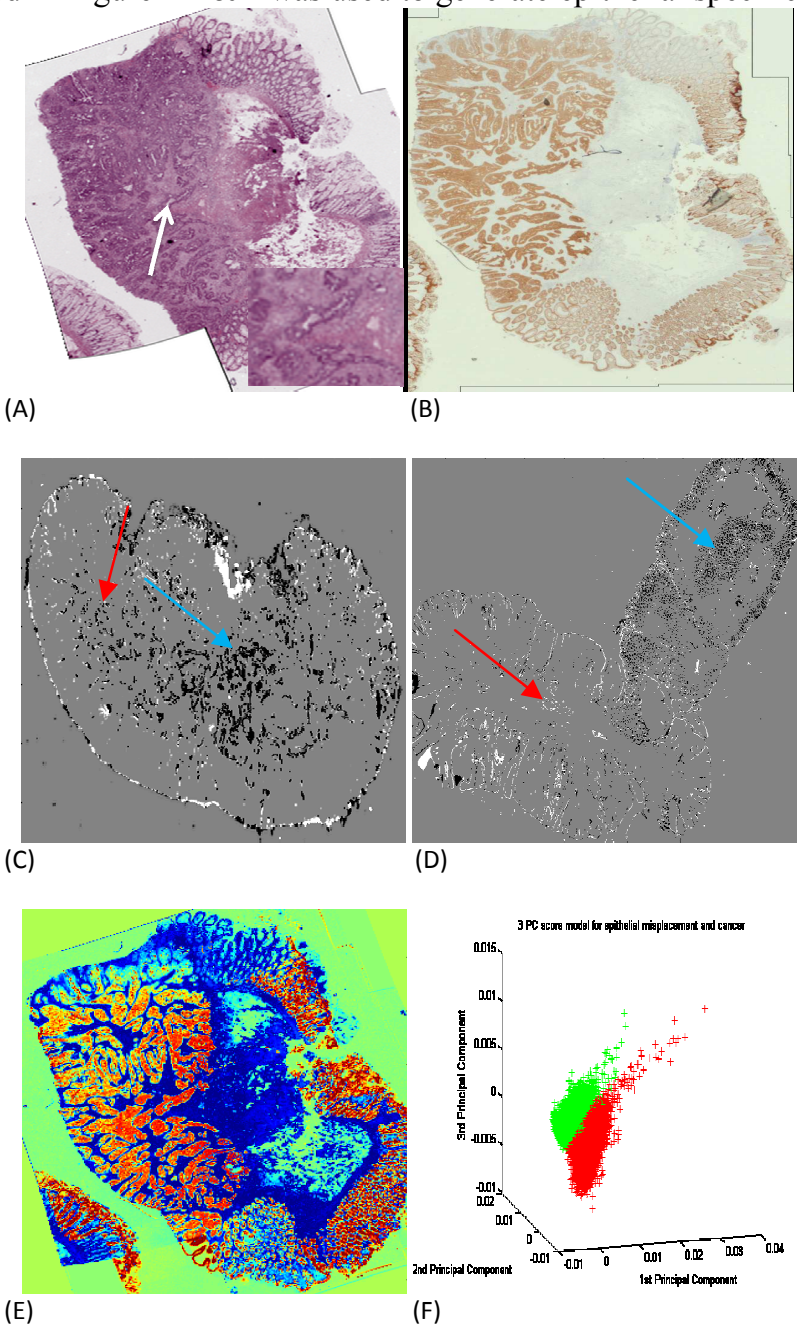

Figure 2. A depiction of how epithelium specific spectra can be collected and used: (A & B)Unregistered MNF116 and H&E histology images stitched together from images acquired from the Leica camera microscope (the white arrow indicates an epithelial island which has been enlarged in the bottom of A); (C & D)the differences between the multimodal images after application of an nonlinear transformation (blue and red arrows indicate the observer criteria used for accuracy);(E)Result of the CPCA algorithm acting on combined MNF116 and H&E images; (F)3D PCA score plot made from the epithelial specific spectra (red points are PC spectra and the green points are EM spectra).

This method allowed for 9,804 spectra to be collected from six samples, three from each pathology group and roughly half of the spectra were contributed from each group. These spectra were used within a PCA model so that the variation between the different pathology groups could be explored. This exploration proved fruitful as is evident from

Figure 2F and potentially confirms that epithelial specific spectra, the green points in Figure 2F, can indeed be used to differentiate epithelial misplacement from cancer, the red points in Figure 2F.

4 Conclusion

A procedure for the intermodal registration of digital histology images and IR spectroscopic images is described and successfully applied to 6 samples. An image segmentation algorithm was then applied to the registered images so that epithelial islands could be located and their associated spectra collected. PCA was then used to show the potential of these abstracted spectra in discriminating EM from PC (Figure 2F). This type of approach could lead to improved cancer diagnostics and reduce the number of EM patients receiving unnecessary treatments [12].

The enlargement of the sample size is an important aspect of future work along with automating the classification. The automated discrimination of EM from cancer can be achieved with Linear Discriminant Analysis (LDA) or Support Vector Machine (SVM) [13] techniques and will help validate the PCA model presented in Figure 2F. Another important future aspect to consider is the method of error analysis used to determine the accuracy of registration. As the collection of specific spectra rests on the accurate registration of images more advanced consistency methods between the modalities of a tissue sample will be evaluated in the future [14].

[1] V. C. Petersen, a L. Sheehan, R. L. Bryan, C. P. Armstrong, and N. A Shepherd, "Misplacement of dysplastic epithelium in Peutz-Jeghers Polyps: the ultimate diagnostic pitfall?," *The American journal of surgical pathology*, vol. 24, no. 1. pp. 34-9, Jan-2000.

[2] C. Kendall et al., "Vibrational spectroscopy: a clinical tool for cancer diagnostics.," *The Analyst*, vol. 134, no. 6, pp. 1029-45, Jun. 2009.

[3] J. T. Kwak, S. M. Hewitt, S. Sinha, and R. Bhargava, "Multimodal microscopy for automated histologic analysis of prostate cancer.," *BMC cancer*, vol. 11, no. 1, p. 62, Jan. 2011.

[4] J. J. Wood, C. Kendall, G. R. LLoyd, N. A Shepherd, T. A Cook, and N. Stone, "Infrared spectroscopy to estimate the gross biochemistry associated with different colorectal pathologies," *Imaging*, vol. 8087, p. 80870P-80870P-9, 2011.

[5] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*. vol. 60, no. 2, pp. 91-110,2004.

[6] V. Rankov, "An Algorithm for image stitching and blending," *Proceedings of SPIE*, vol. 5701, pp. 190-199, 2005.

[7] J.-hao Xue and M. D. Titterington, "t-Tests, F-Tests and Otsu's Method for Image Thresholding," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2392-2396, 2011.

[8] J.C .Lagarias, J. A. Reeds, M. H. Wright and P. E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," *SIAM Journal of Optimization*, vol. 9, no 1, pp. 112-147, 1998.

[9] I. Arganda-carreras, C. O. S. Sorzano, and R. Marabini, "Consistent and Elastic Registration of Histological Sections using Vector-Spline Regularization," *Computer Vision Approaches to Medical Image Analysis In Computer Vision Approaches to Medical Image Analysis*, vol. 4241, pp. 85-95, 2006.

[10] J. A. Westerhuis, T. Kourti, and J. F. Macgregor, "Analysis of multiblock and hierarchical pca and pls models," *Journal of Chemometrics*, vol. 321, pp. 301-321, 1998.

[11] R. Brereton, "Chemometrics. Data Analysis for the Laboratory and the Chemical Plant," *Wiley,* pp 194, 2004

[12] F. L. Greene, "Epithelial Misplacement in adenomatous polyps of the colon and rectum," *Cancer*, vol. 33, no. 1, pp. 206-217, Jan. 1973.

[13] J.W. Han, T.P. Breckon, D.A. Randell, and G. Landini, "The Application of Support Vector Machine Classification to Detect Cell Nuclei for Automated Microscopy", *In Machine Vision and Applications*, Springer, vol 23, no. 1, pp. 15-24, 2012.

[14] E. T. Bender and W. A. Tome, "Utilization of consistency metrics for error analysis in deformable image registration," *Medical Physics*, vol. 54, no. 18, pp. 5561-5577, 2009.