AMSI VACATION RESEARCH SCHOLARSHIPS 2020–21

Get a Thirst for Research this Summer

Statistical learning for time-to-relapse of colorectal cancer patients

Nauvoo Perez

Supervised by Dr. Hien Nguyen

La Trobe University

Vacation Research Scholarships are funded jointly by the Department of Education, Skills and Employment and the Australian Mathematical Sciences Institute.

AMSI Australian Mathematical Sciences Institute

Australian Government
Department of Education, Skills and Employment

Table of Contents

Abstract 3

1 Introduction 4

1.1 Statement of Authorship 5

2 Data Description 5

3 Image Analysis 6

3.1 Feature Extraction 6

3.1.1 Gradient magnitude matrix 6

3.1.2 Texture matrix 7

3.1.3 Colour matrix 8

3.2 Feature Clustering 8

3.2.1 K-Means clustering 8

3.2.2 Markov's random fields 10

4 Results and Discussion 10

5 Conclusion 11

References 12

Abstract

Colorectal cancer (CRC) heavily affects a large proportion of the Australian working population. As patients who undergo curative surgery are more likely to fatally relapse, predicting those who are likely to relapse could help improve the survival rate. This project aims to create a tool that can aid non-clinicians identify lymphocytes and various regions in a colorectal histology. It can later be applied to help determine whether tumour infiltrating lymphocytes (TILs) are predictors of relapse in CRC patients. Image information related to edges, texture, and colour were obtained using the Canny edge detection algorithm, Grey Level Co-Occurrence Matrix (GLCM), and the RGB colour space. The features were clustered using K-means and Markov random fields (MRF). Preliminary findings suggest that the tool assists a non-specialist identify TILs and non-TILs (lymphocytes not located within the tumour region) provided they are familiar with the characteristics of a lymphocyte.

1 Introduction

Colorectal cancer (CRC) has been identified as the second deadliest cancer in Australia since 1982 by the Australian Institute of Health and Wellbeing [1]. In addition, it is the fourth leading cause of death in adults between the ages of 45 and 64 [2]. Given that the working population of Australia is between 25 and 64 years of age, CRC is not only a medical issue, but also an economic concern. While mortality rates have been declining, it is still concerning that 1 in 3 patients with CRC who undergo curative surgery have fatal recurrences [3]. This brings to the forefront the importance of predicting which patients are likely to relapse, and the time-to-relapse of patients at risk. Currently, physicians use different colorectal nomograms to make these predictions. However, in comparison to other nomograms used in oncology, colorectal nomograms are underdeveloped [4]. Limited validation studies and strict patient range are but two examples of issues that need to be addressed. TILs have recently been shown to be significant predictors of relapse in non-small cell lung cancer (NSCLC) [5]. As high levels of immune reaction are known predictors of positive prognosis in CRC patients [6], it is logical to investigate whether the phenomenon observed in NSCLC can be replicated in CRC.

In order to assist the proposed investigation, this project set out to create a tool that can help non-specialists identify TILs and non-TILS, in stage II and III CRC histology slides. Multiple journal articles have found success in combining mathematical image analysis techniques and artificial intelligence, which motivated this study [7-11].

In this work, feature information relating to edges, texture, and colour were extracted from histology images. These information were respectively obtained by implementing a portion of the Canny edge detection algorithm [7], calculating the GLCM [8-10], and extracting the RGB values of the raw image [9]. K-means clustering [10] and Markov random fields (MRF) [11] were then used to group features of interest. Figure 1 depicts the workflow developed.

As far as the researcher is aware, there is currently no published paper that uses the combination of image analysis techniques used in this project. The paper is set out as follows – firstly, this report will describe the data and software used in this project. Secondly, it will discuss the two sections of image analysis used in this research, feature extraction and image clustering. The results of this study will be discussed after which a summary of the project will be presented.

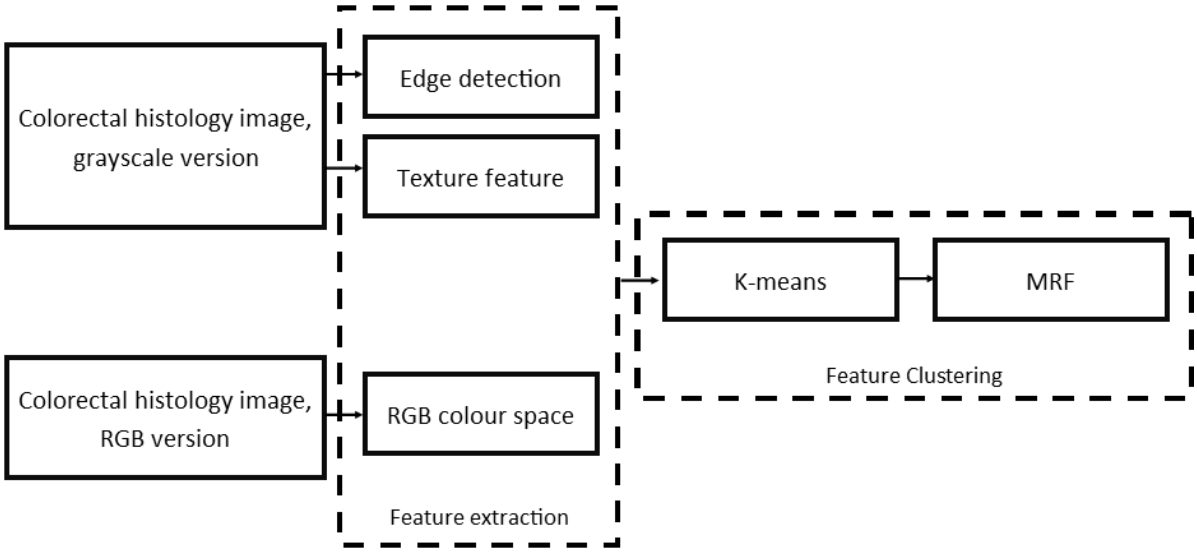

Figure 1. Image processing overview.

1.1 Statement of Authorship

The project idea was presented by Dr. David Williams from the Olivia Newton-John Cancer Research Institute. All raw images and annotations were also provided by Dr. Williams. The scope of the project that was undertaken during the 6-week AMSI Summer Vacation Research Internship was formulated by Nauvoo Perez and Dr. Hien Nguyen. Guidance and supervision were provided by Dr. Nguyen. The project was funded by AMSI and the Australian Department of Education.

2 Data Description

The raw image used in this project is a representative section of a colorectal tissue obtained using the imaging software QuPath [12]. The image was taken at 26x magnification and is 2302 x 1436 pixels. The PNG image format was used throughout the entire project to minimise the loss of information through image compression.

The image was stained using haematoxylin and eosin (H&E), which is generally used to identify the overall cellular and tissue structure components [13]. This technique stains the tissue in different combinations, shades, and hues of pink and blue where components that share similar features are stained the same colour.

The entire project was coded in R using RStudio [14]. The `load.image` function from the `imager` [15] library was used to load in all images used in the project, and the `ggplot` function from the `ggplot2` [16] library was used to graph the images.

3 Image Analysis

3.1 Feature Extraction

Feature extraction is the process of extracting interesting elements of an image. This information can then be used in additional processes to make conclusions based on the intended outcome. The three image features that were of interest in this project were edges, texture, and colour.
Two versions of the image were used in this section. The original image was used to extract the RGB colour space. A grayscale version of the original was used as input for both the edge detection and GLCM. The matrices generated are used as input for the clustering algorithm.

3.1.1 Gradient magnitude matrix

As it is useful in image segmentation, edge detection was implemented. In addition to identifying the boundaries of features that are of interest, it also decreases or removes information that is not required. Although there are numerous ready-to-implement edge detection functions in R, this project manually implemented the Canny edge detection algorithm's first two stages – noise reduction and calculation of the gradient magnitude [17]. This decision was taken to provide the researcher the opportunity to explore and better understand the processes involved in detecting feature edges in an image.

A. Image noise reduction

Denoising is an integral component of image processing as it allows the technology to ignore superficial information and concentrate on the key features. For this project a van Vliet-Young filter, a recursive approach used to approximate a Gaussian filter, was applied [18]. The filter approximated is the Gaussian blur which is given as

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{s\sigma^2}}$$

where x and y respectively denote the horizontal and vertical distance from the origin. The standard deviation of the distribution used throughout is $\sigma = 2$.

B. Calculating gradient magnitude

The gradient of an image is obtained by calculating the change in brightness or colour of an image. By calculating the rate of change in gradient, in essence the magnitude, information such as a change in orientability, depth, or lighting can be obtained.

The gradient of an image is a vector containing partial derivatives with respect to x and y. It is given as

$$\nabla \mathrm{f} = \begin{bmatrix} \mathrm{G_x} \\ \mathrm{G_y} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix},$$

where $\frac{\partial f}{\partial x}$ is the horizontal partial gradient and $\frac{\partial f}{\partial y}$ is the vertical partial gradient. This project made use of the existing `get_gradient` function from the R library `imager` [15]. The magnitudes were calculated using the formula

$$\text{Magnitude} = \sqrt{{G_y}^2 + {G_x}^2}.$$

3.1.2 Texture matrix

GLCM is a method of extracting textural information from an image by taking into account the relationship between a pixel and its immediate right neighbour [19]. The matrix is composed by tabulating the frequency of different grey level pairings occurrences in the image. Texture information is then obtained by calculating different statistics using the matrix. There are no set rules when deciding which texture information measurements would provide the optimal output, though studies suggest that combining different information tend to generate better results [20]. Of the 14 image features originally proposed by Haralick, this project uses the variance, given by

$$\text{variance} = \sum_i \sum_j (\mathrm{i} - \mu)^2 \, \mathrm{p(i, j)},$$

where $p(i, j)$ is the (i,j)th entry of the generated matrix. The variance is primarily calculated to determine the similarity of pixel intensity within a region. This was implemented using various functions from the `raster` and `glcm` R packages [21, 22].

3.1.3 Colour matrix

When working with histology slides, colour is an important feature as different staining techniques produce different coloured slides depending on the intended outcome. For example, immunohistochemistry is the technique used to identify a specific protein in a tissue. Depending on the reagent used, the slides may be stained a different colour, though brown is the most common. To capture this information the image is converted into a data frame, from which the colour space RGB values were extracted.

3.2 Feature Clustering

Image segmentation is the process of grouping information so that the pixels within a category share similar characteristics. This project applied K-means clustering to automatically segment the different features of interest, which include the tumour region, non-tumour region, lymphocytes, and non-lymphocytes. MRF is then applied to refine the image by reducing noise and improving pixel grouping.

3.2.1 K-Means clustering

K-means clustering is an unsupervised method of partitioning n observations $(x_1, x_2, \dots, x_n)$ into k groups $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$, that share similar features using only input vectors. Given an initial set of k means $m_1^{(1)}, \dots, m_k^{(1)}$, each observation x_n is assigned to to the closest $m_i^{(t)}$ calculated based on the Eucledian distance

$$S_i^{(t)} = \left\{ x_n : \left(x_n - m_i^{(t)} \right)^2 \leq \left(x_n - m_j^{(t)} \right)^2 \forall j,\ 1 \leq j \leq k \right\}, \quad *$$

where x_n is strictly assigned to only one $S^{(t)}$. The mean of each group is then re-calculated using

$$\mathrm{m}_\mathrm{i}^{(\mathrm{t}+1)} = \frac{1}{\left|\mathrm{S}_\mathrm{i}^{(\mathrm{t})}\right|} \sum_{\mathrm{x_j} \in \mathrm{S}_\mathrm{i}^{(\mathrm{t})}} x_j. \qquad **$$

Equation (*) and (**) are repeated until the algorithm converges, in the sense that $m_i^{(t)} = m_i^{(t+1)}$, for some t. This algorithm is commonly known as Lloyd's algorithm [23]. The algorithm is implemented using the function `kmeans` from the R base library [14].

This process can be plainly described as follows:

1. k numbers of centroids are randomly placed.
2. Observations are then assigned a group based on their distance from each centroid.

3. The centre or mean of each group is calculated, this then becomes the new centroid.
4. 2 and 3 are repeated until either the centroid no longer shifts (local minimisation of within-group sums is achieved) or the number of iterations is reached.

It should be noted that minimisation of within-group variation is not guaranteed under the K-means clustering algorithm, and that only the convergence to a local optimum is ensured.

To evaluate the optimum number of clusters, this project used the elbow method [24]. This approach involves computing the within-group variability depending on the number of clusters, which we take between one to fifteen. The optimal number, k, of groups is chosen based on the number of clusters whereby increasing the group size by one would not drastically improve the within-group variability, in essence where the graph bends as seen in Fig 2. While the graph shows the ideal number of groups to be either three or four, $k = 6$ was chosen as there were additional features, such as the white spaces and the general pink background, that the researcher wanted to keep separated. The image produced by the algorithm was undistinguishable to the original image through human perception. This was addressed by converting the default six colours into more distinguishable colours, such as primary and secondary colours.

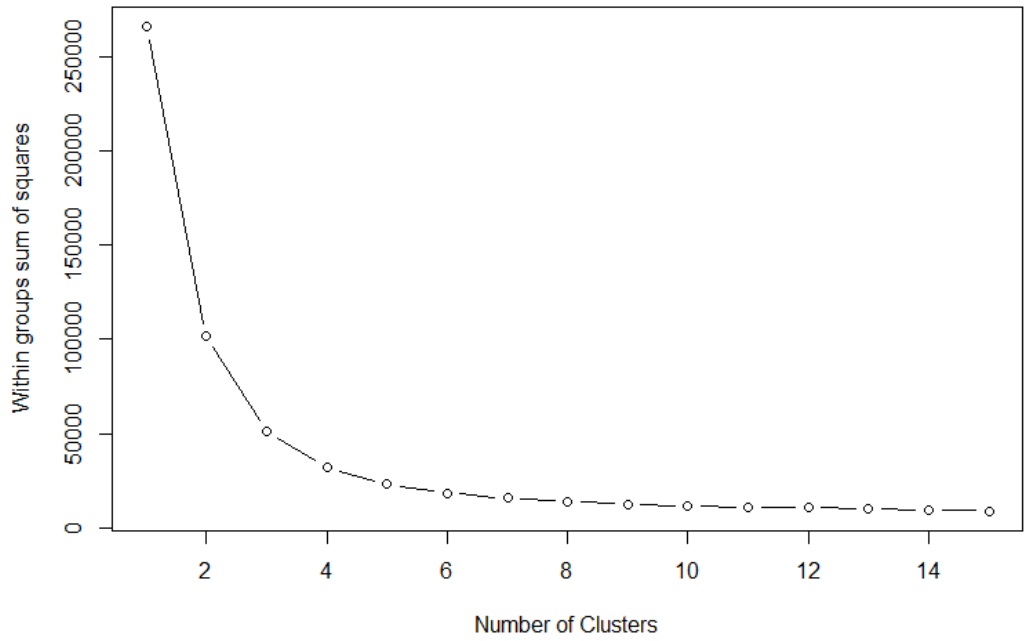

Figure 2. Graphical representation of reduction of within-group sum of squares as number of k increases.

3.2.2 Markov random fields

The MRF model assumes that images are generally smooth, essentially neighbouring pixels tend to look similar except for areas with high gradient magnitude. If an image has high variability, such as the image produced by the K-means algorithm, MRF can be applied to reduce spatial image noise, resulting in a more cohesive image [25]. This process involves using a probability function to predict which group the pixel of origin belongs to, based on its neighbouring cells.

The MRF model in this project was implemented as follows:

A. Obtain the Moore Neighbourhood, N_z, of distance $d = 1$ around the coordinate z, where $z = (x, y)\ \epsilon\ \mathbb{Z}^2$. Moore neighbourhood, also known as the rook's case, are the 8 surrounding pixels z [26]. N_z is given as

$$N_z = \{(i,j): max\{|x-i|, |y-j|\} \leq 1, (i,j) \neq (x,y)\}.$$

B. Let $C_z = \{C_\zeta : \zeta \epsilon N_z\}$ be a set of observations in the neighbourhood of $\boldsymbol{z}$. Also let $\eta_z = \eta(\mathrm{C}_z)$ be a function of the neighbourhood.

C. Construct a multinominal logistic regression MRF model using the K-means image clusters as the response variable and η_z as the predictor variables.

D. Use multinominal logistic regression to predict the responses.

E. Plot the response predictions.

4 Results and Discussion

Based on the results from Section 3, it was determined that the tumour regions contained high concentrations of cytoplasmic material that surround lymphatic cells. This made a significant enough difference that the tumour and non-tumour regions can be easily differentiated. As the purpose of this study was to create an aiding tool, this visual distinction was sufficient. It is possible to take this finding one step further by calculating whether the difference in ratio of this material is statistically different between tumour and non-tumour regions. This has the potential to provide a mathematical means of distinguishing between both regions.

In addition, the clustering algorithm made it easier to identify potential lymphocytes, although those were clustered with features that were clearly not lymphatic cells. As such, the colour ratio cannot be used to determine the density of lymphocytes within a given area. Applying the MRF model assisted in reducing confusion when it came to distinguishing between lymphatic and non-lymphatic cells. However, it did not eliminate the clustering issue. An issue encountered with MRF

was the computational power required to implement the model. As such it was only applied to segments that were a max size of 100 by 100 pixels.

There did not appear to be any human discernible differences when a comparison was made between an RBG-only matrix and gradient magnitude + GLCM + RGB matrix clustered image. These findings contradicted a study that found a combination of GLCM + RGB to produce more accurate classification results compared to using only RGB information [9]. Although it should be mentioned that the study implemented a support-vector machine (SVM) for the classification process. As SVMs were not used in this project, this study is unable to quantify the advantage of using gradient magnitude + GLCM + RGB matrix over an RGB-only matrix if machine learning classifying algorithms are utilised.

5 Conclusion

This paper set out to create a tool that can assist non-clinicians identify different features in a colorectal histology slide with minimal specialist input. It implemented various image analysis methods that were found to be successful in processing medical images, though the specific combination used have not been previously explored. Preliminary findings show that the developed technology can assist non-specialists identify a lymphocyte provided they are familiar with its features – small, spherical, and solid colouring. In addition, the image segmentation methods provided results that made it possible to differentiate between regions within the tissue. Finally, the project suggests that more sophisticated classification methods are required to make better use of the feature information extracted from the image.

References

1. Australian Institute of Health and Welfare 2020, *Cancer data in Australia*. AIHW: Canberra.
2. Australian Institute of Health and Welfare 2020, *Deaths in Australia*. AIHW: Canberra.
3. Kievit, J & Bruinvels, D 1995, *Detection of recurrence after surgery for colorectal cancer.* European Journal of Cancer. vol 31, no. 7-8, p. 1222-1225.
4. Kawai, K, et al. 2015, *Nomograms for colorectal cancer: a systematic review.* World journal of gastroenterology. vol 21, no. 41, p. 11877.
5. Corredor, G, et al. 2019, *Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer.* Clinical cancer research. vol. 25, no. 5, p. 1526-1534.
6. Huh, JW, Lee, JH, & Kim, HR 2012, *Prognostic significance of tumor-infiltrating lymphocytes for patients with colorectal cancer.* Archives of surgery. vol 147, no. 4, p. 366-372.
7. Hamad, YA, Simonov, K, & Naeem MB 2018, *Brain's tumor edge detection on low contrast medical images*. in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*. IEEE.
8. Htay, TT & Maung, SS 2018, *Early stage breast cancer detection system using glcm feature extraction and k-nearest neighbor (k-NN) on mammography image* in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE.
9. Kavitha, J & Suruliandi, A 2016, *Texture and color feature extraction for classification of melanoma using SVM*. in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. IEEE.
10. Wei, L, Gan, Q, & Ji ,T 2017, *Cervical cancer histology image identification method based on texture and lesion area features.* Computer Assisted Surgery. vol 22, p. 186-199.
11. Xu, X et al. 2019, *Automated brain region segmentation for single cell resolution histological images based on markov random field.* Neuroinformatics, p. 1-17.
12. Bankhead, P et al. 2017, *QuPath: Open source software for digital pathology image analysis.* Scientific reports. vol. 7, no. 1, p. 1-7.

13. Chan, JK 2014, *The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology.* International journal of surgical pathology. vol. 22, no. 1, p. 12-32.
14. R Core Team 2020, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.
15. Barthelme, S 2020, *imager: Image Processing Library Based on CImg*.
16. Wickham, H 2016, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
17. Canny, J 1986, *A computational approach to edge detection.* IEEE Transactions on pattern analysis and machine intelligence, p. 679-698.
18. Young, IT & Van Vliet, LJ 1995, *Recursive implementation of the Gaussian filter.* Signal processing. vol. 44, no. 2, p. 139-151.
19. Haralick, RM, Shanmugam, K & Dinstein, IH 1973, *Textural features for image classification.* IEEE Transactions on systems, man, and cybernetics, p. 610-621.
20. Rosenberger, C & Cariou, C 2001, *Contribution to texture analysis*. in *Proceedings of the International Conference on Quality Control and Artificial Vision*.
21. Zvoleff, A 2020, *glcm: Calculate Textures from Grey-Level Co-Occurrence Matrices (GLCM)*.
22. Hijmans, RJ 2020, *raster: Geographic Data Analysis and Modeling*.
23. Lloyd, S 1982, *Least squares quantization in PCM.* IEEE transactions on information theory. vol. 28, no. 2, p. 129-137.
24. Syakur, M et al. 2018, *Integration k-means clustering method and elbow method for identification of the best customer profile cluster*. in *IOP Conference Series: Materials Science and Engineering*. IOP Publishing.
25. Geman, S & Graffigne, C 1986, *Markov random field image models and their applications to computer vision*. in *Proceedings of the international congress of mathematicians*. Berkeley, CA.
26. Ménard, A & Marceau, DJ 2005, *Exploration of spatial scale sensitivity in geographic cellular automata.* Environment and Planning B: Planning and Design. vol. 32, no. 5, p. 693-714.