Automated segmentation and classification of nuclei in histopathological images

Sanjay Vincent and J. Chandra*

Department of Computer Science,
CHRIST (Deemed to be University),
Bangalore, India
Email: sanjay.vincent@cs.christuniversity.in
Email: chandra.j@christuniversity.in
*Corresponding author

Abstract: Various kinds of cancer are detected and diagnosed using histopathological analysis. Recent advances in whole slide scanner technology and the shift towards digitisation of whole slides have inspired the application of computational methods on histological data. Digital analysis of histopathological images has the potential to tackle issues accompanying conventional histological techniques, like the lack of objectivity and high variability. In this paper, we present a framework for the automated segmentation of nuclei from human histopathological whole slide images, and their classification using morphological and colour characteristics of the nuclei. The segmentation stage consists of two methods, thresholding and the watershed transform. The features of the segmented regions are recorded for the classification stage. Experimental results show that the knowledge from the selected features is capable of classifying a segmented object as a candidate nucleus and filtering out the incorrectly identified segments.

Keywords: histopathological images; whole slide images; digital image analysis; segmentation; nuclei; annotated; nuclear; computer-assisted diagnosis; machine learning; classifier; deep learning.

Reference to this paper should be made as follows: Vincent, S. and Chandra, J. (2022) 'Automated segmentation and classification of nuclei in histopathological images', *Int. J. Biomedical Engineering and Technology*, Vol. 38, No. 3, pp.249–266.

Biographical notes: Sanjay Vincent is in his final-year as Master of Science student in Computer Science at the CHRIST (Deemed to be University), Bangalore, India. He received his Bachelor of Science in Computer Science, Mathematics and Electronics from the Christ University, Bangalore in 2017. His research interests include image processing on medical images. In the future, he hopes to continue research in the area of health informatics for the computer-assisted diagnoses of diseases.

J. Chandra is an Associate Professor of Computer Science in the Department of Computer Science, CHRIST (Deemed to be University), Bangalore, India. She received her MPhil from the Vinayaka Missions University, India in 2009, and her PhD from the Hindustan University, India in 2016. She has

over 20 years of industry and teaching experience. Her research interests include data mining, big data, machine learning, data analytics, business intelligence, and image processing.

1 Introduction

There have been significant advances in the area of object recognition in various fields like the automotive and social media industries, involving image processing in its stages of infancy and extending up to computer vision, assisted by machine learning. With algorithms achieving quintessential values in terms of accuracy, it is only logical that the concepts of image processing and machine learning be extended to various other areas that can accommodate them. One such area is the medical sector, which is highly demanding, and a sector which is experiencing, although very slowly, a change in the ways of identification of diseases, and their diagnoses and treatments (Madabhushi and Lee, 2016). Although sectors like the automotive industry and the financial sector have seen much higher rates of technological advancements towards automation, it may be argued that the medical sector has not been able to match this pace due to several reasons including the regulations that pervade the industry and the higher accountability that is necessitated. In any case, computer-assisted diagnosis is the first step in the direction of automation of services in the medical sector in that it can greatly help doctors in the prompt identification of diseases and decision-making towards their diagnoses (Madabhushi, 2009).

A particular case is that of the disease of cancer, the impact of which is felt by millions across the globe, and which has a high mortality rate. It is also known that its early identification and treatment can save lives. There are various stages in the identification and diagnosis of cancer. Its classification requires the examination of the tissue or growth by a histopathologist. Histopathology is the microscopic study of diseased tissue. Human histopathological examination of tissues starts with surgery, biopsy, or autopsy. The most common biopsy techniques are fine needle aspiration, core needle biopsy, vacuum-assisted biopsy, and surgical biopsy. The tissue is removed from the body, and post expert dissection, is placed in a fixative. A pathologist grosses or cuts up the tissue by choosing the part most likely to yield an accurate diagnosis of the disease. The slide is then dyed using stains to differentiate between regions in the tissue. Hematoxylin and Eosin (H&E) is the most popular staining method where hematoxylin, a basic dye, stains the nuclei blue due to its affinity towards nucleic acids, and eosin, an acidic dye, stains the cytoplasm pink (Aswathy and Jagannath, 2017). The specimen is left to settle in order for features to become apparent. The slide is then viewed under a microscope, and the knowledge of the histopathologist aids in the understanding of the relationship between various features in the extracted portion, and finally, in the classification of cancer (Tosta et al., 2017).

Today, cancer is primarily diagnosed by pathological examination of the whole slide images of the tissue, a practice that is considered the ‘gold standard’ for the diagnoses of various diseases (Rorke and Balian, 2000). However, the visual evaluation of tissues is not perfect due to the differences in the knowledge level and experience of pathologists which give rise to inter-pathologist variability (Lozanski et al., 2013). Computer-assisted diagnosis powered by machine learning can help the experts in making prompt and

better decisions, given correct information about the data at hand. The advances in technology have allowed for the capture of whole slide images at varying magnification levels, ranging from 40X to 400X, and for their storage in digital formats. The most common type of such an image is the .svs format which is an embedded image type, containing several copies of the same image at different magnification levels in a single file. Notably, the sheer size of the image, which ranges between several tens of megabytes and a few gigabytes, is a roadblock for applying several image processing techniques. When using a software that can read the image in its raw format, i.e., as an array of intensity values at corresponding locations, the size of the array can easily extend up to 100,000 $\times$ 100,000 (pixels or values). Hence, applying various image processing techniques on the images is not an easy task. To give the readers a better perspective, one of the most commonly used unsupervised clustering methods, the k-means clustering- a colour deconvolution technique which separates the stains used on the tissue, is a computationally expensive process and a present-day computer with above average specs (32 GB RAM and an Intel i7 processor) cannot process the entire image. In fact, most of the times, the image is sampled down to a third of its highest magnification before applying such computationally expensive processing techniques. However, different magnification levels often result in different knowledge. Hence, sampling may be able to segment the image by colour of the stains used but the same cannot be applied when the regions of interest (ROI) are the cells or the nuclei in the histopathological image.

Segmentation is one of the most encountered challenges in histopathological image analysis, involving the automatic recognition of regions like nuclei and glandular structures in the image, which is key in various diagnoses (Gurcan et al., 2009). The differences between general images which have strong defining features that help in their automatic identification as opposed to pathological images which show repetitive patterns of minimum components (usually cells) make histological image analysis strenuous (Haggerty et al., 2014). The large size of the images is associated with a proportional number of ROIs and thus stimulates the application of numerous image processing techniques on the images in order to obtain varying levels of knowledge (Komura and Ishikawa, 2018). This is in contrast to radio-graphical images where the identification of cancer subtypes is practically impossible. However, histological image processing and analysis have several difficult issues because of the muddled structure of histopathological images. An explanation for this is the inconsistency of the images owing to variations in image quality, irregularities encountered in the image acquisition and subsequent stages (e.g., during compression), and most importantly the inevitable overlapping of the nuclei (He et al., 2012).

In this paper, we apply image processing techniques to segment the nuclei from histopathological whole slide images and record their features for their automated classification. We use the technique of thresholding and the watershed method to segment the ROIs. Finally, we record the features of the correctly segmented and the incorrectly segmented objects and compare the results of different machine learning and deep learning classification methods in the correct identification of nuclei.

The objective of this paper is to provide a framework that may be adopted to retrieve nuclear information from histopathological images. Most prevalent methods in practice that address the requirement include in them, thresholding, which is mainly empirical, and the watershed transform which comes with the disadvantage of over-segmentation. Empirical thresholding is not scalable because of the need for finding an optimal

threshold value of the intensities for each image manually. Images may be visually comparable but may vary vastly owing to a generic offset introduced during the stage of image acquisition, and several other variables like ambient light (in the case of light microscopes), and temperature. The watershed transform comes with the problem of over-segmentation where an individual nucleus may be recognised as multiple nuclei, therefore presenting an incorrect discovery that induces discrepancies into the dataset. The approach proposed in this paper aims to reduce the effect of said discrepancies in the data by excluding such records that do not fit in the nuclei spectrum. The experiments were conducted on histopathological images obtained from The Cancer Genome Atlas (TCGA), a National Cancer Institute (NCI) funded project begun in 2005 to help the cancer research community to improve the prevention, diagnosis, and treatment of cancer, and is publicly available (Grossman et al., 2016). Whole slide images of 24 different patients were downloaded for the purposes of this research. An automated thresholding algorithm that finds an optimal threshold improves the efficiency of digital histopathological image analysis. The proposed classification stage uses knowledge from the dataset of annotated nuclei to establish a benchmark for a nucleus. The training of a classifier on a large and accredited dataset can aid in improving the segmentation results, hence reducing inconsistencies.

The remaining of this paper is organised as follows. Section 2 discusses the related works. Section 3 discusses the materials and methods used to perform the experiment. Section 4 presents the experiment and the results obtained and Section 5 interprets the research findings. Finally, Section 6 summarises the contributions and discusses directions for future work.

2 Related work

This section focuses on previous works by experts on digital histopathological image analysis and its applications. It attempts to bring to the foreground the well-established problems in the field, the approaches used to solve them, and hence, the motivation for further study of the field.

Lu et al. (2013) proposed a hybrid technique consisting of a combination of mean-shift segmentation and local region recursive segmentation for the discovery of melanocytes in skin histopathological images. They use a local double ellipse descriptor to filter out nuclei that are falsely segmented, by monitoring the ellipticity of the nuclei.

Kothari et al. (2013) proposed a shape-based classification of renal tumour using Fourier shape descriptors to classify the tumours into its multi-class subtypes in place of morphological, textural, and wavelet-based features for which an average accuracy of 77% was achieved.

Lu and Mandal (2014) proposed a technique for the detection of mitotic cells in histopathological images by generating a discriminative image using ten different spectral band images and performing segmentation by Bayesian modelling and thresholding, following which a comprehensive feature set was extracted for candidate classification.

Zheng et al. (2018b) proposed a framework to automatically process whole slide images to recognise malignant regions which become the search query for content-based histopathological image retrieval (CBHIR). Regions with similar content are obtained through retrieval, which aids the pathologist in a more reasonable diagnosis. They

propose the use of a designed neural network in place of scale invariant feature transform (SIFT) descriptors, following which the features are binarised for efficient retrieval from a database.

Zheng et al. (2018a) proposed a novel framework for CBHIR system by modelling whole slide images into binary-coded matrices. The approach optimises precision, computation costs, and size variability by binarising the query slide and basing the retrieval on efficient table lookups along with a size monitored ranking system.

To overcome the incapability of the bag of features and dictionary learning representations based on randomly extracted fixed size patches of 8×8 pixels in completely describing relationships between pathological elements available in the image, Romo-Bucheli et al. (2015) proposed the use of raw feature descriptors extracted at different scales following which they are concatenated into a single vector, using maximally stable extreme regions (MSER) for nuclei candidate detection.

Vu et al. (2016) proposed an automatic feature discovery framework for histopathological image classification through discriminative class specific dictionaries. The proposed discriminative feature-oriented dictionary learning (DFDL), which learns class specific dictionaries and emphasises inter-class differences while minimising intra-class differences, resulted in enhanced classification accuracies when applied to three datasets, intra-ductal breast lesions, brain cancer images, and tissue images from kidney, lung, and spleen.

Qu et al. (2014) generated morphological feature vectors from segments obtained by pixel-wise support vector machine SVM classifier along with marker-controlled watershed method in order to predict the survival of patients.

Niazi et al. (2017) proposed a method of grading prostate cancer using visually-meaningful feature sets that contain both luminal and architectural features, with the aim of presenting clinically interpretable features to pathologists. This was done by extracting the lumen, nuclei, and stroma post texture-based segmentation, from whole slide images which were used to classify test images from a Gleason-scored dataset.

Spanhol et al. (2016) asserted that although increasing research is being carried out in the area of histological image analysis, experiments being performed on non-standardised and relatively miniature datasets make it difficult to compare the results obtained, however remarkable, hence making it unviable to introduce a standardised computer-assisted classification system into practice with any confidence. A dataset of breast cancer histopathological images including both benign and malignant images was created and was made publicly available to address this issue. Automated classification of the images yielded accuracy rates ranging from 80% to 85% using 6 feature vectors, showing that much room is left for improvement.

Pang et al. (2010) proposed the application of convolutional neural networks, which eliminates the problem of domain specificity, trained using gradient descent techniques for cell nuclei segmentation in histopathological images. The method was proved to be better in performance when compared to two popular classification methods, SVM and FLDA.

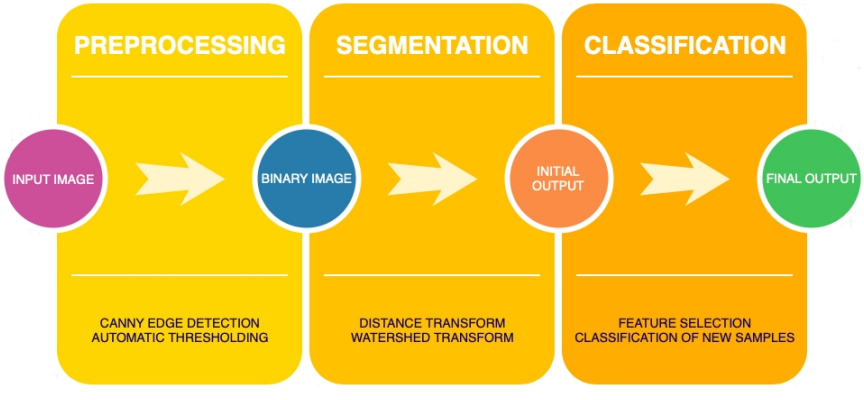
The literature demonstrates the idea of segmentation of various ROIs including nuclei and melanocytes, automatic feature discovery, and the classification of ROIs in order to establish comprehensive frameworks in various areas of computer-assisted diagnosis. The applications of the extraction of ROIs include content-based image retrieval system, apart from using the knowledge from the features in the direct classification of malignancy. The most commonly chosen feature vectors include

morphological, texture, and colour characteristics of the desired ROIs from the result of the segmentation methods. Thresholding and watershed methods are proven to be capable methods in the area of segmentation. The literature also demonstrates the use of masks of fixed size to capture the ROIs which later become the input to a CBHIR system. However, a fixed size mask will result in the unnecessary inclusion of details (noise) along with the true ROIs. The search query, as well as the database, may be optimised to use lesser space and also to improve efficiency if the precise characteristics of the ROIs can be captured. In this paper, we propose a framework to segment nuclei from histopathological images by using a hybrid method involving thresholding and the watershed transform method along with a classification model in order to filter out incorrectly identified segments.

3 Materials and methods

The following section discusses the source of the data used in the experiments conducted, and the methods used on the data. We propose the inclusion of a classification stage post the process of segmentation for improved segmentation results. This section discusses in detail, the methods used in this paper, the rationale for using the methods, and how the methods overcome the issues that necessitated their application on the data. We have a sequential framework in place for the extraction of nuclei from the histopathological images and the flowchart is illustrated in Figure 1. The framework consists of three stages. The first and the second stage in the framework contribute to the segmentation of the ROI, with the first stage involving a global thresholding method and the second, the watershed transform method. The third stage is the classification stage where a combination of morphological and colour features of the ROIs is used to classify the candidate nuclei.

Figure 1 Proposed framework (see online version for colours)



3.1 Dataset

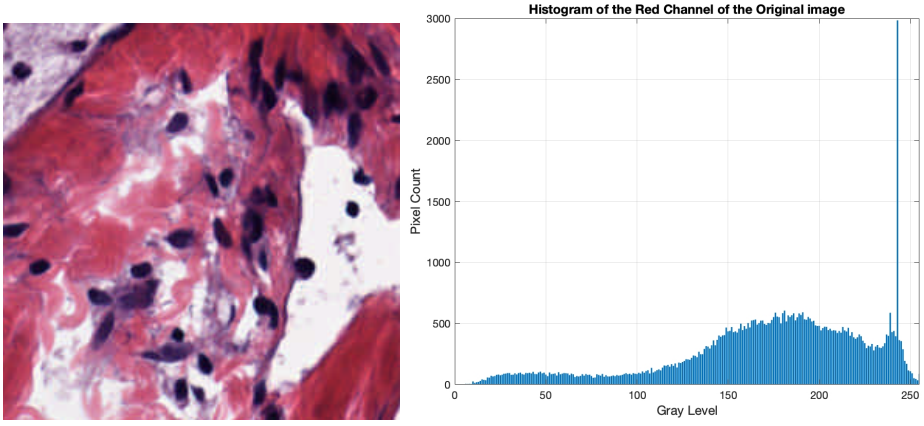
The dataset is downloaded from National Cancer Institute (2018). The histopathological whole slide images of 24 thyroid cancer patients are used for the purposes of this

research. Each image contains over 40,000 × 40,000 values, hence a block processing method is used to break down an image into blocks of 256 × 256 values each. To account for the irregularity of stains in the images, two blocks with varying levels of stains are used from each whole slide image.

3.2 Segmentation

Figure 2 shows a histopathological image and the histogram of the red channel of the image. The skewness is typical in the case of a majority of histopathological images due to the fact that the nuclei are small, dark oval-shaped blob-like structures that exist over masses of lighter supporting tissue. In many cases, the whole slide image has several concentrations of this supporting structure as well as spaces that are unfilled. Due to the staining approaches used, the nuclei being dark blue, occupy the lower end of the available 256 bins in a histogram of the image in grayscale, whereas the majority of the pixels occupy the higher end of the bin locations. The Otsu's method is one of the most popular thresholding methods applicable to several kinds of images (Al-Kubati et al., 2012). It is most effective when the histogram of the image is bimodal. However, it is observed that the Otsu's method fails for images with skewed histograms (Yang et al., 2012).

Figure 2 An instance of a histopathological image and its red-channel histogram (see online version for colours)



(a) (b)

3.2.1 Canny edge detection auto thresholding

The Canny edge detector can be used to obtain the global threshold value for images with skewness. The Canny edge detector reduces noise from the image using a smoothing Gaussian filter before finding the edges (Canny, 1986). The equation for a Gaussian filter kernel of size $(2m+1) \times (2m+1)$ is given by:

$$H = \frac{1}{2\pi\sigma^2} e^{\left(-\frac{(i-(m+1))^2+(j-(m+1))^2}{2\sigma^2}\right)}; 1 \leq i, j \leq (2m+1) \quad (1)$$

An edge detection operator (such as Roberts, Prewitt or Sobel) returns the local gradient and the edge direction at each point in the image.

$$G = \sqrt{{G_x}^2 + {G_y}^2} \tag{2}$$

$$\theta = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{3}$$

If A is defined as the source image, G_x and G_y are two images which at each point contain the horizontal and vertical derivative approximations. The corresponding computations of G_x and G_y of the Sobel operator are given below.

$$G_x = \begin{pmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{pmatrix} * A \text{ and } G_y = \begin{pmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} * A \tag{4}$$

The algorithm then performs a non-maximal suppression by checking at every pixel if the pixel is a local maximum in its neighbourhood in the direction of the gradient. The resulting ridge pixels constituting the edges are thresholded using two thresholds T_1 and T_2, where $T_1 < T_2$ (Kaur et al., 2012). Pixels greater than T_2 constitute the strong pixels while those bound by the range given by T_1 and T_2 constitute the weak edge pixels. These edge points are linked by incorporating the weak points to the strong points subject to a weak point being 8-connected with a strong point.

3.2.2 Watershed transform

The watershed is a transformation defined on an image which identifies 'catchment basins' or 'watershed ridge lines' in the image. The image is viewed as a surface where lighter pixels represent higher elevations and darker pixels represent lower elevations (Meyer, 1994). Such a viewing results in a topographic map of the image. The watershed transform identifies regions of high-intensity gradients, called watersheds, that divide neighbouring local minima, called basins. The idea of the watershed transform is that if the surface is flooded from its minima and if the merging of the waters from different sources is prevented, the image is partitioned into two classes, the catchment basins, and the watershed ridge lines (Salman, 2006). A distance transform of the binary of the image is used to assist the watershed transform in identifying the focal points of each object. The distance transform labels each pixel of the image with the distance to the nearest obstacle pixel. The Euclidean distance transform which uses the Euclidean distance metric is used in this step. The Euclidean metric is the straight line distance between two points in Euclidean space. The distance between points $p(x_1, y_1)$ and $q(x_2, y_2)$ is defined as:

$$d(p, q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

The watershed transform then segments the image on the basis of the focal points discovered using 8-connected neighbourhoods. The aim of using the watershed transform in this stage is to segment larger objects, identified because of overlapped nuclei, and/or irregular shape surrounding the nucleus due to excess amount of stain

in the region, from the thresholding method. However, the watershed transform is also known to be prone to over-segmentation. In the context of histopathological image processing in the segmentation of nuclei, over-segmentation implies the breaking up of an individual nucleus into two or more smaller objects, when such segmentation is not necessary.

3.3 Classification

Statistical classification is a machine learning problem that is used to categorise new probabilistic observations into corresponding categories. A binary classification algorithm classifies the elements of a set into two classes or groups. The performance of the most commonly used statistical binary classifiers is assessed for the data collected post the feature selection step in order to classify a segmented object as 'nucleus' or 'not-nucleus'.

Naive Bayes classifiers are a family of probabilistic classifiers based on Bayes' theorem which is stated as:

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{6}$$

where $P(\theta|\mathbf{D})$ is the likelihood of event θ occurring given that $\mathbf{D}$ is true, $P(\mathbf{D}|\theta)$is the likelihood of event $\mathbf{D}$ occurring given that θ is true, and $P(\mathbf{D})$ and $P(\theta)$ are the probabilities of observing $\mathbf{D}$ and θ independently of each other.

The decision tree is a type of supervised learning algorithm that splits the population or sample into homogeneous sets based on the most significant differentiator in input variables.

Random decision forest is an ensemble learning method that functions by building several decision trees at the time of training. The output class is determined by the mean prediction of individual trees.

The linear discriminant analysis (LDA) approaches classification by an estimation of the probability that a new set of inputs belongs to each class. It assumes the data to be Gaussian and each attribute to have the same variance.

Support vector machines (SVMs) are supervised training models that belong to non-probabilistic binary linear classifiers. The samples are represented as points in space for which a hyperplane is constructed to segregate categories that are separated by a clear gap.

Deep learning models have seen significant success in classification due to their ability to automatically select features and to 'learn' from large databases. Artificial neural networks (ANN) are information processing paradigms that are loosely modelled after the structure of the neurons of the mammalian cerebral cortex. The centre of the neuron is known as the nucleus which is associated with other nuclei by a synaptic connection of the dendrites and the axon, where the dendrites carry information to the nucleus, and the axons carry information away from the nucleus. The general mathematical definition of such a model of the neuron is given by:

$$y(x) = g\left(\sum_{i=0}^{n} w_i + x_i\right) \tag{7}$$

where $y(x)$ is the single output axon of the neuron x, $x = (x_0...x_n)$, a neuron with n input dendrites where $w = (w_0...w_n)$ determine the weights of corresponding inputs and g is an activation function that decides the strength of the output of the neuron. The most commonly used activation functions are the threshold, sigmoid, and hyperbolic tangent. Equations 8 to 10 describe the three activation functions, threshold, sigmoid, and hyperbolic tangent respectively.

$$f(x) = \begin{cases} 1 & if \quad x + t > 0 \\ 0 & if \quad x + t \leq 0 \end{cases} \tag{8}$$

$$g(x) = \frac{1}{1 + e^{-2s(x+t)}} \tag{9}$$

$$g(x) = tan^{-1}(s(x + t)) \tag{10}$$

A dataset containing information pertaining to the shape and colour of each object is created for the purpose of classification. A comprehensive list of features is selected in order to classify an object into the category of 'nucleus' or 'not-nucleus' and the performance of various classifiers is evaluated. The shape descriptors may be identified from a binary image using the RegionProps feature available in the image processing toolbox provided by MATLAB (The Mathworks Inc., 2018).

3.3.1 Feature selection

A combination of 13 shape descriptors is used in order to capture the features of each object. To extract the colour constituents of these objects, the pixels constituting an object are identified. The illumination levels of the three channels of the original RGB image are then extracted by traversing the obtained pixel locations in the three colour channels. The mean, median, mode, standard deviation, and variance of the values obtained from each channel are recorded. Hence, a comprehensive feature vector of 28 variables is used to capture an object, and the feature vectors are stated below:

1 Area – The total number of pixels in the object.

2 ConvexArea – Convex area is the total number of pixels in the convex image with all pixels within the hull filled in. The convex hull is the smallest convex polygon that can contain the region.

3 Eccentricity – The ratio of the number of pixels between the foci of the ellipse and its major axis length.

4 EquivDiameter – The diameter of a circle with the same area as the region.

5 EulerNumber – The number of objects in the region minus the number of holes in those objects.

6 Extent – The proportion of the pixels in the bounding box that are also in the region. The bounding box is the smallest rectangle containing the region.

7 FilledArea – The number of on pixels in the logical image of the same size as the bounding box of the region.

8 MajorAxisLength – The number of pixels in the major axis of the ellipse that has the same second-moments as the region.

9 MinorAxisLength – The number of pixels in the minor axis of the ellipse that has the same second-moments as the region.

10 Orientation – The angle (in degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.

11 Perimeter – The total number of pixels in boundary of the region.

12 Roundness – An estimate of how close the object is to a circle. For a circle this value is 1, and decreases as the circularity decreases.

13 Solidity – The proportion of the pixels in the convex hull that are also in the region.

In addition to these morphological features, five statistical values describing the colour information of the object are registered for each channel of the original RGB image.

1 Mean – The average intensity value of intensities in the region defined by the area of the object.

2 Mode – The value of intensity that occurs the most in the region.

3 Median – The intensity value that partitions the list of intensities in the region into two sets, one containing values higher than it and the other containing values lower than it.

4 Variance – Sum of squared differences from the mean divided by number of pixels in the object.

5 Standard deviation – The squared root of variance.

3.3.2 Classification metrics

A dataset is created using the 28 features mentioned above with the ratio of 'nucleus' and 'not-nucleus' being 60:40. This is initially split into 65% training and 35% test data. Six state-of-the-art binary classifiers are used for the classification stage. Four metrics, precision (positive prediction value), recall [true-positive-rate (TPR)], accuracy (proportion of correct predictions), and F1 Score (harmonic mean of precision and recall) are used to evaluate the performance of the classification. The definitions of the metrics are given below:

$$Recall(r) = \frac{TP}{TP + FN} \tag{11}$$

$$Accuracy(a) = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$Precision(p) = \frac{TP}{TP + FP} \tag{13}$$

$$F\_Score = \frac{2 * p * r}{p + r} \tag{14}$$

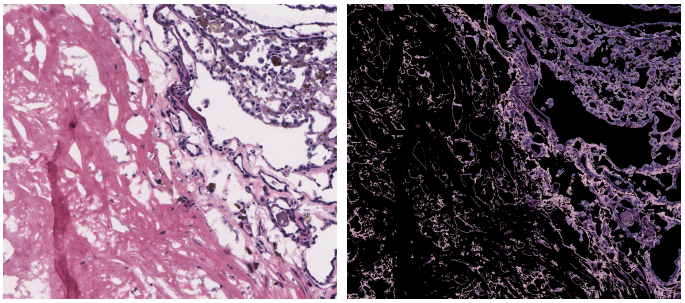
4 Results

This section describes the results of the application of the methods in detail. The section is further divided into three subsections, namely, thresholding, watershed transform, and classification, which are performed in the same order as they are described. Each subsection describes the results of the application of the corresponding method.

4.1 Thresholding

Several works of literature suggest the use of a preprocessing step involving the deconvolution of the colours before applying a threshold. The most commonly used colour-based segmentation include the k-means clustering algorithm which uses the CIE L*a*b* colour space to automatically segment colours. In a typical H&E stained image, the algorithm results in three images or clusters, the third cluster being the one containing just the nuclei. However, this step is found to result in a loss of detail with respect to the original image. Hence, we avoid this step in order to preserve the original colour detail and sharpness of the images. The original image and the resulting image (hematoxylin component) of a deconvolution technique that separates the hematoxylin component from the eosin component are shown in Figure 3. Visible degradation in the image is evident from the appearance of light spots over the nuclei which contribute to weaker edges and hence the loss of nuclear shape. The red component of the RGB image is more discriminative and provides higher contrast between the hematoxylin stained nuclei and the eosin stained stroma. Hence, we use this channel over the generalised RGB to grey conversion which computes a weighted average of the three channels.

Figure 3 (a) H&E stained histopathological image (b) Colour decovolved image showing the hematoxylin component (see online version for colours)

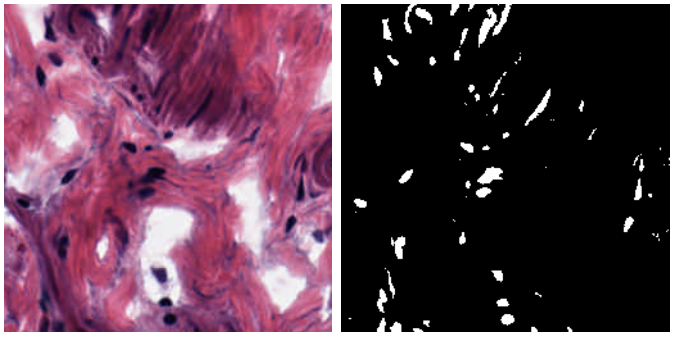


(a) (b)

The Canny edge detection algorithm is applied to the red channel of the original image. The result of this step is a binary image containing the edges and a single dimensional array consisting of two values, the low threshold value, and the high threshold value. A combination of the values is used as a global threshold value to binarise the entire image. The mean of the values with an uncertainty of 50% is found to be a good estimator for such a threshold. Figure 4(a) shows an instance of a histopathological image. Subsequent images depict the result of various methods applied to this image.

Figure 4(b) shows the resultant image of the binarisation stage. The threshold captures the objects that occupy the lower end of the 256 grey level bins. However, this binary image cannot be used as the ultimate descriptor of the nuclei because of two reasons. The method captures the overlapping nuclei in the image as a single object with a larger and non-oval shape. The variation in the amount of stain at particular regions results in dark patches surrounding the nuclei, thus registering shapes that are not representative of nuclei.

Figure 4 (a) H&E stained histopathological image (b) Binary image obtained after threhsolding (see online version for colours)



(a) (b)

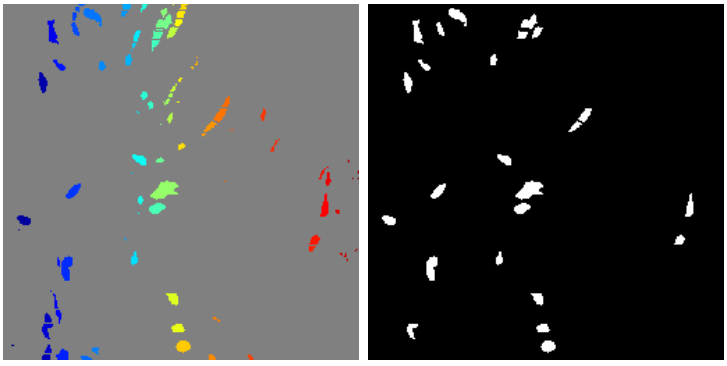
4.2 Watershed transform

The watershed transform is used to improve the segmentation result. The complement of the binary image obtained from the step described above is used to find the distance transform of the image. This is performed so that pixels that do not belong to the objects are forced to be at an infinite distance from the pixels constituting the object. The watershed transform is then applied to obtain the final result of the segmentation process. Figure 5 shows the resulting RGB label image and the binary image after the application of the watershed transform. Although the watershed transformation is able to further segment objects, it comes with a problem of over-segmentation.

4.3 Classification

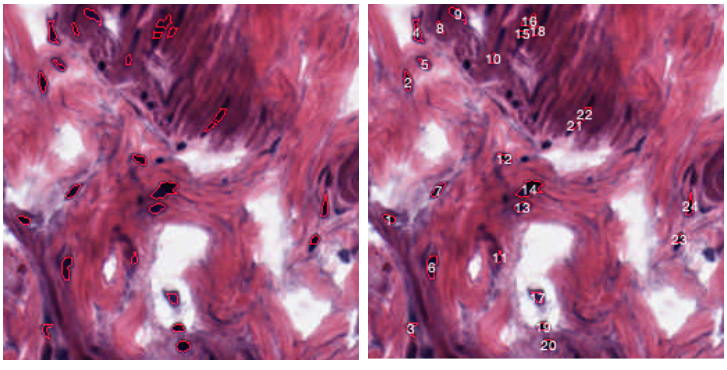
The result of the segmentation method is a binary image which captures the nuclei in the histopathological image as white blobs on a black background. We use the term object to denote individual blobs in the image. In order to choose the objects identified that truly denote nuclei, the perimeter of each object is identified from the binary image and is overlaid onto the original RGB image. Each of these objects is then numbered for identification. The objects that capture the entire nucleus are labeled as ‘nucleus’ and the ones that do not capture the entire nucleus and instead capture multiple overlapped nuclei are labeled as ‘not-nucleus’. Figure 6 shows two images, one with the perimeter of the object outlined on the original image and the other where the identified objects are numbered.

Figure 5 (a) Label colour image (b) Binary image obtained after watershed transform (see online version for colours)



(a) (b)

Figure 6 (a) Objects oulined on original image (b) Objects numbered for candidate selection (see online version for colours)



(a) (b)

Table 1 Comparison of classification results

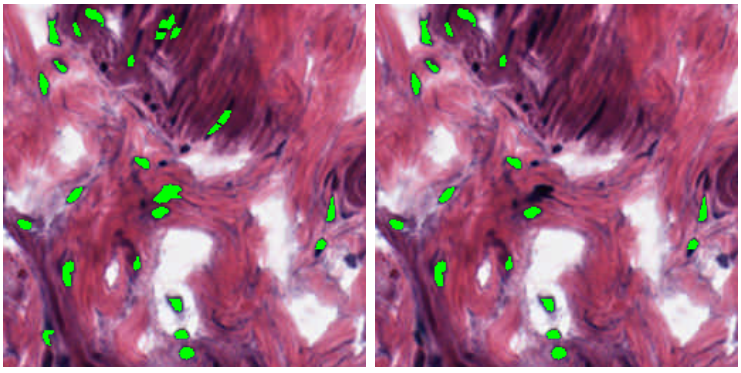
Method	*Accuracy*	*Precision*	*Recall*	P*F1_Score*
ANN	0.9171	0.8571	0.8889	0.8727
LDA	0.8740	0.8962	0.8284	0.8637
Bayes	0.8654	0.8505	0.8404	0.8641
Random Forest	0.8646	0.8526	0.8375	0.8618
Decision Tree	0.8073	0.7903	0.7703	0.8033
SVM	0.6962	0.3481	0.5000	0.5719

Table 1 shows the comparison of the classification results on the dataset using the metrics mentioned above averaged over 20 iterations. The ANN performed well with the highest accuracy, which is a desirable feature considering the distribution of classes

in the data is in the ratio of 60:40. The LDA provided better precision than the ANN. However, the SVM performed rather poorly for the particular data.

The output of the classifier is necessary for filtering out segments that do not capture the nuclei well. When a database of sufficient size is created, the classifier can classify a segment as 'nucleus' or 'not-nucleus' on the basis of its characteristics. Figure 7(a) shows the objects identified by the segmentation method and Figure 7(b) shows the resulting objects after the classification step. The area of the nuclei is made green for easy identification. It can be seen that the classifier filters out certain segments that cannot qualify as candidate nuclei based on the training knowledge.

Figure 7 (a) Initial output (b) Output post classification (see online version for colours)



(a) (b)

5 Discussion

Digital histopathological image analysis is a field that has room for continued development and one that motivates further study of the field. It is established that Otsu's thresholding method fails when the image in question has a skewed histogram. In this paper, we use the threshold found by the Canny edge detection, and the watershed transform to overcome the problem. The method demonstrated success on a wide array of histopathological images. Region growing may be used to enhance the results if the threshold does not fully capture the nuclei, by using the objects discovered as seed points. We included a stage of classification which improves the segmentation results. The information of each object identified by the segmentation methods is stored for this purpose. While the abundance of nuclei in histopathological images encourages the application of various methods for knowledge discovery, the same also increases bandwidth for errors during the segmentation stage. However, this also means that a significantly large dataset can be formed using comparatively fewer images making it suitable for the training of a classifier. The classification results show that a classifier can be trained to discover a benchmark for a nucleus and such a benchmark obtained from an accredited dataset of annotated nuclei can significantly improve segmentation results.

6 Conclusions

Histopathological analysis of tissue samples is a time consuming and labourious task that is critical in the detection and diagnoses of various diseases. Computer-assisted diagnosis makes it easier for experts to arrive at conclusions from the data available. This paper presents a framework for the automated segmentation of nuclei using a hybrid methodology involving segmentation and classification in order to filter out falsely classified nuclei and achieved good accuracy results of 85% $\pm$ 6%. The performance of the classifier depends on the size of the data and the quality of the annotated nuclei used for training. The quantitative analysis of nuclei can help experts in the prompt diagnoses of diseases. Given enough information that accurately describes the nuclei, the system can be adopted for the filtering of non-nuclei from the segmented objects. The segmented nuclei may be exploited for usage in CBHIR systems in diagnoses, education, and research. Future work will focus on classification of malignancy in diseases which require the analysis of the characteristics of the nuclei for the diagnoses, and optimisation in CBHIR systems.

References

Al-Kubati, A.A.M., Saif, J. and Taher, M.A. (2012) *Evaluation of Canny and Otsu Image Segmentation*.

Aswathy, M. and Jagannath, M. (2017) 'Detection of breast cancer on digital histopathology images: present status and future possibilities', *Informatics in Medicine Unlocked*, Vol. 8, pp.74–79 [online] http://www.sciencedirect.com/science/article/pii/S235291481630034X.

Canny, J.F. (1986) 'A computational approach to edge detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, pp.679–698.

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A. and Staudt, L.M. (2016) 'Toward a shared vision for cancer genomic data', *New England Journal of Medicine*, Vol. 375, No. 12, pp.1109–1112, PMID: 27653561 [online] https://doi.org/10.1056/NEJMp1607591.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M. and Yener, B. (2009) 'Histopathological image analysis: a review', *IEEE Reviews in Biomedical Engineering*, Vol. 2, pp.147–171 [online] https://www.ncbi.nlm.nih.gov/pubmed/20671804.

Haggerty, J.M., Wang, X.N., Dickinson, A., O'Malley, C.J. and Martin, E.B. (2014) 'Segmentation of epidermal tissue with histopathological damage in images of haematoxylin and eosin stained human skin', *BMC Medical Imaging*, Vol. 14, No. 1, p.7 [online] https://doi.org/10.1186/1471-2342-14-7.

He, L., Long, L.R., Antani, S. and Thoma, G.R. (2012) 'Histology image analysis for carcinoma detection and grading', *Computer Methods and Programs in Biomedicine*, Vol. 107, No. 3, pp.538–556 [online] https://www.ncbi.nlm.nih.gov/pubmed/22436890.

Kaur, J., Agrawal, S. and Vig, R. (2012) 'Article: a comparative analysis of thresholding and edge detection segmentation techniques', *International Journal of Computer Applications*, Vol. 39, No. 15, pp.29–34.

Komura, D. and Ishikawa, S. (2018) 'Machine learning methods for histopathological image analysis', *Computational and Structural Biotechnology Journal*, Vol. 16, pp.34–42 [online] http://www.sciencedirect.com/science/article/pii/S2001037017300867.

Kothari, S., Phan, J.H., Young, A.N. and Wang, M.D. (2013) 'Histological image classification using biologically interpretable shape-based features', in *BMC Medical Imaging*.

Lozanski, G., Pennell, M., Shana'ah, A., Zhao, W., Gewirtz, A., Racke, F., Hsi, E., Simpson, S., Mosse, C., Alam, S., Swierczynski, S., Hasserjian, R.P. and Gurcan, M.N. (2013) 'Inter-reader variability in follicular lymphoma grading: conventional and digital reading', *Journal of Pathology Informatics*, Vol. 4, No. 30, pp.30–30.

Lu, C. and Mandal, M.K. (2014) 'Toward automatic mitotic cell detection and segmentation in multispectral histopathological images', *IEEE Journal of Biomedical and Health Informatics*, Vol. 18, pp.594–605.

Lu, C., Mahmood, M., Jha, N. and Mandal, M.K. (2013) 'Automated segmentation of the melanocytes in skin histopathological images', *IEEE Journal of Biomedical and Health Informatics*, Vol. 17, pp.284–296.

Madabhushi, A. and Lee, G. (2016) 'Image analysis and machine learning in digital pathology: challenges and opportunities', *Medical Image Analysis, 20th Anniversary of the Medical Image Analysis Journal (MedIA)*, Vol. 33, pp.170–175 [online] http://www.sciencedirect.com/science/article/pii/S1361841516301141.

Madabhushi, A. (2009) 'Digital pathology image analysis: opportunities and challenges', *Imaging in Medicine*, Vol. 1, No. 1, pp.7–10 [online] https://www.ncbi.nlm.nih.gov/pubmed/30147749.

Meyer, F. (1994) 'Topographic distance and watershed lines', *Signal Processing*, Vol. 38, No. 1, pp.113–125.

National Cancer Institute (2018) *The Cancer Genome Atlas (TCGA) Database* [online] http://cancergenome.nih.gov.

Niazi, M.K.K., Yao, K., Zynger, D.L., Clinton, S.K., Chen, J., Koyutürk, M., LaFramboise, T. and Gurcan, M. (2017) 'Visually meaningful histopathological features for automatic grading of prostate cancer', *IEEE Journal of Biomedical and Health Informatics*, Vol. 21, No. 4, pp.1027–1038.

Pang, B., Zhang, Y., Chen, Q., Gao, Z., Peng, Q. and You, X. (2010) 'Cell nucleus segmentation in color histopathological imagery using convolutional networks', *2010 Chinese Conference on Pattern Recognition (CCPR)*, pp.1–5.

Qu, A., Chen, J., Wang, L., Yuan, J., Yang, F., Xiang, Q., Maskey, N., Yang, G., Liu, J. and Li, Y. (2014) 'Two-step segmentation of hematoxylin-eosin stained histopathological images for prognosis of breast cancer', in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.218–223.

Romo-Bucheli, D., Moncayo, R., Cruz-Roa, A. and Romero, E. (2015) 'Identifying histological concepts on basal cell carcinoma images using nuclei based sampling and multi-scale descriptors', *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp.1008–1011.

Rorke and Balian, L. (2000) 'Pathologic diagnosis as the gold standard', *Cancer*, Vol. 79, No. 4, pp.665–667 [online] https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0142%2819970215%2979%3A4%3C665%3A%3AAID-CNCR1%3E3.0.CO%3B2-D.

Salman, N. (2006) 'Image segmentation based on watershed and edge detection techniques', *Int. Arab J. Inf. Technol.*, Vol. 3, pp.104–110.

Spanhol, F.A., de Oliveira, L.E.S., Petitjean, C. and Heutte, L. (2016) 'A dataset for breast cancer histopathological image classification', *IEEE Transactions on Biomedical Engineering*, Vol. 63, pp.1455–1462.

The Mathworks Inc. (2018) *MATLAB*, Natick, Massachusetts, 9.4.0.813654 (R2018a).

Tosta, T.A.A., Neves, L.A. and do Nascimento, M.Z. (2017) 'Segmentation methods of h&e-stained histological images of lymphoma: a review', *Informatics in Medicine Unlocked*, Vol. 9, pp.35–43 [online] http://www.sciencedirect.com/science/article/pii/S2352914817300175.

Vu, T.H., Mousavi, H.S., Monga, V., Rao, G. and Rao, U.K.A. (2016) 'Histopathological image classification using discriminative feature-oriented dictionary learning', *IEEE Transactions on Medical Imaging*, Vol. 35, No. 3, pp.738–751.

Yang, X., Shen, X., Long, J. and Chen, H. (2012) 'An improved median-based otsu image thresholding algorithm', *AASRI Procedia, Conference on Modelling, Identification and Control*, Vol. 3, pp.468–473 [online] http://www.sciencedirect.com/science/article/pii/S2212671612002338.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H. and Zhao, Y. (2018a) 'Histopathological whole slide image analysis using context-based CBIR', *IEEE Transactions on Medical Imaging*, Vol. 37, pp.1641–1652.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Ma, Y., Shi, H. and Zhao, Y. (2018b) 'Size-scalable content-based histopathological image retrieval from database that consists of WSIS', *IEEE Journal of Biomedical and Health Informatics*, Vol. 22, pp.1278–1287.