An automated pattern recognition system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies

Simon Hodgson [a] Robert F. Harrison [a,*] Simon S. Cross [b]

[a] *Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK.*

[b] *Academic Unit of Pathology, Division of Genomic Medicine, School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield S10 2RX, UK.*

Abstract

This paper presents an automated system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies. Initially, features are extracted from colour-corrected biopsy images at positions of interest identified by adaptive thresholding and clump decomposition. A sequential floating search method and principal component analysis are used to reduce dimensionality. Manually annotated training images allow supervised training. The performance of Gaussian parametric and mixture models is compared when used to classify regions as either inflammatory or healthy. The system is optimized using a response surface method that maximises the area under the receiver operating characteristic curve. This system is then tested on images previously ranked by a number of observers with varying levels of expertise. These results are compared to the automated system using Spearman rank correlation. Results show that this system can rank 15 test images, with varying degrees of inflammation, in strong agreement with five expert pathologists.

Key words: Liver biopsy analysis. Feature extraction. Pattern recognition. Bayesian Decision Theory. Gaussian Mixture Models. Sequential Forward Floating Search.

* Corresponding author. Tel.: +44 (0)114 222 5139 Fax.: +44 (0)114 222 5661.

Email addresses: cop01sh@shef.ac.uk (Simon Hodgson), r.f.harrison@shef.ac.uk (Robert F. Harrison), s.s.cross@shef.ac.uk (Simon S. Cross).

1 Introduction

The World Health Organisation (WHO) estimates that 170 million people, 3% of the world's population, are currently infected with the hepatitis C virus (HCV) [1]. This virus is usually transmitted by exposure to the blood or blood products of an infected person. In the majority of cases infected people do not develop symptoms for a number of years, leaving them totally unaware of their situation [2]. Liver damage is not caused by the virus itself but by the body's immune response to the attack. This damage can be extremely serious, resulting in liver failure and death of the patient. The current treatment for HCV, according to the UK clinical guidelines, is with a combination therapy of two drugs, Interferon–alpha and Ribavirin [3]. A major factor in prescribing combination therapy is that both drugs produce side effects in most people [3]. The cost of combination therapy is between £3000 and £12000 per patient per year [3]. It is generally thought that treating patients with expensive drugs with potentially serious side-effects may be inappropriate unless there is evidence of disease activity [1]. A liver biopsy is currently the only method available to assess HCV activity. The biopsy, involves removing a small core of tissue, approximately 15mm in length by 2-3mm in diameter. This core is then processed in paraffin wax, cut into slices along its length and stained. At this stage a trained histopathologist will examine the samples under a light microscope and use his/her experience, combined with a detailed definition, to assess the level of damage. The damage can normally be categorized into two types and it is common to assign a numerical score relative to the level of damage for each type. One of the most widely used scoring methods is the Ishak system [5], which can be summarised as

(1) Inflammation: assigned a necroinflammatory [2] (activity) score from 0–18.
(2) Scarring: assigned a fibrosis [3] (stage) score in the range 0–6.

Scarring is an indication of long-term disease activity and as a result remains relatively constant. For this reason, it is the assessment of inflammation that is normally the determining factor for a patient to receive treatment. The scoring process is time consuming and requires highly experienced and qualified personnel. Studies have shown that it is often difficult for observers to agree on activity and stage scores when evaluating the same samples, and it is common for the same observer to assign different scores at a later date [6]. This inter- and intra-observer variability has been studied in depth in [6], which found that observer agreement was far better for the assessment of fibrosis (stage) than for inflammation (activity). This finding, together with emphasis

[1] HCV is particularly likely to be associated with chronic disease[4]; for 20% of people with this form, liver disease will slowly progress to cirrhosis of the liver during the first 10 to 20 years.

[2] Cell death caused by the body's inflammatory response.

[3] The formation of fibrous tissue.

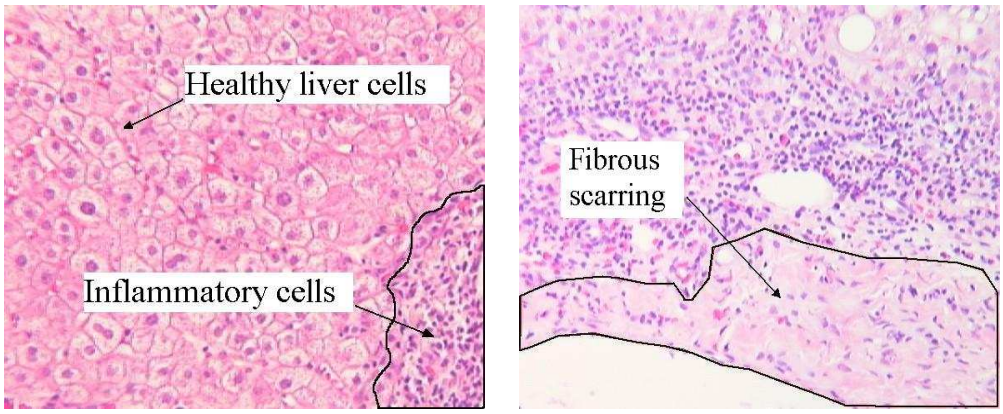
on inflammatory activity when considering treatment, stresses the urgent need for improved reliability in the assessment of inflammation. It is proposed that an automated system could be developed using image processing and pattern recognition techniques to assess, systematically, the level of inflammation in liver biopsies.

There are many different approaches to image analysis of histological sections. In this study we use a systems-based approach which does not seek exactly to reproduce the presumed segmentation and cell-labelling that occurs during a human assessment of a slide but does correlate with a recognized feature in the slide (the amount of lymphocytic inflammation). This avoids, for instance, the substantial computational burden incurred by popular techniques such as "active contours" [7]. Other studies have used different approaches. Some have used immunohistochemistry to label a specific element in a slide which has some relationship to the disease process. A number of studies have used an antibody against proliferating cell nuclear antigen to assess the proliferative response of hepatocytes to the damage caused by the hepatitis virus [8–10]. Other investigators have used histochemical stains for fibrotic tissue to measure the amount of fibrosis in hepatitic liver biopsies [11,12]. These different methodologies produce different information about the biopsies and are likely to be complementary to each other but the aim in this initial study was to automate currently accepted procedures using conventional haematoxylin and eosin stained paraffin sections that are readily available in histopathology laboratories worldwide. To introduce new methods would require a huge amount of development work to validate them sufficiently as a test that could be used in therapeutic decisions in the clinical context.

We present research on the design, optimization and testing of an automated pattern recognition system, to quantify, reliably, the amount of liver inflammation. Initially, the liver biopsy is examined in more detail with particular consideration given to the colour variation in biopsy samples. Next, a new pattern recognition system is presented and a method of system optimization is then outlined. Finally, the optimized system is tested using images previously evaluated by human observers.

2 Liver Biopsy Interpretation

This section introduces the image characteristics of an HCV-infected liver biopsy and discusses the colour variation between biopsy samples. To understand this investigation in more detail it is first necessary to consider the histopathological elements of a normal liver biopsy and the different forms of damage. A microscopic view of a standard liver biopsy from a healthy person shows liver cells (hepatocytes) forming interconnecting walls created by the



(a) (b)

Fig. 1. Microscopic (10× objective) view of a liver biopsy. (a) Healthy and inflammatory liver cells (b) A region of fibrous scarring.

close contact of cell membranes [13], as shown in figure 1(a). The nucleus is the dark mass located at the centre of each cell. The array of hepatocytes is only interrupted by other structural elements of the liver, such as portal tracts [4], hepatic veins [5] and bile ducts [6] (not shown). The damage caused by HCV alters this structure and can normally be categorized into two types:

(1) Inflammation - Cell death (necrosis), caused by the viral attack, evokes an inflammatory response which is manifested by the appearance of inflammatory cells. The majority of these inflammatory cells are lymphocytes [14]. Figure 1(a) shows a region of lymphocyte cells. The lymphocyte cells are generally smaller, with the cell nuclei smaller and darker than those of hepatocyte cells.
(2) Fibrosis - The death of small groups of hepatocytes may leave the reticulum (cell membrane system) intact and the resulting regeneration will repair the damage. However if the reticulum is damaged, healing can only occur by scar and will lead to fibrosis. If scars are produced throughout the liver the lack of blood circulation leads to cirrhosis [14]. Figure 1(b) shows an example of scarring. The remainder of the cells are lymphocytes.

As explained in section 1, the focus of this work is to measure the degree of inflammation relative to the amount of other tissue, not including the background. This means the main task of this system is to group tissue into two classes, inflammatory (C_1) and healthy (C_2). Scar tissue will therefore be classified as 'healthy' for our purpose.

[4] A tract of the portal system of the liver, which is a network of veins that begin and end in capillaries.
[5] Blood vessel in the liver that returns blood to the heart.
[6] Pathway for the transportation of bile from the liver to the gallbladder.

The biopsies used in this study are all stained using haematoxylin and eosin. This usually causes lymphocyte nuclei to appear dark purple, the hepatocyte nuclei to appear light purple and the background to appear white. Haematoxylin and eosin stain is commonly used by many pathology departments. This method can produce high colour variability across different samples as the stain mixture varies at different hospitals and laboratories. Another factor producing image variability is the illumination at the time of image capture. A system must be robust to these factors in order to interpret, adequately, new images.

2.1 Colour Correction

Cardei et al [15] propose a method of colour correction to counteract illumination variability. This involves using the difference between the background of images viewed under different illumination to colour-correct the whole image. This simple technique can be expanded to correct colour variation in tissue caused by stain and illumination change. In brief, a reference image is selected by eye, using the natural human ability to determine mid-range colour attributes. A raw image requiring colour correction is also selected. $\mathbf{Q}_{raw}$ is the raw RGB image reshaped into an $N \times 3$ matrix, where N is the total number of pixels in the image. Similarly, $\mathbf{Q}_{cc}$ is a matrix containing values of the colour-corrected image. Applying the diagonal model of illumination change [15] shows that

$$\mathbf{Q}_{cc} = \mathbf{Q}_{raw}\mathbf{M} \tag{1}$$

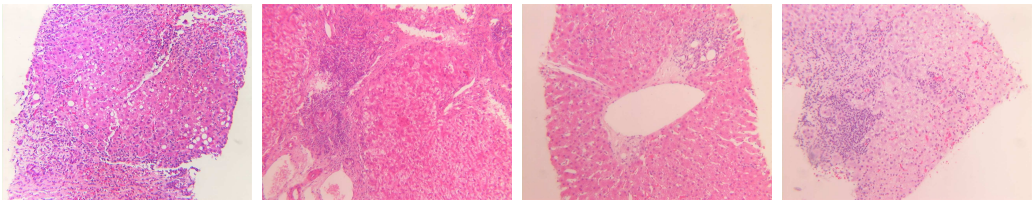
where

$$\mathbf{M} = \text{diag}\left(\frac{\lambda_{ref}^{R}}{\lambda_{raw}^{R}}, \frac{\lambda_{ref}^{G}}{\lambda_{raw}^{G}}, \frac{\lambda_{ref}^{B}}{\lambda_{raw}^{B}}\right). \tag{2}$$

Thresholding each image to remove the background leaves only the tissue portion, a region whose elements are defined by R_j^i where $i \in \{R, G, B\}$ and $j \in \{ref, raw\}$. The mean values, λ_j^i, are then derived from the tissue portion by

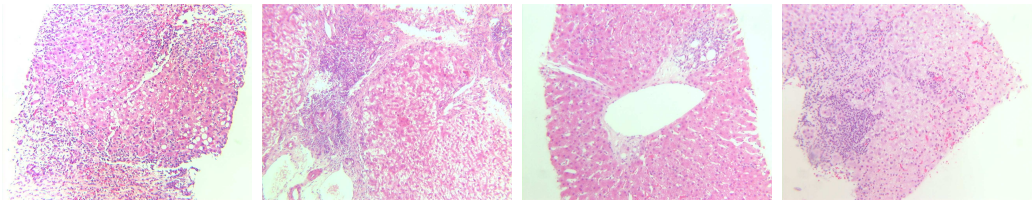
$$\lambda_j^i = \frac{1}{M_j} \sum_{m=1}^{M_j} (R_j^i)^m \tag{3}$$

where M_j is the number of pixels in each region and $(.)^m$ denotes the mth value. Figure 3 shows the result of colour correction on the images presented



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

Fig. 2. A sample of the training images before colour correction



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

Fig. 3. The images presented in figure 2 after colour correction

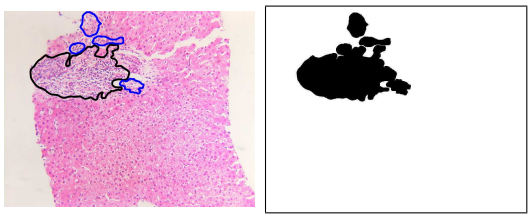
in figure 2. Qualitatively, the colour-corrected images can be seen to be more similar than the raw images.

3 Pattern Recognition System

In this section we present details of the images used during the training process and introduce a new pattern recognition system for identifying inflammation.

3.1 Training Images

To train the system, two sets of 86 colour images of liver biopsies are used. Set 1 contains the raw images and set 2 contains an annotated version of the raw images. Annotated images show regions of inflammation, as demonstrated in figure 4(a). To simplify the time consuming manual annotation process, inflammation was only annotated for close groups of six or more inflammatory (lymphocyte) cells. Each image is a 1000 $\times$ 1280 pixels bitmap of red,green,blue (RGB) layers, taken at 10$\times$ objective magnification and shows only a part of the whole liver biopsy. The images have been selected to show a cross-section of inflammatory and healthy cells, with variation in stain and illumination. The liver biopsy images were supplied and annotated by the third author, SSC, a consultant pathologist in the Academic Unit of Pathology, at



(a) (b)

Fig. 4. Method to generate the binary mask images for supervised training. (a) Annotation showing regions of inflammation. Portal tract inflammation is bordered with black and non-portal tract inflammation with blue. For this study, black and blue regions are considered of equal interest.(b) The binary mask image derived from the annotated image.

the University of Sheffield, UK. During pre-processing, the closed annotated regions shown in figure 4(a) are converted into binary masks, as demonstrated in figure 4(b). This is later overlaid on the raw image to supervise the training process.

3.2 New Approach

After completing the pre-processing steps of colour correction and the creation of the binary masks, the system is trained using the steps detailed in sections 3.2.1 to 3.2.4. The evaluation of new images using the trained system is then discussed in section 3.2.5.

3.2.1 Thresholding

To determine the position and extent of the individual cells, the image is first thresholded to highlight the cell nuclei. We call these the points of interest (POI) within each region since they provide an initial estimate of the location of whole cells. The method of thresholding uses histogram analysis. The histogram of grey levels [7] taken across the whole image is either unimodal or bimodal, depending on the amount of background included in the original image. An example histogram derived from one of our samples is shown in figure 5. The method uses the histogram lobe corresponding to the tissue region to calculate the threshold level. Therefore, it is first necessary to identify the tissue lobe and tissue lobe maximum. This is done by hill climbing a smoothed version of the original greyscale histogram, starting at the zero greyscale value (left side of figure 5). Once a peak is found, a simple search of

[7] Computed using the Matlab [16] function rgb2gray by transforming RGB to NTSC coordinates and retaining only the luminence values.

its immediate neighbourhood is performed to ensure this is the true lobe maximum. With the tissue lobe identified, thresholding at a suitable value within the lobe allows the darker cell nuclei to be segmented from the other tissue. Through experimentation, it was found that thresholding at 1.2 standard deviations (σ) below the tissue lobe maximum produces the best segmentation of cell nuclei across all training images. This method is illustrated in figure 5. Because of the non-gaussian form of the original histogram, σ is calculated by mirroring the lower half of the tissue lobe about the tissue lobe maximum and assuming a gaussian distribution.

3.2.2 Clump decomposition

Thresholding produces a binary representation of the cell nuclei. These are often touching or merged. This prevents identification of the true cell centroid and makes it impossible accurately to quantify the number of cells. Clump decomposition involves separating merged parts by analysis of the morphology of combined or clumped parts [17]. In this study, a method of uniform recursive erosion is implemented for clump decomposition, based on the well-known *watershed* [18] technique. This method is used to identify merged or marginally touching nuclei by splitting clumps at narrow points within the component. This method is demonstrated in figure 6 and discussed in more detail below:

(1) Initially, the morphological *opening* operator (with a 3×3 uniform, square structuring element) is used to remove noise from the thresholded image (see figure 6(b)). An *opening* consists of an erosion followed by a dilation [19].
(2) Each component (or clump) is then labelled using connected component analysis (CCA) [20]. This means each pixel within the image is allocated to an individual component and each component centroid is identified.

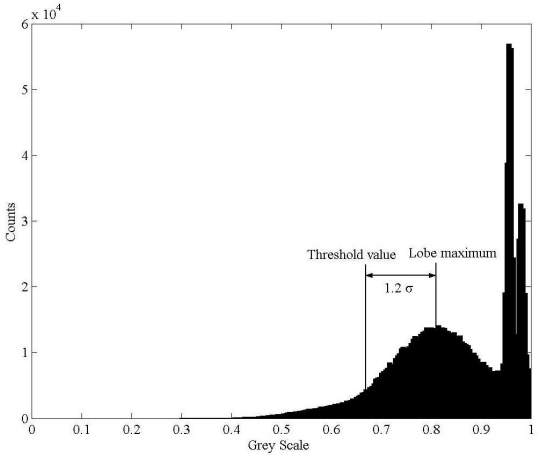


Fig. 5. Experiments show that thresholding at 1.2 standard deviations below the tissue lobe maximum produce the optimum thresholding results.

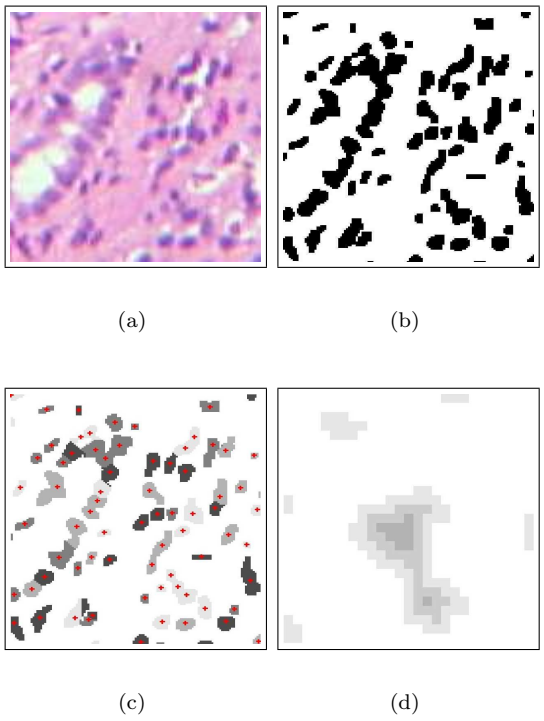


Fig. 6. Illustration of the clump decomposition technique. (a) Raw image. (b) Thresholded image after opening. (c) After clump decomposition. The markers show the new cell centroids and the patch-work effect illustrates the pixel allocation about them. (d) A close-up of the erosion technique. The clump of cells is eroded until the component disappears. Any resulting break-up of the component is used to generate new or 'child' components. Using these components, pixels from the original thresholded image are then re-allocated to the closest centroid.

(3) The binary image is then uniformly eroded using the same 3×3 structuring element. The erosion splits the components at narrow sections which correspond to marginally touching cells. If a component splits by this process, CCA is used to calculate the centroids of the newly created 'child' components. If a component does not split, the original 'parent' centroid remains.
(4) The image is then recursively eroded, using the methods described in step 3, until no more components remain (see figure 6(d)).
(5) The final list of component centroids, containing the resulting mixture of 'parent' and 'child' details, is superimposed onto the image produced in step 1. Each pixel in this binary image is then re-allocated to the nearest component centroid, thus creating the patch-work effect illustrated in figure 6(c).

Although more complex clump decomposition methods are available (see [17,21,22]), this study has found the erosion technique effective and computationally efficient.

3.2.3 Extracting Image Features

Outputs from the clump decomposition process (the component centroids and the surrounding patch-work) are used to identify the POI within the colour biopsy image. Features are then calculated from image data extracted about these POI, using one of two methods: (1) $k \times k \times 3$ blocks of image data centred on each component centroid; (2) all pixels belonging to the region supplied by the surrounding patch-work. For this study, a collection of image features is used to generate a D-dimensional feature vector to discriminate between inflammatory and healthy tissue. The notation used to define the features is presented in table 1 and the feature definitions are presented in table 2. Because the hyper-volume of the feature space increases exponentially as a function of dimensionality and most problems possess only a limited amount of data this rapidly leads to sparsely populated, high-dimensional spaces which are difficult to characterise [23]. For this reason, dimensionality reduction is seen as a key step in any pattern recognition system. Our system uses two main approaches to dimensionality reduction.

Table 1
Notation used for defining the image features in table 2

S	=	A region of RGB data defined by a block of $k \times k \times 3$ pixels centred on each POI.
L	=	A region of the image defined by all pixels belonging to a component, segmented by clump decomposition.
$i \in \{R, G, B\}$	=	Each layer of RGB data.
$\text{cell\_density}(O, r)$	=	The number of cells contained within a circle of radius r pixels, centered at the centroid of O.
$\|O\|$	=	number of pixels in O.
$O(x)$	=	The value of O at x.
$\text{sort}(O)$	=	A sorted version of O
$CH(O)$	=	A binary image representing the convex hull of O.
$\text{greyscale}(O)$	=	A grey scale representation of O.
$ELL(O)$	=	An ellipse with the same second-moments as region O.
$\text{major\_axis}(E)$	=	The length in pixels of the major axis of ellipse E.
$\text{minor\_axis}(E)$	=	The length in pixels of the minor axis of ellipse E.

Table 2
Feature definitions for $i \in \{R, G, B\}$ (see table 1 for the notation used)

Feature No.	Definition	Description
1–3	$\mu_i^{block} = \frac{1}{M}\sum_{m=1}^{M} S_i^m$, where $M = k \times k$	Block mean
4–6	$\sigma_i^{block} = \left(\frac{1}{M-1}\sum_{m=1}^{M}(S_i^m - \mu_i^{block})^2\right)^{\frac{1}{2}}$	Block standard deviation
7	$A = \frac{\lvert L\rvert}{3}$	Component area
8	$ECC = \frac{\text{major\_axis}(ELL(L))}{\text{minor\_axis}(ELL(L))}$	Component eccentricity
9	$ED = \sqrt{\frac{4A}{\pi}}$	Equivalent circle diameter
10	$SOL = \frac{A}{\lvert CH(L)\rvert}$	Solidity
11	$CD = \text{cell\_density}(L, r)$	Component density
12–14	$\mu_i^{cell} = \frac{1}{N}\sum_{n=1}^{N} L_i^n$, where $N = A$	Component mean
15–17	θ_i^{cell}	Component median
18–20	$\sigma_i^{cell} = \left(\frac{1}{N-1}\sum_{n=1}^{N}(L_i^n - \mu_i^{cell})^2\right)^{\frac{1}{2}}$	Component standard deviation
21	$\mu^{grey} = \frac{1}{N}\sum_{n=1}^{N} \text{greyscale}(L)^n$	Grey scale mean
22	θ^{grey}	Grey scale median
23	$\sigma^{grey} = \left(\frac{1}{N-1}\sum_{n=1}^{N}(\text{greyscale}(L)^n - \mu^{grey})^2\right)^{\frac{1}{2}}$	Grey scale standard deviation

3.2.3.1 Feature Selection In feature selection a subset of input features is selected for its suitability to a classification problem. This reduces dimensionality and the computational cost of feature gathering. The only guaranteed method of finding an optimal subset of d features from an original D-dimensional feature vector, is to perform an exhaustive search of all $D!/d!(D-d)!$ subsets of the reduced feature vector [24]. However, this is impractical because the number of subsets grows combinatorially. A number of suboptimal selection methods are available which are discussed in [24]. Of these, Jain et al [25] found that the sequential forward floating search (SFFS) [26] method produced the best results, performing close to the optimal, and demanding lower computational resources than other methods. The SFFS method is a bottom-up search procedure, where the term *floating* identifies that the number of features changes dynamically, with one feature included and/or excluded, at each iteration. The SFFS method is used here. To summarise the method, $X_d = \{x_i | i = 1, 2, \ldots, d, x_i \in Y\}$ is a subset of d features taken from a set $Y = \{y_j | j = 1, 2, \ldots, D\}$ of D available features. $J(X_d)$ is the criterion function used to evaluate the effectiveness of X_d. For this study J is chosen to be AUROC – the area under the receiver operating characteristic (ROC) curve, a commonly used test of classifier performance, discussed in [27] and section 4.1. The algorithm is initialised with an empty feature subset $X_0 = \emptyset$. The most significant feature from Y, $\arg\max_{y_j \in Y}$, is then added to the subset X_0. This step is then repeated once more, taking the most significant feature from the remaining available features $Y - X_1$. The following steps are then performed:

(1) The most significant feature from $Y - X_d$ is added to the current subset, X_d.
(2) The least significant feature, $\arg\max J(X_d - \{x_i\})$, is conditionally excluded from the current subset, X_d. If the newly added feature is the least significant or joint least significant with another feature, then step 1 is repeated. Otherwise, the least significant feature from the current subset, X_d, is excluded and step 3 is performed.
(3) This step is a continuation of the conditional exclusion in step 2. Once again the least significant feature, x_i, from X_d is located. If the resulting subset $X_d - \{x_i\}$ is better than the previous best subset of the same cardinality [8], then feature, x_i, is excluded from X_d and the current step is repeated. Otherwise, the feature is retained and step 1 is repeated.

If the cardinality of X_d returns to 2 at either exclusion step (2 or 3), then the algorithm goes to step 1. The algorithm terminates when the required cardinality is achieved. Through experimentation, a final cardinality not exceeding 12 provides the best results here. Table 3 demonstrates the progression of the SFFS algorithm and presents the final reduced subset of features in bold-face.

3.2.3.2 Feature Extraction Although, there are many feature extraction techniques available (see [24] for a review), this study implements principal component analysis (PCA) [28], one of the most widely used methods. The dimensionality of the d-dimensional ($d = 12$) feature vector derived from

[8] A record is kept of all previous subsets and their associated performance to enable this comparison.

Table 3
The feature subsets considered by the SFFS method. The final subset is in bold-face, the previous subsets demonstrate the floating nature of the SFFS technique. The features are defined in table 2

Iteration	Feature subset
1	{4}
2	{4, 11}
3	{4, 11, 3}
4	{4, 11, 3, 16}
5	{4, 11, 3, 16, 1}
6	{4, 11, 3, 16, 1, 13}
7	{4, 11, 3, 16, 1, 13, 21}
8	{4, 11, 3, 16, 1, 13, 21, 22}
9	{4, 11, 3, 16, 1, 13, 21, 22, 6}
10	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18}
11	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18, 14}
12	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18, 14, 7}
13	{4, 11, 3, 16, 1, 13, 21, 22, 6, 14, 7}
14	{4, 11, 3, 16, 1, 13, 21, 22, 6, 14, 7, 2}
15	{4, 11, 3, 16, 1, 13, 21, 22, 14, 7, 2}
16	{4, 11, 3, 16, 1, 13, 21, 14, 7, 2}
17	{4, 11, 3, 16, 1, 13, 14, 7, 2}
18	{4, 11, 3, 1, 13, 14, 7, 2}
19	{4, 11, 3, 1, 13, 14, 7, 2, 16}
20	{4, 11, 3, 1, 13, 14, 7, 2, 16, 21}
21	{4, 11, 3, 1, 13, 14, 7, 2, 16, 21, 22}
22	{**4,11,3,1,13,14,7,2,16,21,22,20**}

feature selection is further reduced using PCA. PCA seeks to project the high-dimensional input data into lower dimensional space [29]. In simple terms this means that new features are created from a linear transformation of the input features. The feature vector, $\vec{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$, is first normalised for all N data points using

$$\vec{y}^n = \mathbf{\Phi}^{-1}(\vec{x}^n - \overline{\vec{x}}), \quad (n = 1, \ldots, N) \tag{4}$$

where $\overline{\vec{x}} = (\overline{x}_1, \ldots, \overline{x}_d)$; $\overline{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n$; $\mathbf{\Phi} = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ and $\sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_i^n - \overline{x}_i)^2$. This normalisation is intended to counter the intolerance of PCA to data with different orders of magnitude [29]. To conduct PCA, the following are then computed for the normalised feature vector, $\vec{y}^n$.

$$\overline{\vec{y}} = \frac{1}{N}\sum_{n=1}^{N} \vec{y}^n \tag{5}$$

$$\mathbf{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(\vec{y}^n - \overline{\vec{y}})(\vec{y}^n - \overline{\vec{y}})^{\mathrm{T}} \tag{6}$$

The eigen-decomposition of $\mathbf{\Sigma}$

$$\mathbf{\Sigma}\vec{u}_t = \lambda_t \vec{u}_t, \quad (t = 1, \ldots, d) \tag{7}$$

is then calculated and sorted according to decreasing eigenvalue. Owing to the definition of $\mathbf{\Sigma}$ its eigenvalues are real and non-negative [29]. In most cases a small number of eigenvalues will dominate, indicating the inherent dimensionality of the data [28]. By forming a matrix, $\mathbf{U}$, whose columns are the $pc < d$ eigenvectors corresponding to the pc largest eigenvalues $\mathbf{U} = (\vec{u}_1, \ldots, \vec{u}_{pc})$, it is possible to define $\vec{z} = \mathbf{U}^{\mathrm{T}}(\vec{y} - \overline{\vec{y}})$ a pc-dimensional vector of linearly transformed variables. The choice of the number of principal components, pc, will be discussed in section 4. Although, in principle, PCA should provide optimal dimensionality reduction without feature selection. The prohibitive cost of generating large numbers of features makes the inclusion of initial feature selection desirable for this study.

3.2.4 Probability Density Estimation

The conclusion of the training process is to derive the class-conditional probability densities, $p(\vec{z}|C_j)$. The density estimate can then be used for the Bayesian classification discussed in section 3.2.5.1. The binary masks (see figure 4) can be overlayed onto the output from the clump decomposition process

to provide the class (C_j, $j = 1, 2$) labels for each transformed feature vector, $\vec{z}_j$. As a result, we can approximate the required probability distribution for each class. In this study two methods of density estimation are compared:

(1) Gaussian parametric model (GPM)—This is a parametric method where a gaussian probability density function (PDF) is assumed. This technique is easy to compute and simple to implement. For the multivariate case, the PDF takes the form

$$p(\vec{z}|C_j) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_j|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{z} - \overline{\vec{z}}_j)^{\mathrm{T}} \mathbf{\Sigma}_j^{-1} (\vec{z} - \overline{\vec{z}}_j)\right\} \tag{8}$$

where the class covariance matrix $\mathbf{\Sigma}_j$ and mean vector $\overline{\vec{z}}_j$ are estimated from the transformed feature vectors of the training set.

(2) Gaussian mixture model (GMM)—This is a semi-parametric method [29] where mixtures of gaussians are used to build more complex models e.g. multimodal PDFs [28]. For the multivariate case, the probability density function for each class is estimated by a linear combination of K_j ($j = 1, 2$) gaussian basis functions of the form

$$p(\vec{z}|C_j) = \sum_{k=1}^{K_j} P_{jk} p_j(\vec{z}|k), \quad (\vec{z} \in C_j) \tag{9}$$

where

$$p_j(\vec{z}|k) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_{jk}|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{z} - \overline{\vec{z}}_{jk})^{\mathrm{T}} \mathbf{\Sigma}_{jk}^{-1} (\vec{z} - \overline{\vec{z}}_{jk})\right\} \tag{10}$$

where $\mathbf{\Sigma}_{jk}$ is the covariance matrix for the k^{th} gaussian for the j^{th} class and $\overline{\vec{z}}_{jk}$ is the mean vector for the k^{th} gaussian for the j^{th} class. Typically, the GMM parameters are determined using the *expectation-maximisation* (EM) algorithm [30] as is used here.

The performance of each estimator is evaluated during the optimization process outlined in section 4.

3.2.5 Evaluating new images

When a new image is presented to the system, it is first necessary to perform the following, previously discussed, steps:

- Colour-correct the new image (using the original reference image).
- Threshold the image.
- Apply clump decomposition.
- Extract the feature vector, $\vec{x}$, from each POI.
- Apply PCA to produce $\vec{z}$.

The estimated densities (GPM or GMM) are then used to provide density estimates of $\vec{z}$, to give the likelihood of the region surrounding each POI belonging to a particular class. Bayes theorem (11) can then be used to calculate the posterior probability of class membership, which allows a decision to be made regarding class membership of a particular region. This method is discussed in more detail below.

3.2.5.1 Classification Bayes theorem permits the posterior probability, $P(C_j|\vec{z})$, to be expressed in terms of the prior probability, $P(C_j)$, the class-conditional probability density function, $p(\vec{z}|C_j)$, and a normalisation factor, $p(\vec{z})$ [28], thus

$$P(C_j|\vec{z}) = \frac{p(\vec{z}|C_j)P(C_j)}{p(\vec{z})} \tag{11}$$

$P(C_j)$ is the probability of each class occurring based on *a priori* knowledge of the training set. $p(\vec{z}|C_j)$ is the class-conditional probability density function. In practice, an *estimate* of the probability density function for each class is required, as discussed in section 3.2.4. By assuming that new regions belong to one of the two classes C_1–inflammatory or C_2–healthy, then the posterior probabilities obey

$$P(C_1|\vec{z}) = 1 - P(C_2|\vec{z}) \tag{12}$$

Each region may then be assigned class membership according to a user-defined classification threshold, T, as follows

$$\begin{aligned} P(C_1|\vec{z}) > T, \text{ then assign to } C_1 \\ P(C_1|\vec{z}) < T, \text{ then assign to } C_2 \end{aligned} \tag{13}$$

where $0 < T < 1$. The method of selecting a suitable classification threshold is discussed in section 4.1.

4 Optimization

The role of optimization in this study is to select a good set of the adjustable system parameters: block size; number of priciple components; the method of density estimation, and, for the GMM only, the number of mixture components. To determine the optimum system performance it is necessary to

evaluate images with pre-classified cells. As the training images discussed in section 3.1 already provide pre-classified cells, the system is optimized by the m-fold cross-validation ($m = 10$) of these training images. This simply means the training images are randomly divided into m equally sized subsets [28]. With the remainder used for training, the system then evaluates one subset. This operation is then repeated until all subsets have been evaluated. It can be shown [28] that averaging m performance measures gives an estimate of the true system performance. A method of quantifying system performance from the results of m-fold cross-validation is discussed in section 4.1. The optimization method and the final optimized system parameters are presented in section 4.2.

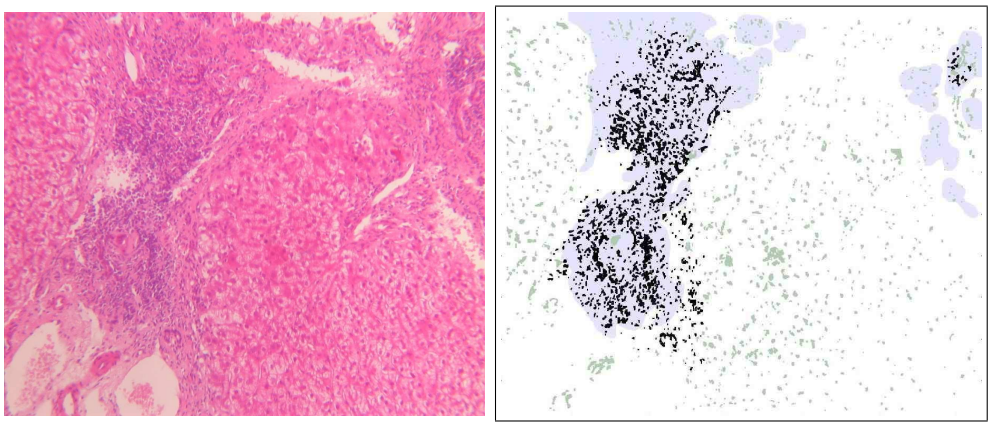
4.1 The Receiver Operating Characteristic Curve

System performance may be evaluated using contingency table data derived from the m-fold cross-validation of the training set. The contingency table is defined in table 4 and an example of the results obtained from cross-validation of the training set is shown in figure 7.

Table 4
Contingency table definition. The observer results are derived from the annotated test images and the test results are derived from cell classification at a particular classification threshold.

	Test results	
Observer results	Class 1	Class 2
Class 1	#True Positive (TP)	#False Negative (FN)
Class 2	#False Positive (FP)	#True Negative (TN)

For a two-class problem it is possible to evaluate the system more robustly by plotting the receiver operating characteristic (ROC) curve, a technique commonly used in medical imaging [31]. The ROC curve is constructed from contingency table data by plotting *sensitivity*, $TP/(TP + FN)$, against one minus *specificity*, $FP/(TN + FP)$, as the classifier threshold varies from 0 to 1. The area under the ROC curve can be considered to be a measure of the overall quality of the classification model [27]. Maximising the area by changing important system variables leads to the optimum classifier. As each point lying on the curve corresponds to a different threshold, the final classification threshold may be chosen based on desired levels of sensitivity and/or specificity. Using the Neyman-Pearson criterion (NPC) for this purpose [27], a maximum false positive rate (one minus specificity) is specified by the user, as shown in figure 8. The final classification threshold, T, is then selected with the highest false positive rate less than the NPC. For the purpose of illustration in this study, the maximum permitted false-positive rate is set at 0.1.



(a) Raw image (b) Experimental result

Fig. 7. Experimental result for training image RAW00067 generated from 10-fold cross-validation of the training set. Tissue classed as inflammatory is shown in black, healthy in grey and the regions identified, by SSC, as inflamed, in blue. This result shows that the majority of nuclei within the largest annotated region have been classified correctly. However, classification is less successful in the outlying regions.

4.2 Response Surface Methodology

Response surface methodology (RSM) is a technique to reduce the cost of optimization by searching for combinations of variables (factors) that maximise the performance of the system [32]. In this study RSM is used to maximise the AUROC by searching for the optimum value of key system factors. The first stage of this technique is to develop a strategy for gathering experimental data, known as the 'design of experiments' (DoE) [33]. This involves identifying factors that have a significant effect on the response of the system, a procedure normally carried out by screening out insignificant factors during the development process. Once identified, the factors are constrained to an allowable range by identifying suitable upper and lower limits for each factor. The range is then discretised at equal spacing to generate *levels* within the allowable range. A common approach when considering a small number of input factors (less than five) is to evaluate the system at all combinations of factors and corresponding levels. This approach is known as a *full-factorial* design [33]. The next stage of RSM is to develop a model of the system response. This model can then be searched to find the maximum predicted response and thus the optimal factor values. For an example with two factors, the *full-factorial* design provides a 2-dimensional grid of system evaluation points. The model can then be visualised as a response surface constructed on the grid. Considering a system more formally

$$y = f(\vec{v})$$

where y is the system response, f is an unknown function and $\vec{v} = (v_1, v_2, \ldots, v_q)$ is a vector of q independent factors. It is common to construct a model of this system by fitting a low-order (either linear, quadratic or cubic) polynomial to the experimental data. For a cubic polynomial with p combinations of factors and levels, this takes the form

$$y^n = b_0 + \sum_{i=1}^{q} b_i v_i^n + \cdots \sum_{i=1}^{q}\sum_{j=1}^{q} b_{ij} v_i^n v_j^n + \cdots \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} b_{ijk} v_i^n v_j^n v_k^n \qquad (14)$$

where b_0, b_i, b_{ij} and b_{ijk} are the unknown polynomial coefficients. y^n is the experimental observed response value and $n = 1, \ldots, p$. This model can be written in matrix notation as $\vec{y} = \mathbf{V}\vec{b}$, where $\vec{y} = (y^1, y^2, \ldots, y^p)$,

$$\vec{b} = (b_0, b_1, \ldots, b_q, b_{11}, \ldots, b_{qq}, b_{111}, \ldots, b_{qqq})$$

and $\mathbf{V}$ is the experimental design matrix constructed from p rows of $\hat{\vec{v}}^n$, a vector corresponding to the factor terms in 14, of the form

$$\hat{\vec{v}}^n = (1, v_1^n, \ldots, v_q^n, v_1^n v_1^n, \ldots, v_q^n v_q^n, v_1^n v_1^n v_1^n, \ldots, v_q^n v_q^n v_q^n)$$

The coefficients, $\vec{b}$, can then be estimated using the least squares method $\vec{b} = \mathbf{V}^\dagger \vec{y}$ where $\mathbf{V}^\dagger$ is the *pseudo-inverse* of $\mathbf{V}$ [29]. The value of the input variables that provide the maximum system response can then be derived from the polynomial given by $\mathbf{V}\vec{b}$. Considering the DoE for this study, the following system factors have a significant effect on the system response.

(1) Block size (k)—Features 1–6 defined in table 2 rely on square blocks of data extracted from around each cell centroid. This factor represents the block size and is constrained between 1 and 81 pixels. The discretised levels of this allowable range are $\{21, 41, 61, 81\}$.
(2) Number of Principal Components (pc)—This factor governs the number of dimensions that the reduced feature vector is mapped to and is constrained between 1 and 12 (levels = $\{3, 6, 9, 12\}$).
(3) Density estimation method—Either GPM or GMM, as discussed in section 3.2.4. For GMM only, the following extra factor requires optimization:
    (a) Number of basis functions (bf)—The number of gaussian basis functions to fit the data. This is constrained to be between 1 and 8 (levels = $\{2, 4, 6, 8\}$).

Table 5
Table showing the optimized system factors for both GPM and GMM density estimation methods and the corresponding AUROC.

Factor	Value (GPM)	Value (GMM)
Block size (k)	55	53
Number of principal components (pc)	5	6
Number of gaussian basis functions (bf)	N/A	5
AUROC	0.9559	0.9619

The final cardinality of the feature selection method, discussed in section 3.2.3.1, should also be treated as a factor. However, this is impractical owing to the high computational cost of combining the SFFS technique with the full-factorial design. To compare the GPM and GMM density estimation (DE) methods, two response surfaces are generated. This allows the system to be independently optimized for each DE method. The optimized systems are then compared for the maximum system response by re-evaluating the AUROC. A cubic polynomial is used to model the response in both cases. With this in mind, the two forms of DE will now be considered.

- GPM—only variables k and pc are applicable. Evaluating these variables using a full-factorial design requires 16 evaluations of the AUROC.
- GMM—variables k, pc and bf are applicable. Evaluating this set using a full-factorial design requires 64 evaluations of the AUROC.

As all the factors under discussion can take only integer values within the constraints discussed previously, the maximum predicted response may be derived by evaluating the model at all variable combinations between the upper and lower bounds for each variable. Although this is a combinatorial problem, the task of evaluating the model is computationally trivial in comparison to evaluating the AUROC. Searching each model for the maximum response using this method provides the factor values listed in table 5. Evaluating the AUROC at these parameters shows that GMM provides the optimum DE method. However the improvement gained by using the GMM method rather than the computationally more efficient GPM method is only marginal (0.65%). As the intended end-users of this system are pathologists, it is thought that adopting the conceptually simpler GPM method will aid the understanding and trust of this system by medical professionals who may not be familiar with pattern recognition theory. Therefore the GPM method is focused on in what follows.

The ROC curve for the optimum GPM configuration is illustrated in figure 8. By applying the NPC technique discussed in section 4.1, the final classification threshold of 0.22 can be derived from the curve.

5 Testing

It is common to test a system of this type against a 'gold standard', a set of universally accepted accurately quantified test images. However, no 'gold standard' exists for liver biopsy inflammation. The closest alternative is the Ishak scoring system [5] discussed in section 1, but as previously stated, this suffers from high inter- and intra-observer variability. With this in mind, our system is tested using a separate group of 15 test images previously evaluated in a study by Cross at al [34]. In this study, 25 observers (including 5 consultant pathologists, 4 trainee pathologists and 16 control observers – medical students with no previous experience of histology or pathology) were asked to compare 15 liver biopsy images with varying degrees of inflammation. It can be assumed that consultant pathologists have the most experience in identifying cell inflammation, followed by the trainee pathologists and finally the 16 control observers. The images are named '*mild1*, ..., *mild5*, *mod1*, ..., *mod5*, *sev1*, ..., *sev5*' but this only indicates a preliminary approximation of the level of inflammatory activity. Each of the 15 images is compared to each other image producing 105 pairs. The observers are then asked to identify the image containing the most inflammation from each pair. A rank order of images is then produced for each observer using a ranking algorithm normally used to rank competitive chess players [35]. Finally, Spearman rank correlation (SRC) [36] is used to assess, statistically, the level of agreement between the observers. As the results show good inter-observer agreement (SRC=0.95) using this comparison technique, it is proposed that image ranks from each observer can be used to test the proposed, automated system.

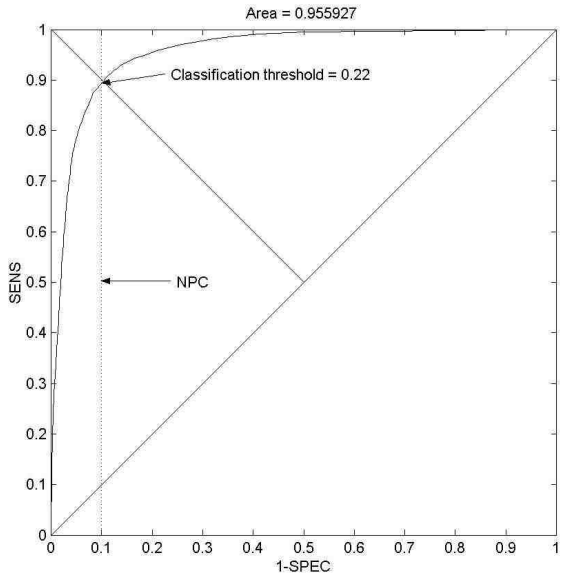


Fig. 8. ROC curve for the optimized system. For the purpose of illustration the maximum false-positive rate is set at 0.1.

Initially, the optimized system is trained on all 86 training images. Each test image is then processed using the methods outlined in section 3.2.5 and using the classification threshold determined by NPC (see figure 8). The test images are the same size and magnification as the training images. Quantification of the inflammation is carried out by counting the number of regions classified either C_1–inflammatory or C_2–healthy. The proportion of inflammatory tissue is then computed. The time taken to process, fully, each *new* image is variable and dependent on a number of factors. The two most significant of these, are the number of clumped nuclei and the final number of fully segmented nuclei. For the test images, the mean computational processing time per image and standard error is 57.31 ± 4.51 seconds. This system was developed using the Matlab software environment [16] and the test was performed on a standard desktop PC (Pentium P4 - 1.6 GHz processor) using the Microsoft Windows XP operating system (Microsoft Corporation, WA).

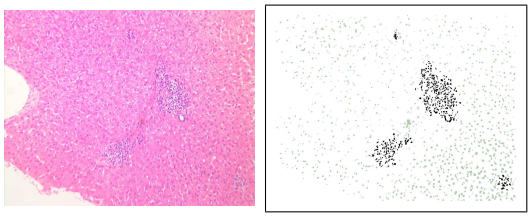
Quantifying the inflammation in percentage form, allows the images to be placed in rank order of severity and compared to the image ranks from the previous study [34]. As in [34], SRC can then be used to assess the level of agreement between the image rank produced by this automated system and the image ranks produced by the 25 observers. The SRC coefficient is defined by

$$r_s = 1 - \frac{6 \sum_{n=1}^{N} d_n^2}{N(N^2 - 1)}$$

where d_n is the difference between each pair of ranks and N is the number of paired observations. The resulting values of r_s show good agreement between the observers and the automated system. Using a null hypothesis that there is no correlation between any of the ranks, the *significance* of each r_s value can be determined by calculating the probability (P-value) that this hypothesis is true [36]. For $N > 10$ ($N = 15$ in this case), r_s has an approximately Normal distribution with a mean of zero and a variance of $1/(n-1)$ [37]. To test the significance of r_s, the z value is first calculated as follows [36]

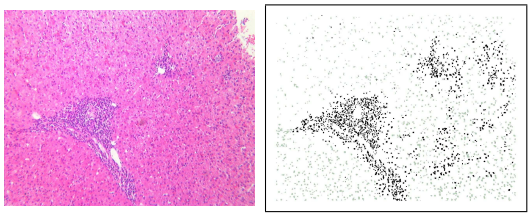
$$z = \frac{r_s}{\sqrt{\frac{1}{(n-1)}}} = r_s \sqrt{(n-1)}$$

The P-value is then determined from z, using tables of the area under the Normal distribution curve [37]. Table 6 shows the r_s values corresponding to the relationship between consultant pathologists and the automated system. The probability that the null hypothesis is true, for each of these r_s values, is $P < 10^{-3}$. P-values were also calculated for all other observers (not shown). $P < 10^{-3}$ is also the maximum when considering the correlation between our system and the trainee pathologists. However the P-value rises to $P < 10^{-2}$,



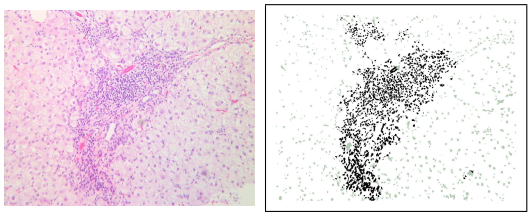
(a) Raw image (‘mod 1’)

(b) Test result (‘mod 1’). Showing 22.3% inflammation.



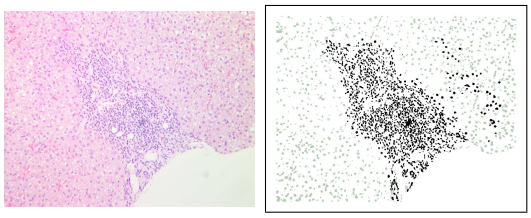
(c) Raw image (‘mild 1’)

(d) Test result (‘mild 1’). Showing 49.4% inflammation.



(e) Raw image (‘mild 4’)

(f) Test result (‘mild 4’). Showing 57.1% inflammation.



(g) Raw image (‘sev 3’)

(h) Test result (‘sev 3’). Showing 53.2% inflammation.

Fig. 9. A sample of the experimental test results. Inflamed cells are shown in black and healthy cells are grey.

when considering the correlation between control observers and our system. Historically $P < 10^{-2}$ or a ‘one percent probability level’ suggests that the null hypothesis may be rejected [38]. Although, this means that the null hypothesis

Table 6
The SRC coefficients relating the automated system (computer) and five consultant pathologists. For SRC: -1 = systematic disagreement; 0 = no connection; 1 = strong agreement. The *significance* (P-value) of all the results shown is $P < 10^{-3}$.

	Consultant 1	Consultant 2	Consultant 3	Consultant 4	Consultant 5	Computer
Consultant 1	1.000	0.946	0.936	0.936	0.971	0.943
Consultant 2	0.946	1.000	0.950	0.968	0.971	0.989
Consultant 3	0.936	0.950	1.000	0.911	0.943	0.971
Consultant 4	0.936	0.968	0.911	1.000	0.975	0.964
Consultant 5	0.971	0.971	0.943	0.975	1.000	0.971
Computer	0.943	0.989	0.971	0.964	0.971	1.000

may be rejected for all observer groups, the low P-values between consultants and this system, suggest a strong correlation. It can also be shown that consultants have the lowest inter-observer variability with each other [34]. This means our system can rank 15 test images in correlation to five consultant pathologists, who in turn agree more strongly with each other than with the other observers, indicating this system has an expert capability in this test.

For completeness the above tests were also conducted using the GMM density estimation method discussed in sections 3.2.4 and 4. Although the GMM method produces a better system performance (greater AUROC) when considering 10-fold cross-validation of the training set, results here show a similar performance to the GPM method when considering the correlation between our system and the three observer groups. This confirms the decision, made in section 4.2, to adopt the GPM for density estimation on the basis of its simplicity.

6 Conclusions

An effective and systematic method of evaluating the liver biopsies of patients with hepatitis C will become increasingly important owing to the large number of people currently infected with the disease. Previous approaches to similar cell classification problems do not adequately address the specific issues associated with the automatic segmentation and classification of inflammatory cells in HCV-infected liver biopsies. The system outlined in this study offers a fully automatic pattern recognition solution to quantify the amount of inflammatory tissue. The simplicity of the pattern recognition techniques used aids the understanding and trust of this system by pathologists and facilitates the implementation of this system on a standard desktop PC in a pathology laboratory. Important steps forward have been made in: (1) colour-correcting images for stain and illumination variability, (2) the segmentation of individual cells via clump decomposition and (3) the application of feature selection and extraction methods to reduce the cost of feature gathering and the di-

mensionality of the feature vector.

The comparison of two commonly used density estimation methods, GPM and GMM, shows that the simpler GPM technique provides an equivalent system performance when considering novel images. The implementation of the GPM method shows the system can rank a set of 15 previously unseen test images in correlation to five consultant pathologists and four trainee pathologists with a level of significance of $P < 10^{-3}$. Although the level of significance is reduced when considering the relationship between this system and the control observers ($P < 10^{-2}$), the consultants have the lowest inter-observer variability and have the most experience of interpreting biopsy images. Therefore, the correlation between consultants can be considered to be the closest thing to a 'gold standard' for this test.

The results of this study show that the system has a capability which is at least as good as the current scoring systems which are applied subjectively by consultant histopathologists. It may be that the system has a performance which exceeds that of the consultants but it has not been possible to identify a 'gold standard' with which the system can be compared. A future study could include using this system on a retrospective series of biopsies where the outcome of disease progression (as measured by the amount of fibrosis in a subsequent liver biopsy) was known and then to examine how accurately the system predicts this outcome on a test series of cases. Since the current system shows promise for a clinically-useful system then consideration needs to be given as to how such a system would integrate into working practices. The level of agreement in assessment of liver inflammation between consultant histopathologists using the current scoring systems is relatively low so it is not difficult to justify the use of an automated image analysis system if its performance is more accurate and/or reproducible. The current system needs to be developed so that it would measure inflammation in an entire biopsy, rather than an isolated field of view, but there is currently available scanning technology that would allow this to be done.

Acknowledgements

SH would like to thank the UK EPSRC for its financial support for this work.

References

[1] WHO, Hepatitis C global prevalence (update), Weekly Epidemiological Record (World Health Organisation) 74 (1999) 421–428.

[2] WHO, Hepatitis C (Fact Sheet No. 164), World Health Organisation, Geneva, 2000.

[3] J. Booth, J. O'Grady, J. Neuberger, Clinical guidelines on the management of hepatitis C, Gut 49 (Supplement 1) (2001) i1–21.

[4] Chief Medical Officer, Annual Report of the Chief Medical Officer 2001, Department of Health, London, 2001.

[5] K. Ishak, A. Baptista, L. B. et al, Histological grading and staging of chronic hepatitis, Journal of Hepatology 22 (1995) 696–699.

[6] R. Goldin, J. Goldin, A. B. et al, Intra-observer and inter-observer variation in the histopathological assessment of chronic viral hepatitis, Journal of Hepatology 25 (5) (1996) 649–654.

[7] T. Würflinger, J. Stockhausen, D. Meyer-Ebrecht, A. Böcking, Robust automatic coregistration, segmentation, and classification of cell nuclei in multimodal cytopathological microscopic images, Computerized Medical Imaging and Graphics 28 (2004) 87–98.

[8] M. Donato, E. Arosio, V. Monti, P. Fasani, D. Prati, A. Sangiovanni, G. Ronchi, M. Colombo, Proliferating cell nuclear antigen assessed by a computer-assisted image analysis system in patients with chronic viral hepatitis and cirrhosis, Digestive and Liver Disease 34 (2002) 197–203.

[9] G. Lake-Bakaar, V. Mazzoccoli, L. Ruffini, Digital image analysis of the distribution of proliferating cell nuclear antigen in hepatitis c virus-related chronic hepatitis, cirrhosis, and hepatocellular carcinoma, Digestive Diseases Sciences 47 (2002) 1644–1648.

[10] K. Werling, Z. Szentirmay, A. Szepesi, Z. Schaff, F. Szalay, Z. Szabo, L. Telegdy, K. David, G. Stotz, Z. Tulassay, Hepatocyte proliferation and cell cycle phase fractions in chronic viral hepatitis c by image analysis method, European Journal of Gastroenterology and Hepatology 13 (2001) 489–493.

[11] T. Caballero, A. Perez-Milena, M. Masseroli, F. O'Valle, F. Salmeron, R. del Moral, G. Sanchez-Salgado, Liver fibrosis assessment with semiquantitative indexes and image analysis quantification in sustained-responder and non-responder interferon-treated patients with chronic hepatitis c, Journal of Hepatology 34 (2001) 740–747.

[12] M. O'Brien, N. Keating, S. Elderiny, S. Cerda, A. Keaveny, N. Afdhal, D. Nunes, An assessment of digital image analysis to measure fibrosis in liver biopsy specimens of patients with chronic hepatitis c, American Journal of Clinical Pathology 114 (2000) 712–718.

[13] P. Scheuer, J. Lefkowitch, Liver Biopsy Interpretation, 5th Edition, W.B. Saunders Company Ltd, London, 1994.

[14] E. Orfei, Review of pathology of the liver, Department of Pathology, Stritch School of Medicine, Loyola University of Chicago, http://www.meddean.luc.edu/, 2003.

[15] V. Cardei, B. Funt, M. Brockington, Issues in color correcting digital images of unknown origin, in: CSCS 12, Bucharest, 1996.

[16] MathWorks, Matlab version 7.3, The MathWorks Inc., Natick, MA, 2005.

[17] T. Teo, X. Jin, S. Ong, Jayasooriah, R. Sinniah, Clump splitting through concavity analysis, Pattern Recognition Letters 15 (1994) 1013–1018.

[18] J. Russ, The image processing handbook, CRC Press, London, 1992.

[19] R. Schalkoff, Digital image processing and computer vision, John Wiley & Sons, Inc., 1989.

[20] R. Haralick, L. Shapiro, Computer and Robot Vision, Vol. I, Addison-Wesley, Reading,MA, 1992.

[21] C. Xu, J. Prince, Gradient vector flow: A new external force for snakes, in: Conference on Computer Vision and Pattern Recognition, IEEE, 1997, pp. 66–71.

[22] L. Liu, S. Sclaroff, Deformable shape detection and description via model-based grouping, Tech. Rep. 98-017, Computer Science, Boston University (November 1998).

[23] J. Hallinan, Detection of malignancy associated changes in cervical cells using statistical and evolutionary computational techniques, Ph.D. thesis, Cytometrics Project, Centre for Sensor Signal and Information Processing, University of Queensland, Australia (1999).

[24] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[25] A. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 153–158.

[26] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.

[27] M. Scott, M. Niranjan, R. Prager, Parcel: feature subset selection in variable cost domains, Tech. Rep. CUED/F-INFENG/TR. 323, Cambridge University Engineering Department (May 1998).

[28] R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd Edition, John Wiley and sons, New York, 2001.

[29] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.

[30] I. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer-Verlag, 2002.

[31] J. Hanley, B. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 143 (1982) 29–36.

[32] R. Myers, D. Montgomery, Response surface methodology: process and product optimization using designed experiments, Wiley-Interscience, New York, 2002.

[33] G. Box, N. Draper, Empirical Model-Building and Response Surfaces, Wiley, New York, 1987.

[34] S. Cross, N. Bashir, P. Hempshall, S. Hodgson, R. Harrison, Comparison of paired images produces a much higher level of interobserver agreement than estimation from a single image in the assessment of inflammation in chronic hepatitis, in: Histopathology 2002, Vol. 41(Suppl 1), 2002, p. 43.

[35] Internet Chess Club, ICC Help File: RATINGS, World Wide Web, http://www.chessclub.com/help/ratings, 2002.

[36] F. Zuwaylif, Applied General Statistics, 3rd Edition, Mass:Addison-Wesley Pub. Co, Reading, 1979.

[37] M. Bland, An Introduction to Medical Statistics, 2nd Edition, Oxford Medical Publications, Oxford, 1995.

[38] M. Kendall, J. Gibbons, Rank correlation methods, Edward Arnold, London, 1990.