Computer Methods and Programs in Biomedicine 140 (2017) 61–68

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A methodology for automated CPA extraction using liver biopsy image analysis and machine learning techniques

Markos G. Tsipouras[a,b], Nikolaos Giannakeas[a,b], Alexandros T. Tzallas[a,b,*], Zoe E. Tsianou[a], Pinelopi Manousou[c], Andrew Hall[d], Ioannis Tsoulos[b], Epameinondas Tsianos[a]

[a] Division of Gastroenterology, Faculty of Medicine, School of Health Sciences, University of Ioannina, GR45110 Ioannina, Greece
[b] Department of Computer Engineering, School of Applied Technology, Technological Educational Institute of Epirus, Kostakioi, GR47100, Arta, Greece
[c] Liver Unit, St Mary's Hospital, Imperial College NHS Trust, London, UK
[d] Department of Histopathology, UCL Medical School, Royal Free Campus, Rowland Hill Street, London NW3 2QG, UK

ARTICLE INFO

Article history:
Received 6 February 2016
Revised 12 November 2016
Accepted 22 November 2016

Keywords:
Collagen proportional area
Liver biopsy image analysis
Machine learning techniques
Clustering
Classification

ABSTRACT

Background and objective: Collagen proportional area (CPA) extraction in liver biopsy images provides the degree of fibrosis expansion in liver tissue, which is the most characteristic histological alteration in hepatitis C virus (HCV). Assessment of the fibrotic tissue is currently based on semiquantitative staging scores such as Ishak and Metavir. Since its introduction as a fibrotic tissue assessment technique, CPA calculation based on image analysis techniques has proven to be more accurate than semiquantitative scores. However, CPA has yet to reach everyday clinical practice, since the lack of standardized and robust methods for computerized image analysis for CPA assessment have proven to be a major limitation.

Methods: The current work introduces a three-stage fully automated methodology for CPA extraction based on machine learning techniques. Specifically, clustering algorithms have been employed for background-tissue separation, as well as for fibrosis detection in liver tissue regions, in the first and the third stage of the methodology, respectively. Due to the existence of several types of tissue regions in the image (such as blood clots, muscle tissue, structural collagen, etc.), classification algorithms have been employed to identify liver tissue regions and exclude all other non-liver tissue regions from CPA computation.

Results: For the evaluation of the methodology, 79 liver biopsy images have been employed, obtaining 1.31% mean absolute CPA error, with 0.923 concordance correlation coefficient.

Conclusions: The proposed methodology is designed to (i) avoid manual threshold-based and region selection processes, widely used in similar approaches presented in the literature, and (ii) minimize CPA calculation time.

1. Introduction

Hepatitis C is one of the most common liver diseases worldwide, with 2%–3% of the world's population living with HCV infection [1]. Patients with chronic hepatitis C and elevated transaminases, which are candidates for treatment, often undergo liver biopsy for determination of hepatic fibrosis, i.e. determination of the collagen, which is the major component of fibrotic tissue. For assessing disease staging and provide diagnosis needle biopsy specimens are cut and stained using several dyes that bind selectively to liver collagen, making the quantification of the fibrosis suitable [2]. The most common dyes are the Masson's Trichrome for standard histological evaluation and the Sirius red for fibrosis detection. Pathologists assess the severity of HCV via microscopy images of liver biopsy specimens.

The liver biopsy evaluation systems, which are actually semiqualitative scores, are based on the assessment of architectural and structural findings in liver tissue. They assume that fibrosis generates around portals tracks and bridged from each portal track to other neighboring portals tracks, leading to the segmentation of liver tissue into dysfunctional areas. In this way, a staging of the disease is provided without taking under consideration the amount or degree of fibrosis. There are four scoring systems for the staging of liver diseases: a) the Knodell Histology Activity Index (HAI) [3], b) the Scheuer HAI [4], c) the Metavir scoring system [5], and d) finally the Ishak HAI [6]. Instead of the above approach, Manousou

* Corresponding author.
E-mail addresses: tsipouras@teiep.gr (M.G. Tsipouras), giannakeas@teiep.gr (N. Giannakeas), tzallas@teiep.gr (A.T. Tzallas), z.tsianou@gmail.com (Z.E. Tsianou), pinelopi.manousou@imperial.nhs.uk (P. Manousou), andrewhall1@nhs.net (A. Hall), itsoulos@teiep.gr (I. Tsoulos), etsianos@uoi.gr (E. Tsianos).

http://dx.doi.org/10.1016/j.cmpb.2016.11.012
0169-2607/

et al. [7], have indicated that CPA computation could provide also an effective way for liver disease staging estimation. The ratio of collagen in whole liver tissue (Collagen Proportionate Area – CPA), which can be measured in images, can quantitate the fibrosis, and has already been associated with the clinical course and outcome in patients with hepatitis C. Image analysis of liver biopsy images for CPA assessment has gained attention is the last decade, and has been used as a marker for fibrosis quantification in several studies.

In the extracted biopsy image, a number of needle tissue specimens are presented. According to the ideal assumption, only liver tissue should be placed on the substrate, as well as only fibrotic tissue should bind the dyes. Unfortunately, actual liver biopsy samples and the respective obtained images include several types of regions (beside liver tissue), with some of them retaining high dye concentrations. These can include: capsule or structural collagen, blood clots, blood vessels, muscles, fat, as long as artifacts such as dye drops, dust and scratches. The existence of all above undesired regions in the image makes the employment of sophisticated image analysis techniques imperative.

Although several works have been presented in the literature employing CPA assessment through image analysis, most of them mainly focus on the medical problem and attempt to correlate CPA with known HAI scores. Thus, include little details regarding the image analysis techniques [2,7–11]. Commonly, CPA analysis methodologies include a segmentation procedure (such as histogram thresholding) for detecting tissue and collagen areas and an additional artifact exclusion procedure, for the rejection of areas that should not be included in CPA calculation. Several research groups utilize image processing software, such as Zeiss KS300 [12], Adobe Photoshop [13], Nexus cube [14], or software that is provided with the microscope facility such as Aperio Image scope [8,9,11]. Some other authors clearly state that their methods are semi-automated, and thus extensive human interaction is needed. Such a method has been presented by Pillete et al. [15], where the image is converted into binary and then undesired areas are eliminated applying a semi-automated technique. Fibroquant, which is introduced by Masseroli et al. [16] employs histogram equalization, and several thresholding techniques [17,18]. This methodology has been also used by Caballero et al. [19]. A recent development for automated liver biopsy image analysis has been presented by Xu et al. [20]. In this study, a new index, namely qFibrosis, is introduced, which is based on extensive image processing, analyzing three area types (portal, fibrous and septa) in the image, and subsequently extracting several features from each of them, which are used for fibrosis staging. Thus, this work attempts to combine image analysis with the philosophy of semi-quantitative scoring systems.

The proposed methodology provides a fully automated liver biopsy image segmentation and region classification to extract CPA. The methodology is based on clustering and classification algorithms, in order to avoid manual threshold setting and image areas selection. Furthermore, it is designed to operate using low-resolution images, which can be easily obtained without any specialized equipment or time consuming manual processes, and it requires low computational effort, thus consuming very little time (image acquisition and processing needs less than a minute). Finally, the proposed methodology utilizes region characterization, in order to exclude image regions of other tissue types and artifacts from CPA extraction, thus only liver tissues contribute to the computation of the ration between fibrosis pixels and liver tissue pixels. Extensive comparative results have been extracted, combining K-means (KM) or Fuzzy C-means (FCM) algorithm for clustering stages, with several different classification algorithms for region characterization.

2. Materials and methods

The proposed methodology consists of three stages. During the first stage, clustering algorithms are used to separate background pixels from pixels belonging to all other image regions (liver tissue and all types of other tissue/artifacts). As a result, an image which includes liver tissue and all types of other tissue/artifacts on zero background is produced. In the second stage, a set of shape and color features are extracted from each region in the image. These features are used for region classification in order to identify the type of each region. Each feature vector has been annotated as liver tissue or non-liver tissue, based on experts' annotation of the corresponding image region. According to classification results, image regions classified as non-liver tissue are discarded, while regions classified as liver tissue are further processed for CPA computation. CPA is assessed is the third stage of the methodology, where clustering algorithms are applied to all pixels belonging to liver tissue regions in the image, in order to group them into two groups; the one group will include fibrosis pixels, while the other will consist of pixels of normal liver tissue. The flowchart of the proposed methodology is presented in Fig. 1.

2.1. Background/Tissue separation

The first stage of the methodology is the background/tissue separation, which is implemented using a clustering approach. To reduce dimension and time complexity stage, the image is divided into 9×9 windows, beginning from the top left pixel of the image. Then, the mean value of each window from all RGB channels is calculated, thus a feature vector with three components is calculated for each window. Then, two centroids are calculated, employing a clustering algorithm on all feature vectors of the image and setting the number of clusters equal to 2 (the first corresponding to background and the second to non-background/tissue regions).

Both KM [21] and FCM [22] clustering algorithms have been tested for the first stage.KM employs a square-error criterion, which is calculated for each of the two clusters. Iteratively, each feature vector (corresponding to a 9×9 window) of the image is assigned to the nearest centroid and then both centroid are recalculated. The iterative procedure ends when the distance between the centroids of two subsequent iterations is minimized. Similarly, FCM is based on the minimization of an objective function; however, each feature vector belongs partially in both centroids according to a membership value. After the iteratively procedure and the centroids generation, all the pixels of the images are assigned to the nearest centroid according to Euclidean distance, producing the segmented image.

With the completion of the clustering step, an image with tissue regions in zero background is produced. Then, all 8-connected regions of the image are detected and labeled, thus processing an image with labeled regions in zero background. Labeled regions with less than 500 pixels are eliminated. In Fig. 2, a liver biopsy image and the respective zero background image (produced after clustering) and the labeled image, are illustrated.

2.2. Region classification

The labeled image includes all non-background regions of the image. However, labeled regions may include several types of artifacts, which must be excluded from the CPA assessment. To address this issue, a region classification procedure has been performed. The classification procedure was designed to classify all image regions as liver tissue and non-liver tissue without differentiating the actual non-liver tissue region class (i.e. blood clot, blood vessel, muscle tissue, fat, structural collagen, dye drop, or artefact), since the main idea is to separate liver tissue regions