Computerized Medical Imaging and Graphics 42 (2015) 51–55

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Heterogeneity assessment of histological tissue sections in whole slide images☆

Philippe Belhomme[a,*], Simon Toralba[a,c], Benoît Plancoulaine[a], Myriam Oger[a,b], Metin N. Gurcan[c], Catherine Bor-Angelier[a,b]

[a] Normandie Université; UNICAEN, CLCC F. Baclesse, PATHIMAGE BioTICLA EA 4656, Caen, France
[b] Pathology Department, CLCC F. Baclesse, Caen, France
[c] CIALab, Department of Biomedical Informatics, OSU, Columbus, OH, USA

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form 10 October 2014
Accepted 10 November 2014

Keywords:
Heterogeneity
Whole slide image
Breast cancer
Dimensionality reduction
Spectral graph theory

ABSTRACT

Computerized image analysis (IA) can provide quantitative and repeatable object measurements by means of methods such as segmentation, indexation, classification, etc. Embedded in reliable automated systems, IA could help pathologists in their daily work and thus contribute to more accurate determination of prognostic histological factors on whole slide images. One of the key concept pathologists want to dispose of now is a numerical estimation of heterogeneity. In this study, the objective is to propose a general framework based on the diffusion maps technique for measuring tissue heterogeneity in whole slide images and to apply this methodology on breast cancer histopathology digital images.

1. Introduction

Research in signal and image analysis is going on for many decades now and is directly linked with the exceptional development of computer technologies. But after all these years, it must be admitted that there are not so many real working applications in practice, especially in the medicine area where the expert's eye is still more accurate and faster than many automated systems dealing with large amounts of data. However, reliable automated systems could really help pathologists in their daily work as the number of pathological cases increases as far as the early screening campaigns do. We know that computerized image analysis (IA) can provide quantitative and repeatable object measurements by means of methods such as segmentation, indexation or classification. When IA is used to analyze images of histological sections in the medical research, the typical objects to be processed are nuclei, vessels, cell groups or tumors. IA allows these structures in histopathological images to be detected, to be analyzed automatically in terms of their size or shape, in order to assess their proportion or to compute their staining intensities. Indeed, IA can contribute to more accurate determination of prognostic histological factors by pathologists [1] but its action is limited to provide quantitative information as no multi-purpose method exists for segmenting or classifying images with a wide range of color or shape configurations, with which pathologists are familiar. The advent of digital scanners leads to generate whole slide images (WSI) from histological sections acquired at a full resolution. A main challenge this technology presents is the size of data to be processed which are generally very large and requires a huge storage capacity and processing needs. Conversely, the main advantage is that the complete tissue structure is easily accessed. In parallel, the aggressiveness of a cancer could result in morphological and architectural changes that can be observed in the tissue structure, and so be characterized by the object distribution on the slide, by the cross relations between objects and by the texture. This kind of information could contribute to evaluate a well-known concept: heterogeneity. Frequently addressed in signal processing, especially in terms of "entropy", but more rarely in the field of imaging [2], the objective here is to propose a framework for measuring tissue heterogeneity in WSI and apply this methodology on breast cancer histopathology digital images. The key idea in this work is to not rely on segmentation (e.g. [1,2]) of individual structures to characterize heterogeneity, but to make use of classification of squared sub-images later called 'patches'. In some previous works dedicated to the development of a computer-aided diagnosis system (CADS) based on image retrieval and classification [3,5], we have used a

☆ Special thanks to the Pathology Department, Caen, France & National Center of Pathology, Vilnius, Lithuania.
* Corresponding author. Tel.: +33 2 33 01 46 29.
E-mail address: philippe.belhomme@unicaen.fr (P. Belhomme).

http://dx.doi.org/10.1016/j.compmedimag.2014.11.006
0895-6111/

method coming from spectral graph theory, the Diffusion Maps (DM) [6], to process WSI split in small squares called 'patches.' The DM algorithm, in which eigenvalues and eigenvectors of a Markov matrix defining a random walk on the data are computed, allows to both cluster non-linear input data thanks to its inner classification properties preserving local neighborhood relationships, but also to reduce the input data dimensionality in a space (usually a 2D or 3D space) where it is therefore possible to compute euclidean distances between the objects to be analyzed [6,8]. To briefly describe the complete CADS we are developing, a first step consists in building a knowledge database involving many features extracted from a set of well-known images; this is an 'off-line' procedure conducted once. These features are represented by vectors of non-linear data acting as a signature. In a second step, signatures are obtained from new unknown images and then compared with those already present in the database; this is an 'on-line' procedure that has to be conducted each time a new image is processed. In a last step, before rendering a qualitative measure of similarity/dissimilarity between the supervised images and new unknown images, a feedback procedure would have to be processed in order to eliminate most of the artifacts that usually corrupt initial knowledge databases. But this general approach, especially with the DM algorithm, can also be derived to analyze a set of image patches coming from any WSI, with just the goal to compare their feature vectors. In this paper, we focus on a way to characterize the tissue heterogeneity at a regional level by working on the projected coordinates of image patches in a reduced 3D space.

2. Materials and methods

Images used to illustrate this study come either from histological sections or tissue microarrays (TMAs) of breast cancer stained according to the Ki67 protocol and the Hematoxylin–Eosin-Saffron protocol (HES). They are acquired at a 20× magnification on a microscopic scanner device (ScanScope CS; Aperio Technologies). The resolution of WSI is 0.5 microns/pixel and the typical image size of any histological section can reach 50,000 × 40,000 pixels, corresponding to an uncompressed file of size 5.6 Gb. They are stored in the TIFF 6.0 file format (compression 30%). In each WSI some regions of interest (ROI), chosen by the pathologist, are then split in connected blocks of size 50 × 50 pixels, called "patches", and each patch is associated with a numerical signature expressed as a feature vector. Tools developed here are written in Python language with the help of specialized modules (PIL: Python Imaging Library, SciPy-Numpy, Matplotlib, Mahotas...).

2.1. Features extraction

The feature vectors embed some statistical measures obtained from color components and some texture parameters. In this study, 61 numerical values per component were calculated from up to two color spaces (RGB for Ki67 images and RGB + H&E color deconvolution for HES images [11]) plus 18 values obtained from the excess-red component (2R-G-B). From any given component, the 9 statistical features are the mean, median, mode, Skewness, Kurtosis parameters with also the 20%–40%–60%–80% quantiles of its cumulated histogram (initially reduced to 64 values). The 52 texture features correspond to the classic 13 Haralick parameters obtained in four directions [12]. We also added to the texture features 18 new values coming from the intrinsic statistical parameters of regionalized variables in three directions (0°–45°–90°). They are derived from the nugget, sill and range of a geostatistical method [13]. Finally, each feature vector thus contains either $F = 201$ ($61 \times 3 + 18$) or $F = 323$ ($61 \times 5 + 18$) numerical values, here with the predominance of texture features in order to be quite independent from the color staining variations encountered between different laboratories (and often inside the same laboratory).

In order to later compare feature vectors, and considering the sparse numerical range of their values, the symmetric Kullback–Leibler distance [3,4] has been retained for its ability to easily manage such a case, while remaining fast to implement. The distance between two vectors p_1,p_2 of length F is given by:

$$D_{KL}(p_1, p_2) = \frac{1}{2}\sum_{j=1}^{F} p_{1j} \cdot \log\left(\frac{p_{1j}}{p_{2j}}\right) + p_{2j}\left(\frac{p_{2j}}{p_{1j}}\right) \quad (1)$$

2.2. Dimensionality reduction

In any classical CADS, one of the key components is a visualization tool showing relationships between supervised images, stored in a knowledge database, and new images that are presented to the system. Typically, these relationships may be expressed as a connected graph in a 2D or 3D space where one hopes to find distinctive clusters corresponding to histological types or sub-types. It is therefore mandatory to reduce dimensionality from F (201 or 323 in our application) to 2 or 3 dimensions. With feature vectors containing non linear data as we are faced with, authors in [7,8] have shown that it was not appropriate to perform a principal component analysis (PCA). Instead, methods based on Spectral Connectivity Analysis (SCA) such as the Diffusion Maps, involving eigenvalues and eigenvectors of a normalized graph Laplacian, are well suited for this task. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n patches that we assimilate to a fully connected graph G, that means a distance function is computed for each pair $\{x_i, x_j\}$. A $n \times n$ kernel P is obtained from a Gaussian function whose coefficients are given by:

$$p(x_i, x_j) = \frac{w(x_i, x_j)}{d(x_i)} \quad (2)$$

with

$$d(x_i) = \sum_{x_k \in X} w(x_i, x_k) \quad (3)$$

and

$$w(x_i, x_j) = e^{-\left(\frac{D_{KL}(x_i, x_j)}{\epsilon}\right)} \quad (4)$$

In fact, $p(x_i,x_j)$ may be considered as the transition kernel of the Markov chain on G. In other words, $p(x_i,x_j)$ defines the transition probability for going from x_i to x_j in one time step. The eigenvectors ϕ_k of P, ordered by decreasing positive eigenvalues, give the practical observation space axes. It must be noticed that ϕ_0 is never used since linked to the eigenvalue $\lambda = 1$ (i.e. the data set mean or trivial solution). A 3D projection can be then obtained along axes (ϕ_1,ϕ_2,ϕ_3). Choosing ε in $w(x_i,x_j)$ is an empirical task which should permit a moderate decrease of the exponential in Eq. (4); some works [8] use the median value of all $D_{KL}(x_i,x_j)$ distances whereas other works [6] use the mean distance obtained in the k nearest neighbors from a subset of X, which finally yields to quite the median value in many cases. We have retained the first solution for its simplicity and to lessen the overall computational time. Fig. 1a–c show the resulting projections obtained from the DM algorithm for 300 patches of fibroadenoma with $\varepsilon = 0.5 \times$ medianValue, $\varepsilon =$ medianValue and $\varepsilon = 2 \times$ medianValue. One can see that the general shapes of point clouds are not the same but, due to the inner classification properties of DM, we have shown that the local neighborhood relationships were preserved in terms of histological types/sub-types [9].