Belhomme et al. Diagnostic Pathology 2011, 6(Suppl 1):S3
http://www.diagnosticpathology.org/content/6/S1/S3

DIAGNOSTIC PATHOLOGY

PROCEEDINGS Open Access

Towards a computer aided diagnosis system dedicated to virtual microscopy based on stereology sampling and diffusion maps

Philippe Belhomme*, Myriam Oger, Jean-Jaques Michels, Benoit Plancoulaine, Paulette Herlin

From The 10th European Congress on Telepathology and 4th International Congress on Virtual Microscopy Vilnius, Lithuania. 1-3 July 2010

Abstract

An original strategy is presented, combining stereological sampling methods based on test grids and data reduction methods based on diffusion maps, in order to build a knowledge image database with no bias introduced by a subjective choice of exploration areas. The practical application of the exposed methodology concerns virtual slides of breast tumors.

Introduction

While pathologist population tends to dramatically dropped, the number of pathological cases to examine increases steadily (mainly due to the new screening campaigns). Fully automated image processing is able to provide a solution to this problem. Indeed, it may help pathologists in their daily practice in finding objective criteria for differential diagnosis or quantifying prognostic markers.

The recent marketing of digitizers now allows visualizing the entire histological slide at high resolution, while limiting time expense and artifacts previously encountered with image tiling methods [1]. More and more introduced in pathology departments, these systems however generate very large images which frequently exceed several Gigabytes. Because of tumor heterogeneity, it is essential to build image knowledge databases containing representative features of the various morphological types of lesions before considering implementing computer-aided diagnosis systems [2]. But, as it is almost impossible for a pathologist to manually segment such a large image, and a fortiori many of them (the estimated time being hundred hours), the current practice consists in manually selecting some 'representative areas'. A bias is then introduced in the process as this choice is obviously subjective. It is then mandatory to find wiser solutions leading to an unbiased collection of image databases. The sampling tools offered by stereology can be of great help in this context [3,4].

Systematic sampling resulting from a random starting point with a fixed periodic interval is able to reduce the area to be analyzed, while preserving the collection of varied and characteristic regions encountered in a virtual slide (VS) of a tumor. However, even if the working area is smaller, the number of selected regions can be very high and can include many redundant elements. A data reduction has then to be conducted in order to keep a proper right number of representative elements. Among these reduction methods, the diffusion maps [5,6] provide a very attractive framework for processing and visualizing huge non-linear bulk data.

This work relates to the medical image processing and retrieval field, with the goal to develop and propose a functional computer-aided diagnosis system based on a knowledge database. The original strategy exposed in this paper consists in starting from a collection of VS, then taking advantage of stereological sampling methods and diffusion maps, to finally compute a knowledge image database containing a small number of image patches that are representative of a given histological type or subtype. The practical application illustrating this framework makes use of VS of breast tumors.

* Correspondence: philippe.belhomme@unicaen.fr
GRECAN EA 1772, IFR ICORE 146, Université de Caen, France
Full list of author information is available at the end of the article

Materials

Images used for illustrating the strategy are VS of histological sections of breast tumors, stained in the same laboratory according to the Hematoxylin-Eosin-Safron protocol and acquired with the same digital scanner. The main goal here is to collect a useful number of image patches corresponding to a given histological type. Its ability to be embedded into a computer-aided diagnosis system (CADS) is illustrated by building an unbiased image database containing representative patches of a benign tumor (Fibroadenoma) and by testing the discrimination between a benign tumor and a malignant tumor (Fibroadenoma vs Comedo carcinoma). Images have been acquired at X20 (0.5 μm per pixel), using a digital slide scanner (ScanScope CS from Aperio Technologies, Inc) and then stored in TIFF 6.0 image file format with a 30% jpeg compression [7]. Their mean size is about 65000x43000 pixels^2 and each holds about 350 MB on a hard disk.

The tools needed for this study were developed in Python language (http://www.python.org) with the help of specialized modules (PIL: Python Imaging Library and SciPy: http://www.scipy.org).

Methods

Stereology

In order to reduce the expertise workload, a stereological test grid for point counting is over imposed onto VS in ImageScope viewer (Aperio Technologies, Inc) [8]. This kind of probe is usually dedicated to estimate the area and volume fractions in a tissue compartment [4]. In our application, the grid step was set to 1000x1000 pixels, ie 2800 points in an image having an average size of 65000x43000. The pathologist has to determine in which histological class must be arranged each area centered on grid points; 30 possibilities are available (breast tumor histological types and sub-types) provided by the annotation tool embedded in Aperio ImageScope. The pathologist is only asked to draw on each point a simple line selected in the overlay layer whose name corresponds to his choice. Each area is then extracted by a dedicated software at the plain resolution and stored as an uncompressed TIFF image file in order to enrich the future knowledge database. These areas (called later 'patches') are squares of size 400x400 pixels. Patch size has been chosen according to the mean size of representative structures encountered in the various histological types of breast tumors [9]. It allows pathologists to expertise only 16% of the whole VS. All patches are then analyzed and sorted in order to storing only the most representative ones. The original image name, the histological type of the patch and its coordinates in the test grid are stored in each filename, for later being used by sending SQL requests to the database.

Patch characterization

For each patch, statistical features are computed and embedded in a vector signature. All these signatures will be used in a later image retrieval process. At this stage of the study, none of the features results from segmentation. All are obtained from global measurements on patches computed on $I_1I_2I_3$ and YCh_1Ch_2 color components which are derived from the *RGB* color system according to the following formulas proposed by Ohta [10] and Carron [11]:

$$I_1 = \frac{R+G+B}{3}, I_2 = \frac{R-B}{2}, I_3 = \frac{2\cdot G-R-B}{4},$$
$$Y = \frac{R+G+B}{3}, Ch_1 = R - \frac{G+B}{2}, Ch_2 = \frac{\sqrt{3}}{2}\cdot(B-G)$$

These color components have been computed from the *RGB* histograms previously reduced to 64 values.

For a given color component whose histogram is called *H*, the computed features are: *H*, *H* reverse sorted, cumulative *H*, 20%, 40%, 60% and 80% quantiles of cumulative *H*, meanH, medianH, modeH, SkewnessH, KurtosisH, PearsonModeSkewnessH, that is a total of 13 data. Three of them are themselves vectors of 64 values, but will provide a single feature after distance measurements between two signatures. Definitions of these statistical features can be found in [12]. With the resulting 5 effective color components (as $Y=I_1$), 65 distance measures will be taken into account but 1010 values will be stored in the signature vector for each patch. Considering the sparse numerical range of features in signatures, the Kullback-Leibler symmetrical distance has been retained for its ability to manage such values, while remaining simple and fast to implement (compared to Mahalanobis or earth mover's distance for example). The symmetric Kullback-Leibler distance between two vectors p_1,p_2 of length n is defined by:

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

The computation time can be reduced using:

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

In order to give the same weights to histogram features (h_i) and scalar features (x_i), the Kullback-Leibler distance is averaged by the number of histogram values while comparing h_1 and h_2. Because of the symmetry of D_{KL}, and with N images to process, the computation time is proportional to $(N\times N-N)/2$ and is parallelized on multi-core/multi-processor computers.

Data reduction

The ultimate goal of this study is to contribute to the development of a computer-aided diagnosis system (CADS) whose one component should be a visualization tool dedicated to knowledge image databases. This tool would be useful for pathologists if results can be visualized in a 2D or possibly 3D space. It is therefore necessary to reduce dimensionality from n (65 dimensions in our example) to 2 or 3. The patches signatures do not necessary contain linear data. Therefore it is not appropriate to perform a principal component analysis (PCA). Belkin [5] and Coifman [6] have shown that methods based on diffusion maps, involving eigenvalues and eigenvectors of a normalized graph Laplacian, are well suited to non linear data.

Let $X=\{x_1,x_2,...,x_N\}$ be a set of N patches. A (NxN) kernel P is obtained whose coefficients are:

$$p(x_i,x_j)=\frac{w(x_i,x_j)}{d(x_i)} \quad \text{where} \quad d(x_i)=\sum_{x_k\in X} w(x_i,x_k)$$

and $w(x_i,x_j)=e^{-(D_{KL}(x_i,x_j)/\varepsilon)}$

The eigenvectors ϕ_k of P, ordered by decreasing eigenvalues, give the axes of the new observation space. It must be noticed that ϕ_0 is never used since linked to the eigenvalue λ=1 (ie the mean of the data set). The projection is then done in (ϕ_1, ϕ_2) for a 2D space or (ϕ_1, ϕ_2, ϕ_3) for 3D. The choice of ε is empirical but should permit a moderate decrease of the exponential; the median of D_{KL} distances is usually chosen [13].

Results

To illustrate the data reduction algorithm, 4 VS coming from different pathological cases have been selected; their storage needs 1,5 GB. A total number of 2967 patches, classified as Fibroadenoma by a pathologist, have been extracted from a stereological test grid. Figure 1 shows their projection in a 2D space. In this reduced space, a classical Euclidean distance can be applied to estimate the similarity between two different patches.

Keeping only 100 representative elements thanks to a regular decimation along the original curve, one obtains a new set of patches to be stored in the knowledge database. Their 2D projection is illustrated in Figure 2. These patches represent the unbiased reference, to which new 400x400 areas, extracted from unknown VS, should be compared.

To illustrate the comparison between a benign and a malignant tumor, these 100 patches obtained from the Fibroadenoma class, were compared to 64 patches extracted from a Comedo carcinoma. Figure 3 exhibits

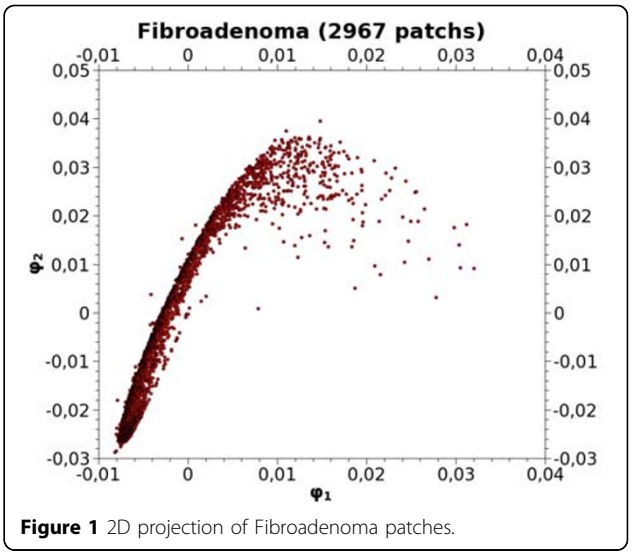

Figure 1 2D projection of Fibroadenoma patches.

the 2D projection of the overall 164 patches. The good discrimination of these two families is obvious, especially according to the axis ϕ_1=0 which usually provides the sharpest separation between object classes [13].

Conclusion

This work relies on an original strategy starting from VS and leading to an unbiased knowledge image database containing reference patches of breast tumors. We have shown that combining stereological sampling and data reduction based on diffusion maps offers an interesting general framework for this purpose. Once the sequence of procedures has been implemented, the only parameters to be tuned are the choice of image features to use for patch signatures, the size of patches and their

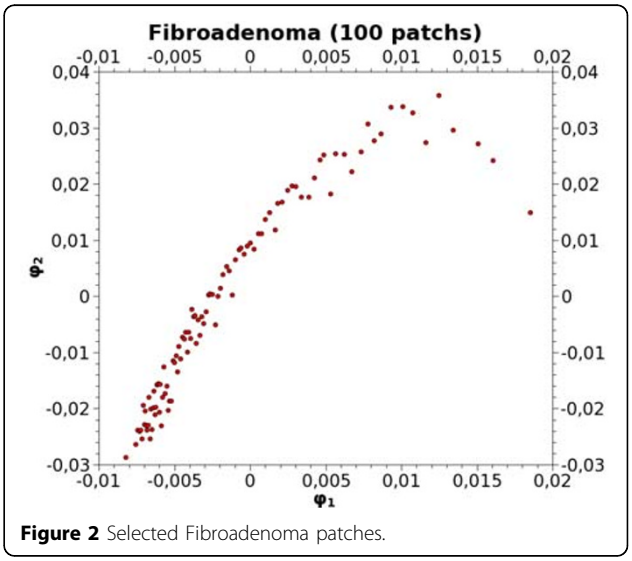

Figure 2 Selected Fibroadenoma patches.

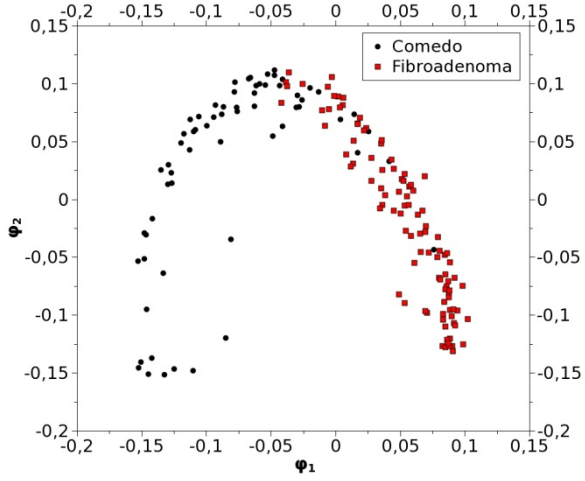

Figure 3 2D projection of Fibroadenoma vs Comedo carcinoma (164 patches).

number to be kept in the knowledge database. At this stage of the study, none of the features come from any segmentation or from any texture measurement. When all steps will be validated, it will be time to consider advantages of adding new parameters or to introduce other color components. Using patches acquired at a lower resolution should be also an interesting issue, depending on the histological type or subtype to be studied. It should be noticed that, up to now, we did not try to adjust parameters in order to be independent from the acquisition conditions, since all images come from the same origin. However, the final goal being the development of a CADS for several laboratories, it should be necessary to take this into account, by computing International Color Consortium profiles for each device used along the process, starting from histological staining up to image acquisition [14].

The results illustrated in this study are preliminary ones. Up to now, 400 high resolution VS of breast tumors are available. The benign and malignant tumors were classified into 30 histological types and subtypes. We plan now to project these 30 classes in the same 3D space, in order to analyze their scattering. This work is in progress in our laboratory.

Acknowledgements
This article has been published as part of *Diagnostic Pathology* Volume 6 Supplement 1, 2011: Proceedings of the 10th European Congress on Telepathology and 4th International Congress on Virtual Microscopy. The full contents of the supplement are available online at http://www.diagnosticpathology.org/supplements/6/S1.

Competing interests
The authors declare that they have no competing interests.

Published: 30 March 2011

References

1. Ortiz JPG, Ruiz V, Garcia I: **Virtual Slide Telepathology Systems with JPEG2000, EMBS 2007.** *29th Annual International Conference of the IEEE* 2007, 880-883.
2. Kayser K, Radziszowski D, Bzdyl P, Sommer R, Kayser G: **Towards an automated virtual slide screening: theoretical considerations and practical experiences of automated tissue-based virtual diagnosis to be implemented in the Internet.** *Diagnostic Pathology* 2006, **1**:10, doi: 10.1186/1746-1596.
3. Elias H: **Stereology.** *Proceedings of the Second International Congress for Stereology* Chicago, New York: Springer-Verlag; 1967.
4. Baddeley A, Jensen EB: **Stereology for Statisticians.** *Chapman and Hall/CRC* 2005.
5. Belkin M, Niyogi P: **Laplacian eigenmaps for dimensionality reduction and data representation.** *Neural Computation* 2003, , **15:** 1373-1396.
6. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker S: **Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps.** *Proceedings of the National Academy of Sciences* 2005, **102(21)**:7426-7431.
7. [http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf].
8. Herlin P: **Computer-Assisted Stereology for Pathology Applications.** *Science Webinar series* 2009 [http://www.aperio.com].
9. Oger M, Belhomme P, Klossa J, Michels JJ, Elmoataz A: **Automated region of interest retrieval and classification using spectral analysis.** *Proceedings of 3rd International Congress on Virtual Microscopy* Toledo, Spain; 2008.
10. Ohta Y-I, Kanade T, Sakai T: **Color Information for Region Segmentation.** *Computer Graphics and Image Processing* 1980, **13(3)**:222-241.
11. Carron T: **Segmentation d'images couleur dans la base Teinte-Luminance-Saturation : approche numérique et symbolique.** Thèse de doctorat, Université de Savoie; 1995.
12. [http://mathworld.wolfram.com/topics/Moments.html].
13. Oger M: **Indexation automatique d'images numériques : application aux images histopathologiques du cancer du sein et hématologiques de leucémies lymphoïdes chroniques.** Thèse de doctorat, Université de Caen Basse-Normandie; 2008.
14. Kayser K, Borkenfeld S, Gortler J, Kayser G: **Image standardization in tissue–based diagnosis.** *Diagnostic Pathology* 2010, **S13, doi:10.1186/1746-1596-5-S1-S13.**

doi:10.1186/1746-1596-6-S1-S3
Cite this article as: Belhomme *et al.*: Towards a computer aided diagnosis system dedicated to virtual microscopy based on stereology sampling and diffusion maps. *Diagnostic Pathology* 2011 **6**(Suppl 1):S3.