Technology and Health Care 8 (2000) 291–297
IOS Press

Content-based retrieval of breast cancer biopsy slides

F. Schnorrenberg[a], C.S. Pattichis[a,*], C.N. Schizas[a] and K. Kyriacou[b]
[a]*Department of Computer Science, University of Cyprus, P.O. Box 20537, CY-1678 Nicosia, Cyprus*
E-mail: {csfranks, pattichi, schizas}@ucy.ac.cy
[b]*Department of Electron Microscopy, The Cyprus Institute of Neurology and Genetics, P.O. Box 23462, Nicosia, Cyprus*

Accepted 4 September 2000

Abstract. The Biopsy Analysis Support System (BASS), previously used for image analysis of immunohistochemically stained sections of breast carcinoma, has been extended to include indexing and content-based retrieval of biopsy slide images from a database of 57 captured cases. Images from histopathological biopsy slides are described and these are accessed in terms of the properties of either individual nuclei or groups of cell nuclei present in the slide. Visual similarity of cases is specified in terms of a diagnostic index, commonly known as the H-score, which incorporates the heterogeneity of nuclear staining intensity, as well as the percentage of nuclei staining at specific intensities. The system provides a platform that can be exploited in telepathology and teleconsultation, but further research is needed to explore its full potential and accuracy in a diagnostic clinical environment.

Keywords: Breast cancer biopsy, content-based retrieval of biopsy slide images, digital medical libraries, optical imaging

1. Introduction

Content-based access to image and video databases is already a reality [1–4]. Recent research work shows that depending on the image data and the desired search mode some approaches are more appropriate than others [5]. In particular, in the medical domain global properties of an image, while at times simple and efficient to compute [6], are often less important compared to structural and semantic descriptions of image content [7]. However, reasoning over structural and semantic details or objects in an image requires specialized functionality for each image domain. Examples where the local structure is important are magnetic resonance (MRI), X-Ray imagery, and biopsy images. The objective of this paper is to present the Biopsy Analysis Support System (BASS) as a means to perform indexing and content-based retrieval of biopsy slide images from a database of captured cases.

In histopathology, breast cancer biopsy slides, that are immunohistochemically stained for steroid receptors, are evaluated for calculating diagnostic indices such as the H-Score. This is a time consuming procedure involving examination of slides by a human expert, namely a histopathologist using a light microscope. The H-Score gives a semi-quantitative assessment of the status of prognostic factors, such

*Corresponding author: Constantinos S. Pattichis, Department of Computer Science, University of Cyprus, P.O. Box 20537, CY-1678 Nicosia, Cyprus. Tel.: +357 2 892244; Fax: +357 2 339062; E-mail: pattichi@ucy.ac.cy.

as estrogen and progesterone receptors, as visualized in breast cancer nuclei populations. The H-Score is computed by classifying nuclei according to staining intensity and then estimating the proportions of nuclei with different staining intensities in the biopsy slide. Commercial computer-aided systems compute their own H-Score as a function of total nuclei area and the optical density. These systems, however, although dedicated, do not use the manual approach of counting individual nuclei, but rather work with specific areas of stained nuclei where global thresholding techniques are used for discriminating between positive and negative [8–11]. In contrast, the BASS system was designed to detect and classify individual nuclei, thus facilitating content based biopsy image retrieval.

Inherent variabilities stemming in part from the use of a variety of assessment schemes, and their dependence on the judgment of a particular expert or the choice of parameters in an algorithm, limit the validity of comparisons. While accurate and detailed biopsy assessment is crucial for determining the most appropriate mode of treatment and the estimated survival of patients, it is difficult to compare assessment results between laboratories or even between different experts. The opinion of multiple experts is often desired, but frequently hard to obtain due to limitations in time and human resources in a single location. Advances in telepathology make it possible for multiple experts, in geographically distributed locations to assess digitized images of specimen [12] or even control microscopes [13]. Furthermore, efforts are underway through the EU project European Pathology Assisted by Telematics for Health (EUROPATH) to develop quality assurance criteria and a database of standardised gynecological cancer biopsy cases (http://europath.imag.fr).

In this paper, we propose content-based retrieval of breast cancer biopsy slide images from a database of cases using BASS. With BASS, intensity images, from stained biopsy sections, can be selectively retrieved from a database based on visual features such as occurrence of a nuclei class, nuclear heterogeneity, and H-Score. The system provides a platform that can be further exploited in telepathology and teleconsultation.

2. Materials and methods

The complete BASS system is illustrated in Fig. 1. It consists of the following three modules: (i) detection of nuclei, (ii) nuclei classification and biopsy scoring, and (iii) retrieval interface. The nuclei detection and classification modules have already been published in [14–16]. The system was developed in IDL data analysis and visualization software (http://www.rsinc.com/). The analysis of one image takes on average less than 25 seconds (450 MHz Intel Pentium PC, 64M bytes RAM).

2.1. Database of cases

A database of cases (DBC) (see Fig. 1) was generated using 41 breast cancer biopsy slides (57 images, over 8300 nuclei) immunolabelled for the prognostic factors estrogen and progesterone receptors (ERICA-kit, Abbot, UK) and counterstained with haematoxylin to contrast positively stained nuclei (brown color) with negative nuclei (blue color). Medical experts selected and digitized up to three regions of interest from each slide at $\times 400$ magnification (Zeiss Axiophot microscope, SONY DXC-980P camera) in 24 bit color and 640*480 pixel spatial resolution. Additional information for each case included a unique registration number (CASENO), which can be used to cross-reference a case in other databases external to BASS, and an H-Score which was manually assigned in the laboratory routine assessment (LABSCORE).

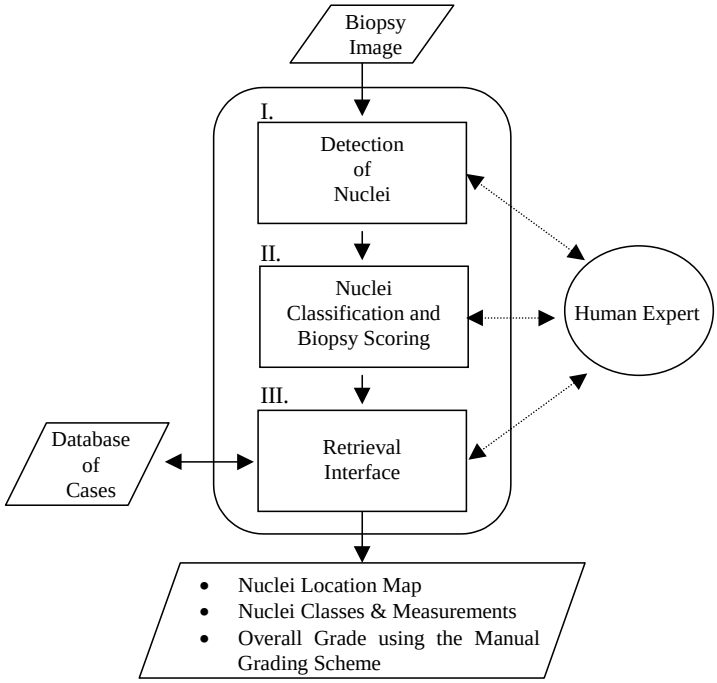


Fig. 1. Diagram of the Biopsy Analysis Support System (BASS).

2.2. Detection of nuclei (*Fig. 1, I*) *[14,15]*

This module is implemented in two stages: (i) standardization, and (ii) adaptive nuclei detection. The standardization stage yields an image which has been corrected to account for illumination differences across the optical field due to the light microscope. The digitized biopsy image is pixelwise and channelwise divided by a background image which in turn was obtained using a blank biopsy slide. All digitized images are transformed from RGB color space to the YIQ color space. In YIQ space the luminance (Y channel) and the chromaticity (I, Q channels) of a color image are clearly separated. BASS uses the luminance image for the detection of nuclei, while for the classification module it utilizes features based on all three channels.

BASS automatically detects individual breast cancer nuclei using receptive fields. The input to the detection stage consists of the luminance channel of the original color image. The image is repeatedly convoluted and the individual intensity values are remapped (squashed) using an on-center-off surround type receptive field filter and a logistic squashing function. After less than five iterations the histogram of the image develops a characteristic bimodal shape, which is used to determine a threshold to obtain a binary image. The centers of the remaining disconnected image structures are determined using morphological operations. These centers are interpreted as the nuclei center locations. The performance of the detection algorithm was compared to that of two experts. The sensitivity and positive predictive value of BASS were found to be 83% and 67.4% respectively.

2.3. Nuclei classification and biopsy scoring (*Fig. 1, II*) *[16]*

BASS's classification stage uses a widely accepted manual grading scheme for accurately assessing staining intensity and heterogeneity of the nuclei populations in the biopsy [17]. According to the grading

Table 1
Content-based query

a. Find *matching* cases	b. Find *similar* cases
SELECT(	*SELECT*(
FROM= dbc,	FROM= dbc,
MATCH(	LIKE= dbc([<IDX>]),
CASENO= [‘1678-92’, . . .],	*USING*(
TYPE= [‘ER’, . . .],	SCORE_DEV= 0/1/2/3
NAME= [‘substring’, . . .],	HETERO= [‘LT’, ‘GT’, ‘EQ’]
LABSCORE= [−1/0/1, −1/0/1, −1/0/1, −1/0/1, −1/0/1],	NUCCLASS= [−1/0/1, −1/0/1, −1/0/1,
SYSSCORE= [−1/0/1, −1/0/1, −1/0/1, −1/0/1, −1/0/1],	−1/0/1, −1/0/1]))
NUCCLASS= [0/1, 0/1, 0/1, 0/1, 0/1]	
HETERO= [‘LT num.’, ‘GT num.’, ‘EQ num.’]))	

‘−1’ indicates ‘no’, ‘0’ stands for ‘don’t care’, and ‘1’ symbolizes ‘yes’.

scheme nuclei are assigned to one of five staining intensity classes (negative, weak, moderate, strong, very strong). Depending on the proportions of the classified nuclei, one final H-Score out of five possible values (0, 1+, 2+, 3+, 4+) is assigned to the biopsy. The input to the classification stage consists of six-dimensional feature vectors. The feature vectors are based on a fixed number of pixels in a small circular neighborhood of each nuclei center location. The features are optical density, two chromaticity values, a variance based texture measure, and the average optical density and variance of all nuclei in the biopsy image. A neural network which obtained 72% correct classification score on the database of cases was implemented. The output of this module describes the content of each biopsy image.

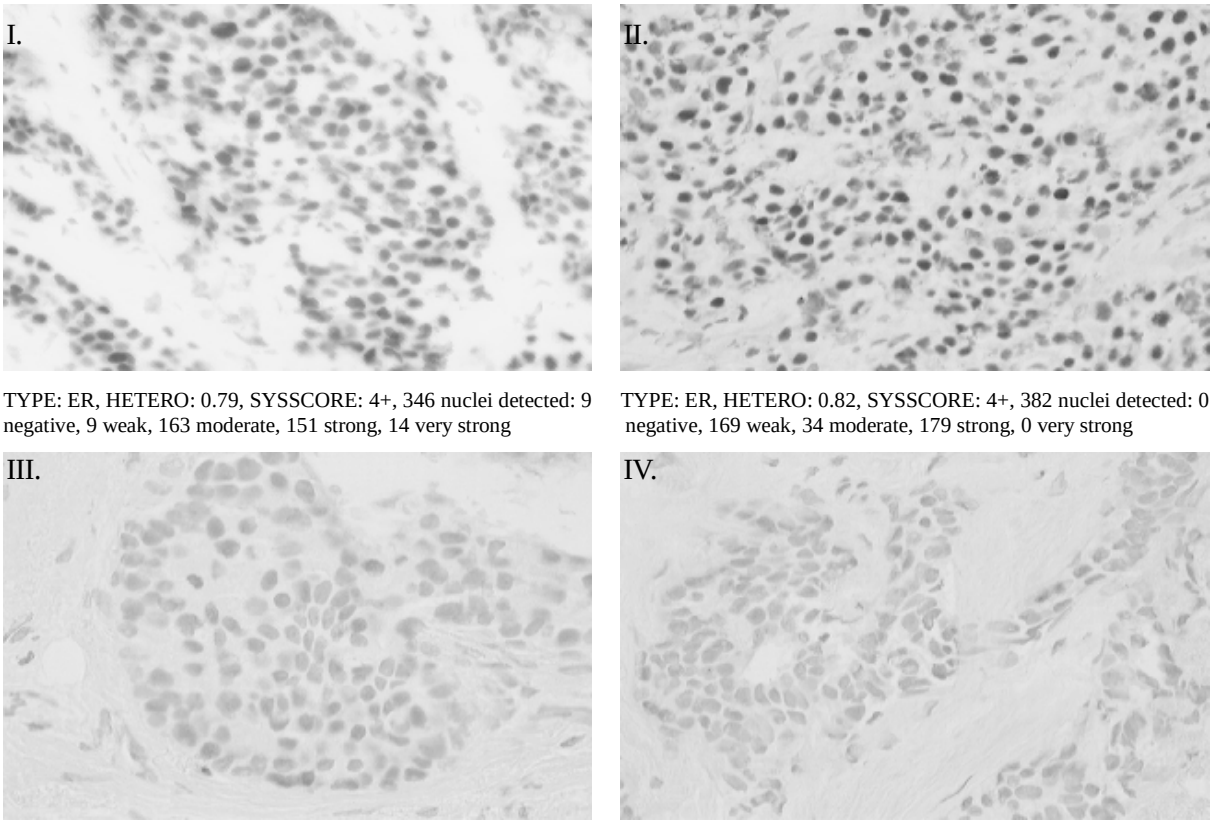
2.4. Retrieval interface (*Fig. 1, III*)

All features used for comparing two biopsy slides are precomputed using BASS in either an interactive or automated fashion. Thus, the actual comparison of cases can be performed without reading any of the images except for display purposes. After the computed features and classification results have been stored in the database of cases, queries can be formulated (Table 1).

The scope of queries that a database of breast biopsy slides has to address is of a perceptual and task-specific nature. The queries take the form of textual compound expressions as illustrated in Table 1. The expressions contain feature value lists and ‘don’t care’ values, based on the above mentioned assessment scheme which classifies nuclei into 5 classes and distinguishes 5 different H-Scores. To find matching cases, one can specify (see Table 1(a)) the unique case number (CASENO), the type of stain (TYPE), the name (NAME), the H-Score manually obtained in the laboratory routine (LABSCORE), the H-Score assigned using BASS (SYSSCORE), occurrence of nuclei classes (NUCCLASS), and a heterogeneity coefficient (HETERO), which is based on a normalized variance measure of all members of the nuclei classes in each image. Similar images can be retrieved by using the construct specified in Table 1(b). A range of biopsy images can be specified to find other images which match, given a maximal deviation from the H-Score (SCORE_DEV), heterogeneity coefficient (HETERO), and occurrence of nuclei classes (NUCCLASS).

3. Results

A database of cases was generated using 41 breast cancer biopsy slides. The cases were selected to cover the full range of the diagnostic index, H-score, which varies from 0 to 4+, including at least 5



TYPE: ER, HETERO: 0.79, SYSSCORE: 4+, 346 nuclei detected: 9 negative, 9 weak, 163 moderate, 151 strong, 14 very strong

TYPE: ER, HETERO: 0.82, SYSSCORE: 4+, 382 nuclei detected: 0 negative, 169 weak, 34 moderate, 179 strong, 0 very strong

TYPE: ER, SYSSCORE: 2+, 156 nuclei detected: 54 negative., 93 weak, 9 moderate, 0 strong, 0 very strong

TYPE: ER, SYSSCORE: 1+, 237 nuclei detected: 24 negative, 213 weak, 0 moderate, 0 strong, 0 very strong

Fig. 2. Results of database queries given in Table 2(a) and (b). Image (I) and image (II) match the query with TYPE=['ESTROGEN'], syscore=[0,0,0,0,1], HETERO>0.75. Image (III) is similar to image (IV) according to the query with TYPE=['Estrogen'], SCORE-DEV=[1], and NUCCLASS=[1,1,0,−1,−1].

Table 2
Query examples: a. Finding matching cases, b. Finding similar cases

a. Find *matching* cases	b. Find *similar* cases
SELECT(	*SELECT*(
FROM= dbc,	FROM= dbc,
MATCH(	LIKE= dbc([18]),
TYPE= ['ER'],	*USING*(
SYSSCORE= [−1, −1, −1, −1, 1],	SCORE_DEV= 1
HETERO= [GT .75]))	NUCCLASS= [1, 1, 0, −1, −1]))

'−1' indicates 'no', '0' stands for 'don't care', and '1' symbolizes 'yes'.

representative cases for each diagnostic index. This was done in order to ensure that BASS received maximum learning and training information, from routinely available cases. Once a biopsy slide that covered the entire range of the diagnostic index was analyzed using BASS, the results were stored in an image independent data-structure. This data-structure can easily be accessed to retrieve and compare any features relevant for a content-based query.

Sample queries supported by the system are given and their results are illustrated in Fig. 2. The query

in Table 2(a) demonstrates a search for matching cases, i.e. images, of staining type 'ER' (estrogen) to which BASS has assigned an H-Score of 4+ (SYSSCORE=[−1,−1,−1,−1,1]) and a heterogeneity coefficient greater than 0.75 (HETERO=[GT.75]). The response of the system is illustrated in Fig. 2, I, and II. In Table 2(b), a sample query is shown, whose goal it is to find similar cases, i.e. images. In particular, the query will flag cases which contain negative nuclei, weakly stained nuclei, but not strongly stained, nor very strongly stained nuclei (NUCCLASS=[1,1,0,−1,−1]), and which may differ up to 1 point in the H-Score (SCORE_DEV=1). The response of the system is given in Fig. 1, III, and IV.

The original images used for Fig. 2 are color images. Gradations in nuclei staining intensity, which are important for classification of nuclei, cannot satisfactorily be resolved in greylevel. Thus, only coarse differences among nuclei populations are visible.

4. Discussion and conclusion

BASS provides a general interface for retrieval and characterization of biopsy slides. Visual similarity of cases can be specified in terms of an H-Score, maximal deviation from the H-Score, heterogeneity of nuclei staining intensity, and occurrence of particular nuclei classes. This is used in the semi-quantitative assessment of the biological features of tumour cells such as the presence of steroid receptors within tumour nuclei. This information is currently required by clinicians since it influences the choice of therapy of individual cancer patients.

Currently, BASS is a specialized tool to analyze nuclei populations in breast cancer biopsy slides. It was previously shown that the detection and classification of individual nuclei in histopathological sections can be reliably performed by the BASS modular neural network system in an accurate and consistent manner [14–16]. BASS also facilitates interaction with experts and such interaction is constructive, since it was demonstrated that the modules correctly detect additional numbers of nuclei which were not initially detected by the experts [15].

The present pilot study was based on the analysis of 41 biopsy slides (57 images, over 8300 nuclei) and involved three histopathologists and two computer experts. Despite the rather limited number of cases used, BASS has proven useful in the following two areas which are of importance in diagnostic histopathology:

a) in creating an unbiased, knowledge based, diagnostic system that performs with good accuracy, minimizing errors due to interobserver and intraobserver variations, and can thus be used as an additional independent expert, and
b) in enabling the efficient identification and retrieval of similar cases ensuring uniformity and standardization of reporting procedures.

In an attempt to improve objectivity and offer rapid analysis speeds some commercial systems, like CAS [18] and SAMBA [19], rely on global discrimination of structures of interest, between nuclei in this case, and background. These systems measure percent stained surface area using global thresholding techniques [8–11]. However, there is disagreement among experts about the optimal selection of global thresholds, the choice being fixed, manual, and automatically set thresholds. BASS avoids the need for global thresholding and area measurements, since it detects, counts, and classifies individual nuclei according to the manual semi-quantitative diagnostic index. Future extensions of this work should include the description of spatial structure of the tumour in terms of the location of tumor nuclei and the application of BASS to semi-quantitatively evaluate the presence of other prognostic factors. The latter

include tumour suppressor genes, or proliferation indices of neoplastic cells in various types of tumours. Furthermore, BASS has the potential to be exploited on a larger scale.

In conclusion it should be stated that content-based access to biopsy slide image databases as provided by BASS, is very useful in telepathology and teleconsultation. Moreover, the world wide web provides a forum for reinforcing standardized assessment of medical images beyond dedicated telemedicine networks. In the case of specialized biopsy slide image databases, BASS not only facilitates convenient access, but can be used as a reference point or indeed as a second local or distant expert.

Acknowledgment

This work is supported through a grant from the research committee of the University of Cyprus, and by a grant from the Cyprus Planning Bureau and the General Secretariat of Research and Technology of Greece.

References

[1] Special issue on content-based image retrieval, *IEEE Computer* **28** (1995), 9.
[2] Special Issue on digital libraries in medicine, *Inter. J. on Digital Libraries* **1** (1997), 3.
[3] The Virage project, http://www.virage.com/.
[4] The IBM QBIC project, http://www.qbic.almaden.ibm.com/.
[5] R.W. Picard, A Society of Models for Video and Image Libraries, MIT Media Laboratory, Perceptual Computing Section, Technical Report No. 360, 1996.
[6] M.J. Swain and D.H. Ballard, Indexing via color histograms, *Image Understanding Workshop,* Pittsburgh, PA, USA, 1990, pp. 623–630.
[7] S.C. Orphanoudakis, C. Chronaki and S. Kostomanolakis, I^2C: A system for indexing, storage, and retrieval of medical images by content, *Journal of Medical Informatics* **19**(2) (1994), 109–122.
[8] G. Brugal, Color processing in automated image analysis for cytology, in: *Quantitative Image Analysis in Cancer Cytology and Histology*, J.Y. Mary and J.P. Rigaut, eds., Amsterdam: Elservier, 1985, pp. 19–33.
[9] S. Bacus and J.L. Flowers, The evaluation of estrogen receptor in primary breast carcinoma by computer-assisted image analysis, *American Journal of Clinical Pathology* **90** (1988), 233–239.
[10] C. Charpin, P.M. Martin, B.D. Victor, M.N. Lavaut, M.C. Habib, L. Andrac and M. Toga, Multiparametric study (SAMBA 200) of estrogen receptor immunocytochemical assay in 400 human breast carcinomas: Analysis of estrogen receptor distribution heterogeneity in tissues and correlations with dextran coated charcoal assays and morphological data, *Cancer Research* **48** (1988), 1578–1586.
[11] A.E. Dawson, Jr. Austin and D.S. Weingerg, Nuclear grading of breast carcinoma by image analysis, *American Journal of Clinical Pathology* **95**(1) (1991), S29–S37.
[12] M. Oberholzer, H.R. Fischer, H. Christen, S. Gerber, M. Brühlmann, M.J. Mihatsch, T. Gahm, M. Famos, C. Winkler, P. Fehr, H.J. Hosch and L. Bächtold, Telepathology: frozen section diagnosis at a distance, *Virchows Arch.* **426** (1995), 3–9.
[13] M.J. O'brien and A.V. Sotnikov, Digital imaging in anatomic pathology, *American Journal of Clinical Pathology* **106**(1,4) (1996), S23–S32.
[14] F. Schnorrenberg, C.S. Pattichis, K. Kyriacou and C.N. Schizas, Computer-aided detection of breast cancer nuclei, *IEEE Transactions on Information Technology in Biomedicine* **1**(2) (1997), 128–140.
[15] F. Schnorrenberg, N. Tsapatsoulis, C.S. pattichis, C.N. Schizas, S. Kollias, M. Vassiliou, A. Adamou and K. Kyriacou, Improved detection of breast cancer nuclei using modular neural networks, *IEEE Eng. in Med. and Biol.* **19**(1) (2000), 48–62.
[16] F. Schnorrenberg, C.S. Pattichis, K. Kyriacou, M. Vassiliou and C.N. Schizas, Computer-aided classification of breast cancer nuclei, *Techology and Health Care* **4**(2) (1996), 147–161.
[17] S. Störkel, T. Reichert, K.A. Reiffen and W. Wagner, EGFR and PCNA expression in oral squamous cell carcinomas: A valuable tool in estimating the patients prognosis, *European Journal of Cancer* **29B** (1993), 273–277.
[18] Cell Analysis Systems Inc., Cell analysis systems: Quantitative estrogen progesterone users manual, Application Version 2.0, Catalog number 210325-00, USA, April 1990.
[19] Alcatel TITN Answare, IMMUNO 4.00: User's guide, First Edition, Grenoble, France, October 1993.