Artificial Intelligence in Medicine 64 (2015) 131–145

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

An unsupervised feature learning framework for basal cell carcinoma image analysis

John Arevalo[a], Angel Cruz-Roa[a], Viviana Arias[b], Eduardo Romero[c], Fabio A. González[a,*]

[a] Machine Learning, Perception and Discovery Lab, Systems and Computer Engineering Department, Universidad Nacional de Colombia, Faculty of Engineering, Cra 30 No 45 03-Ciudad Universitaria, Building 453 Office 114, Bogotá DC, Colombia
[b] Pathology Department, Universidad Nacional de Colombia, Faculty of Medicine, Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia
[c] Computer Imaging & Medical Applications Laboratory, Universidad Nacional de Colombia, Faculty of Medicine, Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia

ARTICLE INFO

Article history:
Received 1 October 2014
Received in revised form 9 April 2015
Accepted 15 April 2015

Keywords:
Digital pathology
Representation learning
Unsupervised feature learning
Basal cell carcinoma

ABSTRACT

Objective: The paper addresses the problem of automatic detection of basal cell carcinoma (BCC) in histopathology images. In particular, it proposes a framework to both, learn the image representation in an unsupervised way and visualize discriminative features supported by the learned model.

Materials and methods: This paper presents an integrated unsupervised feature learning (UFL) framework for histopathology image analysis that comprises three main stages: (1) local (patch) representation learning using different strategies (sparse autoencoders, reconstruct independent component analysis and topographic independent component analysis (TICA), (2) global (image) representation learning using a bag-of-features representation or a convolutional neural network, and (3) a visual interpretation layer to highlight the most discriminant regions detected by the model. The integrated unsupervised feature learning framework was exhaustively evaluated in a histopathology image dataset for BCC diagnosis.

Results: The experimental evaluation produced a classification performance of 98.1%, in terms of the area under receiver-operating-characteristic curve, for the proposed framework outperforming by 7% the state-of-the-art discrete cosine transform patch-based representation.

Conclusions: The proposed UFL-representation-based approach outperforms state-of-the-art methods for BCC detection. Thanks to its visual interpretation layer, the method is able to highlight discriminative tissue regions providing a better diagnosis support. Among the different UFL strategies tested, TICA-learned features exhibited the best performance thanks to its ability to capture low-level invariances, which are inherent to the nature of the problem.

1. Introduction

Digital pathology refers to the set of computational methods and technologies that support the different pathology workflow stages, including digital slide acquisition, computer aided diagnosis, prognosis and theragnosis [1]. The importance and popularity of digital pathology have rapidly grown during the last years thanks to the emergence of fast, cost-effective whole slide image acquisition systems. An important component of digital pathology is automatic image analysis, which is fundamental for tasks such as automatic tumor detection and grading [2]. Automatic image analysis encompasses different kinds of computer vision and pattern recognition problems associated with the detection, segmentation and classification of biological structures (pathological and non-pathological).

The success of any histopathology image analysis method depends on how well it captures morphological and architectural characteristics from nuclei, cells, glands, organs and tissues. In turn this depends on how well the method characterizes the visual content of the histopathology image. This characterization is accomplished by a feature extraction process which typically uses canonical (e.g. wavelet transforms) or hand-engineered features (e.g. SIFT). Different visual features have been proposed and extensively evaluated in different histopathology image analysis problems: (1) object level features to characterize biological structures, e.g. size and shape, radiometric and densitometric, texture, chromatin-specific; (2) spatially related features to represent

---

* Corresponding author. Tel.: +57 1 3165322; fax: +57 1 3165491.
E-mail addresses: jearevaloo@unal.edu.co (J. Arevalo), aacruzr@unal.edu.co (A. Cruz-Roa), vlariasp@unal.edu.co (V. Arias), edromero@unal.edu.co (E. Romero), fagonzalezo@unal.edu.co (F.A. González).

http://dx.doi.org/10.1016/j.artmed.2015.04.004
0933-3657/

the architectural arrangement of cells or other structures, usually derived from graph-based representations and descriptors, e.g. Voronoi tessellation, Delaunay triangulation, minimum spanning graph, connected graph, relative neighbor graph and others [2,3]; (3) multi-scale feature extraction algorithms, e.g. multi-resolution, pyramid and hierarchical representation of images among others [2,3]. Mostly, feature choice is handmade and based on the particularities of the problem at hand. However, recent investigations in computer vision suggests that strategies which learn the image representation directly, and automatically, from image collections may produce better representations resulting in better performance of automatic analysis algorithms [4]. This approach, known as *unsupervised feature learning* (UFL), has been applied to different computer vision and pattern recognition problems with great success [5].

This paper explores the application of UFL strategies to the representation and automatic analysis of histopathology images. In particular, the paper presents a framework for basal cell carcinoma (BCC) image analysis which is able to learn an appropriate representation from a representative set of images for automatic carcinoma detection. The experimental evaluation shows that the proposed framework outperforms state-of-the-art BCC histopathology image representation and classification methods. The main contributions of this work are:

- a novel hybrid method combining the state-of-the-art visual representations and learning techniques to integrate UFL in a complete data-driven approach;
- a strategy to endow the classification method with a visual interpretation layer, which exploits both the hierarchical representation of the model, to highlight those regions of the image that contribute to a higher degree to the model prediction, and the topographic organization of the learned representation, to identify visual patterns associated with tumor and non-tumor images; and
- a systematic evaluation of different UFL strategies on BCC histopathology images, which shows that, in this particular problem, learned features outperform state-of-the-art canonical representations.

The rest of the paper is organized as follows: Section 2 makes a review of the relevant literature, Section 3 describes the unified unsupervised learning framework for BCC histopathology image representation and analysis; Section 4 presents the results of experimental evaluation for BCC detection. Finally, conclusions are presented in Section 5.

2. Related work

Nowadays, the explosion of big data in pathology research is bringing many opportunities and challenges that have given birth to a novel research area known as *digital pathology* [1]. This is possible thanks to the massification of histology slide scanners as part of the pathology routine, the increase on the number of publicly available histology and histopathology image databases, and the development of virtual slide navigation systems [6,7]

Digital or computational pathology is a new emerging area that investigates a complete probabilistic treatment of scientific and clinical workflows in general pathology, i.e. it combines experimental design, statistical pattern recognition and survival analysis within a unified framework to answer scientific and clinical questions in pathology [8]. This area is faced with big challenges such as automatic diagnosis and objective quantification of disease's indicators for personalized diagnosis [1]. Typically, tumor detection and grading are the main goals in pathology digital slide analysis [2]. The main goal of tumor detection is to differentiate between healthy and tumor tissues to support diagnosis. The goal of grading is to quantify architectural and morphological signatures within tumor to classify the grade of aggressiveness, which is very important to determine the appropriate treatment for each patient. These two problems are mainly addressed combining two different, but complementary strategies: (1) hand-crafted features to capture chromatin, morphological and architectural characteristics from nuclei, glands and tissues, and (2) machine learning methods to induce models that use the visual features for detection, segmentation, and classification of biological structures or types of cancer.

There is an extensive literature in automatic histopathology image analysis. Gurcan et al. [3] performed a comprehensive review of state-of-the-art in histopathology image analysis in different pathology problems with 133 references. There, the authors presented the works organized according to type of segmentation, feature extraction, and classification algorithms. For instance, for feature extraction they describe: (i) object-level features to characterize biological structures (e.g. size, shape, radiometric, densitometric, texture, and chromatin-specific [9]); (ii) spatially related features to represent the architectural arrangement of cells or other structures mainly graph-based representations and descriptors, (e.g. Voronoi tessellation, Delaunay triangulation, minimum spanning graph, connected graph, relative neighbor graph, and k-NN [9–12]); (iii) multi-scale feature extraction algorithms (e.g. multi-resolution, pyramid and hierarchical representations [13–15]); and (iv) a common feature selection or dimensionality reduction techniques (e.g. independent component analysis, linear discriminant analysis, principal component analysis, and manifold learning [16–19]). In the case of machine learning algorithms applied to histopathology image analysis, the methods have to deal with large, highly dense datasets, multiple scales and visual features motivating the use of ensemble of classifiers more than a single classifier for tumor detection and grading, e.g. Adaboost learning, graph-based classifiers, and kernel methods like support vector machines [20]. More recently, He et al. [2] presented a new overview of the state of the art in histopathology image analysis by organizing the above different stages and techniques for each of the four most common carcinomas: cervix, prostate, breast and lung. Particularly, the present work focuses in BCC, the most common type of skin cancer. In this sense, automatic image analysis of BCC had been explored previously under traditional approaches by integrating image representation and machine learning methods.

2.1. Representation learning

Representation learning have recently drawn a lot of attention thanks of the impressive results in different computer vision and pattern recognition tasks [5]. In general, representation learning encompasses different methods, most of them based on neural networks, that combine linear and non-linear transformations of the data, with the goal of yielding more abstract and ultimately more useful representations. These methods have rapidly increased in popularity and attracted researchers attention due to a remarkable string of empirical successes, both in academy and in industry, beating traditional approaches in each application domain, such as speech recognition, object recognition and image classification, with significant breakthrough results [4,21–24].

These approaches have already been successfully used in histopathology image analysis tasks. Malon et al. [25] were one of the firsts proposing to use convolutional neural networks (CNN), one of the most popular type of supervised representation learning models, to count mitotic figures in breast cancer histology images, to recognize epithelial layers in the stomach, and to detect signet ring cells. CNN had been also used in segmentation and classification of histopathology images for breast cancer cell