Computerized Medical Imaging and Graphics 35 (2011) 629–645

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Ensemble based system for whole-slide prostate cancer probability mapping using color texture features

Matthew D. DiFranco [a,*], Gillian O'Hurley [b], Elaine W. Kay [c,b], R. William G. Watson [d], Padraig Cunningham [a]

[a] School of Computer Science and Informatics, University College Dublin, Ireland
[b] Royal College of Surgeons in Ireland, Dublin, Ireland
[c] Department of Pathology, Beaumont Hospital, Dublin, Ireland
[d] School of Medicine and Medical Science, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 15 June 2010
Received in revised form 31 October 2010
Accepted 17 December 2010

Keywords:
Ensemble classification
Prostate cancer detection
Random forest feature selection
Histology
Digital pathology

ABSTRACT

We present a tile-based approach for producing clinically relevant probability maps of prostatic carcinoma in histological sections from radical prostatectomy. Our methodology incorporates ensemble learning for feature selection and classification on expert-annotated images. Random forest feature selection performed over varying training sets provides a subset of generalized CIE$L^*a^*b^*$ co-occurrence texture features, while sample selection strategies with minimal constraints reduce training data requirements to achieve reliable results. Ensembles of classifiers are built using expert-annotated tiles from training images, and scores for the probability of cancer presence are calculated from the responses of each classifier in the ensemble. Spatial filtering of tile-based texture features prior to classification results in increased heat-map coherence as well as AUC values of 95% using ensembles of either random forests or support vector machines. Our approach is designed for adaptation to different imaging modalities, image features, and histological decision domains.

1. Introduction

Clinical assessment and prognosis of prostate cancer in patients who undergo radical prostatectomy relies on the evaluation of whole-slide histology sections of prostatic tissue under a microscope. This assessment relies on pathologist interpretation of the Gleason grading scale, a system devised in 1966 to differentiate tumor grades using tissue architecture [1]. Since that time, the landscape of prostate cancer has changed dramatically due to developments such as prostate specific antigen screening, transrectal ultrasound imaging, multi-core biopsies, immunohistochemistry to detect basal cells, and surgical improvements that enable whole-organ resection [2,3]. These advances have led to the increased detection of early stage disease, which has in turn led to an increase in radical prostatectomy procedures. Modifications have been made to Gleason grading to keep pace with clinical understanding [4], yet pathologist errors and inter-observer variability have been consistently shown to be problematic under various conditions [5–11].

Gleason grading is initiated at low microscopic resolutions (4–10×) where regions of suspected carcinoma can often be detected [12]. Haemotoxylin and eosin staining (H&E) provides tissue-type differentiation through the interpretation of color-based architectural features. Tissue architecture differences are interpreted by the observer as textural variations which can often be detected at low-resolutions. Pathologists then perform differentiation and validation of Gleason grades through the examination of morphological and cellular features at higher resolutions.

Using traditional microscopy, clinically available image information is limited by the capabilities of the microscope and the pathologist's own level of knowledge and experience. The emergence of digital pathology is removing these limitations by replacing the microscope with computing, which in turn enables fast, complex and novel processing of microscopic images using existing capabilities from image processing, machine vision, content-based image retrieval, data mining, and human–computer interaction.

As digital slide technologies evolve and costs decrease, whole-slide imaging (WSI) will become integrated into the clinical workflow, augmenting or possibly replacing optical microscopic analysis for certain tasks. At the same time, the potential exists for digital scanning of existing glass-slide archives for subsequent development of complex retrospective studies.

* Corresponding author. Tel.: +43 660 142 7331.
E-mail addresses: matthew.difranco@ucd.ie, mdifranc@gmail.com (M.D. DiFranco), padraig.cunningham@ucd.ie (P. Cunningham).

0895-6111/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compmedimag.2010.12.005

The resolution at which WSI data can be effectively captured, stored and processed represents a frontier in microscopic digital image analysis. For example, when a pathologist requires high-magnification viewing of images (e.g. 40× and above), a digital scan captured at 20× becomes of little use and a light microscope is required. However, an image captured and stored at 100× would provide the required magnification and remove the need for traditional slide viewing.

However, whole-slide images present technological challenges due to their size at full resolution. A whole-section prostate slide scanned at 40× magnification contains billions of pixels. Image compression techniques make it possible to store these images at manageable sizes (<1 GB), but whole-image processing is at present inefficient in terms of time and computing resource utilization. Automated tile-based techniques offer a method for high-volume WSI processing and analysis which incorporates expert input (digital annotations), distributed processing environments (cloud computing and parallel processing), and state-of-the-art machine vision techniques for dimensionality reduction, tumor detection, and data visualization.

In this paper, we present a system for producing reliable and clinically relevant heat-maps of prostate tumor presence probability (see Figs. 15 and 16) using a sparsity of data taken from a limited number of whole-slide training images. Visualization is performed through ensemble classification of tiles in query images. The ensemble of classifiers is built using expert-annotated tiles from training images, and probability scores for the cancer presence are calculated from the responses of each classifier in the ensemble.

2. Background

2.1. Review of existing prostate cancer detection methods

The Gleason scoring system relies on the architecture of prostatic tissue patterns, and the majority of work for the computer aided detection to date has relied in part on architectural or morphological features [13–16]. Extraction of such features generally requires the segmentation or localization of tissue structures such as glandular lumen and cell nuclei [17–19]. Approaches which explicitly identify tissue structures before extracting image features seek to translate the tissue patterns seen in digital images into a representation which fits the language of the Gleason guidelines. However, the generalization and reproducibility of such approaches are hindered by the heterogeneity and variety of prostatic carcinoma (PCa) patterns in histopathology images.

Approaches using textural features have also been applied for histological detection and grading of prostate cancer. Such features include Gabor filters [15,20,21], fractal dimension [22], and wavelet-based features [23]. In their use of fractal dimension features, Huang and Lee [22] report Gleason grading correct classification rate (i.e. classification accuracy) of up to 94.6% but use only 205 images of size 512 × 384 pixels and rely on only red-channel intensities for calculation of texture features. We demonstrate through feature selection analysis that red-channel texture features do not generalize well on larger datasets.

Jafari-Khouzani and Soltanian-Zadeh [23] perform nearest neighbor classification of Gleason grades 2–5 using multiwavelets, wavelet packets and co-occurrence matrices. They report accuracy as high as 97% using multiwavelets, while the best performance with co-occurrence features is 84%. However, optimum results are obtained using (k)-nearest neighbor where $k = 1$, and cross-validation folds allow train and test data to come from the same images, which could lead to over-fitting as shown by DiFranco et al. [24]. In our methodology, we split our test and training sets at the image level to ensure that tiles are not classified using models in which they are included for training.

Prior work has also focused on classification between tumor and non-tumor tissue. Diamond et al. [13] demonstrate tile-based prostate tumor detection using textural and morphological features and report accuracy as high as 88.9%. However, the classifiers or validation schemes implemented are not described. Monaco et al. [16] implement a gland-segmentation approach coupled with probabilistic pairwise Markov models and report sensitivity of 87% and specificity of 90%. Doyle et al. [20] implement an ensemble approach using a hierarchical version of AdaBoost. They report an overall accuracy of 88%. Tahir and Bouridane [14] implement another ensemble technique, round-robin classification, which transforms an (n)-class problem into $n \times (n-1)$ two-class problems, on multi-spectral prostate histopathology images. They couple round-robin classification with an optimization scheme known as tabu search and report accuracy of 98–100% for the classification of prostatic carcinoma versus three non-tumor tissue patterns.

2.2. Ensemble classification and feature selection

Ensemble learning is a technique whereby the outputs of multiple classifiers are used to improve overall classification accuracy [25,26]. The decision of an ensemble can be computed from the output of each classifier using a majority voting scheme [27], and a probability for the ensemble decision can be measured using information from the decisions made by each component classifier. The classifiers that make-up an ensemble represent a collection of experts, all with varying expertise and specialization based on the data and variables on which they have been trained. A clinical parallel would be a consensus panel composed of experts with varying background and experience.

Random forest is an ensemble technique which produces accurate classification results by building many classification and regression tree (CART) classifiers on randomly selected partitions of a data set [28]. Given a set of (N) examples represented by (M) attributes, a bootstrapping approach is used to build (T) training sets of size (N), chosen with replacement from the original (N) examples, where (T) is the number of decision trees in the random forest. The remaining examples, known as *out-of-bag* data, are used as a test set at each tree.

Randomness is further introduced at the nodes of each decision tree. For a given forest, a feature subset size $m \ll M$ is chosen. At each node in a tree (m), features are selected at random and an optimal split of the training data is found for those (m) features based on a Gini split criterion. Trees are grown in this way without pruning and to as large an extent as possible. Examples are classified by majority voting of the classification outputs of all trees in the forest.

The result is an ensemble which randomly varies both the features and the training examples at each node of each tree. In addition, randomly selected feature subsets offer a means of assessing feature interactions. Performance of random forest can be assessed using the error on out-of-bag examples, thereby removing the need for separate train and test data sets. Parameters to set for random forests include the number of trees (*nTree*) to grow and the number of features *mTry* on which to split at each node of the tree.

2.3. Random forest feature selection

Random forest classification also provides a mechanism for embedded feature importance evaluation [28,29]. Like other tree-based ensemble methods, importance can be assessed naively by counting instances when variables are selected over all trees in the forest. More sophisticated approaches assess the impact of indi-