Pattern Recognition 42 (2009) 1104 – 1112

Contents lists available at ScienceDirect

Pattern Recognition

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

Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection

Akif Burak Tosun[a], Melih Kandemir[a], Cenk Sokmensuer[b], Cigdem Gunduz-Demir[a,*]

[a]Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey
[b]Department of Pathology, Hacettepe University Medical School, Ankara TR-06100, Turkey

ARTICLE INFO

Article history:
Received 1 December 2007
Received in revised form 5 June 2008
Accepted 16 July 2008

Keywords:
Cancer detection
Image segmentation
Histopathological image analysis
Object-oriented texture
Colon adenocarcinoma

ABSTRACT

Staining methods routinely used in pathology lead to similar color distributions in the biologically different regions of histopathological images. This causes problems in image segmentation for the quantitative analysis and detection of cancer. To overcome this problem, unlike previous methods that use pixel distributions, we propose a new homogeneity measure based on the distribution of the objects that we define to represent tissue components. Using this measure, we demonstrate a new object-oriented segmentation algorithm. Working with colon biopsy images, we show that this algorithm segments the cancerous and normal regions with 94.89 percent accuracy on the average and significantly improves the segmentation accuracy compared to its pixel-based counterpart.

1. Introduction

In the current practice of medicine, histopathological examination of biopsies is the most commonly used method to locate and classify diseases including cancer. In cancer diagnosis, pathologists visually examine the changes in cell morphology and tissue distribution under a microscope and determine whether a biopsy contains any malignant (cancerous) region and, if so, the cancer type and its malignancy level (grade). However, as it mainly relies on the visual interpretation, this examination may lead to a considerable amount of subjectivity, especially in cancer grading [1,2]. To reduce this subjectivity, it has been proposed to use computational methods that rely on the quantification of a tissue by defining mathematical features [3–7]. Although the very first step in this quantification is the segmentation of a tissue image into homogeneous regions, these studies have not mainly focused on this problem and have extracted features from the tissue image assuming that it is homogeneous. Nevertheless, besides having many heterogeneous regions in a tissue image, the existence of such regions and their ratio help pathologists determine the cancer grade.

It has been proposed to segment an image into homogeneous regions by either connecting adjacent pixels or locating edges according to a homogeneity measure. Numerous studies have defined the homogeneity based on the color information of pixels and/or the spatial relations between the colors (i.e., texture information of pixels) [8,9]. A large subset of them quantizes the pixels of an image into clusters using the color information alone and considers the connected pixels of the same cluster as a homogeneous region [10–15]. One common method to find such clusters is to employ the color histogram of the image. For example, Park et al. [11] perform morphological operators on the histogram, while Shafarenko et al. [12] apply the watershed algorithm to the histogram to detect the clusters. Besides employing the color histogram, different approaches (such as fuzzy [13,14] and genetic [15] approaches) are also used to obtain the clusters. Another subset of studies proposes to use spatial information of pixels in addition to their color information [16–19]. The color and the texture information could be used consecutively or together in segmentation. For example, the JSEG algorithm proposed by Deng and Manjunath [17] first uses pixel color information to quantize them into clusters, without considering their spatial relations. Following the quantization, they define a homogeneity criterion to quantify the texture of the color-quantized pixels. In contrast, in the algorithm proposed by Chen et al. [19], the texture is extracted from the gray-scale of the image and it is used together with the color information to obtain the overall segmentation. As they commonly result in oversegmentation, both of these approaches merge the segmented regions according to their color similarity. These previously proposed segmentation algorithms lead to promising results especially when there is a considerable amount of difference in the color distribution of adjacent regions. However,

* Corresponding author. Tel.: +90 312 290 3443; fax: +90 312 266 4047.
E-mail addresses: tosun@cs.bilkent.edu.tr (A.B. Tosun), melih@cs.bilkent.edu.tr (M. Kandemir), csokmens@hacettepe.edu.tr (C. Sokmensuer), gunduz@cs.bilkent.edu.tr (C. Gunduz-Demir).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.07.007

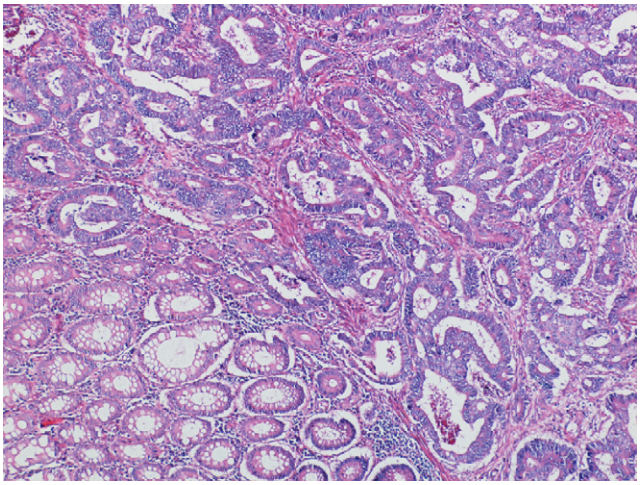

Fig. 1. A histopathological image stained with the hematoxylin-and-eosin technique.

due to the routinely used staining techniques, the color distribution of different regions is very similar for histopathological images. For instance, Fig. 1 shows such a histopathological image stained with the hematoxylin-and-eosin technique, which is routinely used to stain a biopsy in hospitals. As observed in this figure, this image consists of two regions in which the color distributions are very similar (here the upper right of the image contains tumorous regions whereas its lower left contains healthy regions).

In this paper, we introduce a new homogeneity measure and demonstrate a new segmentation algorithm that uses this homogeneity measure to segment a biopsy image, for which the color distribution is very similar in its heterogeneous regions. Our proposed approach relies on the quantification of the spatial relations between the tissue components (e.g., epithelial tissue components, connective tissue components, and luminal structures). For this purpose, we define different types of "objects", which represent different components of a tissue, and make use of the distribution of these objects as well as their spatial relations to define our homogeneity measure. As opposed to the existing algorithms, which rely on pixel-based information (pixel colors and/or pixel-based textures), the proposed segmentation algorithm uses object-based information. Working with colon biopsy images, we demonstrate that our object-oriented segmentation algorithm yields 94.89 percent accuracy on the average and significantly improves the accuracy in locating tumorous regions and other non-cancerous tissue transformations compared to its pixel-based counterpart.

The remainder of this paper is organized as follows. In Section 2, we first provide our method to identify the objects in a tissue image, then describe our homogeneity criterion based on the spatial relations between the objects, and last explain our segmentation algorithm. Subsequently, we present our experiments and discuss their results in Section 3. Finally, we provide a summary of our work and discuss a future research perspective in Section 4.

2. Object-oriented textural segmentation

In a biopsy image, biologically different parts of the tissue are characterized with the spatial organizations of its cellular and connective tissue components. These organizations show differences, depending on the organ from which the tissue is taken. In this work, we focus on colon biopsy images where the tissue components are organized to form glandular structures. In these images, epithelial cells are lined up around a luminal structure, forming a crypt (gland), and lymphoid cells take place in between these crypts. This organization deviates from its regular structure due to the existence of cancer. Furthermore, this deviation is aggravated with the increasing malignancy level. Thus, for the detection of cancer and its malignancy level, the regions containing such tumorous structures should be distinguished from those that contain the normal ones.

In this work, we propose a segmentation algorithm that uses the fact that the structural organization of a tissue (i.e., the spatial distributions of cellular and connective tissue components with respect to each other) changes with the existence of cancer. To this end, we define our homogeneity measure based on the texture of these components. We define primitive objects to represent these components, instead of exactly identifying their locations, since this localization brings about a more difficult segmentation problem even for human experts.

2.1. Object definition

For defining objects, we first run the k-means algorithm on the color intensities of pixels and quantize the pixels into three clusters. These clusters correspond to purple regions (for epithelial and lymphoid cell components), pink regions (for connective tissue components), and white regions (for both luminal structures and connective tissue components[1]). Subsequently, we locate circular primitives on the pixels of each cluster after eliminating small holes and regions in the cluster. Since there are different ways of locating circles on a group of pixels, we make use of a heuristic in which the circular components are iteratively located. For a given set of pixels $\mathscr{P}=\{x_i\}$, this algorithm works as follows:

Step 1: It assigns each particular pixel x_i to the largest possible circle that includes this particular pixel x_i and that is formed by only the pixels $x_j \in \mathscr{P}$.

Step 2: It forms connected components $\mathscr{C}=\{C_1, C_2, \ldots, C_N\}$ from the pixels such that the connected component C_k consists of the pixels that are assigned to the circle k. In this step, it also eliminates the connected components smaller than an area threshold.

Step 3: For each component C_k, it recursively calls Steps 1 and 2 considering only the pixels of this connected component (i.e., in Step 1, $\mathscr{P}$ will be a set of pixels belonging to the component C_k) until there is no change in the pixels of the component. Note that there will be no change when a component is circular.

In our work, we run this iterative algorithm twice for each cluster. In the first run, we consider all of the pixels of the cluster and find the circular primitives. Then in the second run, we consider all of the pixels that belong to the same cluster but not belong to any of the circles found in the first run. Finally, the primitives obtained from the first and the second runs are merged. In Fig. 2, the results of iterations in the first and the second runs along with the final result are shown on a small image.

After this two-step iterative algorithm, we group the circular primitives of each cluster into two object types depending on their sizes. Here the circular primitives with a size smaller than a threshold are defined as one type of the object and those with a size greater than the threshold are defined as the other type. Therefore, at the end of this step, six different object types (depending on both the size of a primitive and the cluster that it belongs to) are defined for the tissue components of a given image. In Fig. 3(b), the object-map for

[1] Here "connective tissue components" is used as a general term, which represents the other components rather than cell nuclei and luminal structures. Pink-like color usually corresponds to the components such as cytoplasm, muscularis mucosa, and cell plate whereas white-like color usually corresponds to postfixative artificial defects in the connective tissue and materials secreted by neoplastic cells such as mucin.