Journal of Biomedical Informatics 43 (2010) 924–931

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A new pivoting and iterative text detection algorithm for biomedical images

Songhua Xu [a,b], Michael Krauthammer [b,*]

[a] Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN 37831, USA
[b] Department of Pathology, Yale University School of Medicine, CT 06511, USA

ARTICLE INFO

Article history:
Received 10 November 2009
Available online 29 September 2010

Keywords:
Text detection
Histogram analysis for text detection
Pivoting and iterative text region detection
Biomedical image mining

ABSTRACT

There is interest to expand the reach of literature mining to include the analysis of biomedical images, which often contain a paper's key findings. Examples include recent studies that use Optical Character Recognition (OCR) to extract image text, which is used to boost biomedical image retrieval and classification. Such studies rely on the robust identification of text elements in biomedical images, which is a non-trivial task. In this work, we introduce a new text detection algorithm for biomedical images based on iterative projection histograms. We study the effectiveness of our algorithm by evaluating the performance on a set of manually labeled random biomedical images, and compare the performance against other state-of-the-art text detection algorithms. We demonstrate that our projection histogram-based text detection approach is well suited for text detection in biomedical images, and that the iterative application of the algorithm boosts performance to an *F* score of .60. We provide a C++ implementation of our algorithm freely available for academic use.

1. Background

1.1. Introduction

Biomedical literature mining is concerned with transforming free text into a structured, machine-readable format, to improve tasks such as information retrieval and extraction. Recent work indicates that there is much interest to also consider image information when mining research articles, as images often depict the results of experiments, and sum up a paper's key findings. There are several obstacles when mining image information. First, there are many different types of images, such as graphs, gel electrophoresis and microscopy images, diagrams or heat maps. There exists no image publication standard, neither with regard to image resolution, or image file format (images are stored at different resolutions, and in a variety of file formats, such as jpeg, tiff etc.). Also, there are no explicit image design guidelines, even though authors seem to follow some universally accepted norms when creating figures such as box plots, heatmaps or gel electrophoresis images.

A unifying element across all biomedical images is image text, i.e. text characters that are embedded in images. Text in images serves several purposes, such as labeling a graph, representing genes in a heat map images, or proteins in a pathway diagram. We have previously shown that extracting image text, and making it available to image search, improves biomedical image retrieval [1]. In this work, we are concerned with optimizing the performance of a critical step in image text extraction—locating text regions in images, which is known as *text detection* in studies on image processing and Optical Character Recognition (OCR).

Generally speaking, text detection is a crucial step in processing textual information in biomedical images. For example, properly finding the text regions is the first stage of a standard OCR pipeline for extracting image text. Determining the location of text is also important for high-level image content understanding, as it is the text location that indicates the meaning of certain image text element, such as the label of the *x*- versus *y*-axis in a graph. Practical applications aside, in this paper, we are exclusively concerned with optimizing the performance of text detection, which is a fundamental research problem in image text processing.

In this work, we introduce a new text detection algorithm suited for biomedical images. We also discuss the methodological details in creating a gold standard biomedical image text detection corpus, and the use of the corpus for evaluating the performance of our algorithm. During the development of the corpus, we laid down clear guidelines on what exactly constitutes an image text region (or element) and how to manually mark the image region linked to the string. We then compared our algorithm against three existing state-of-the-art text detection methods. Even though our algorithm can in principle be applied for processing all images types, it is especially beneficial for images embedded in biomedical publications. Compared to other disciplines, biomedical authors tend to use distributed and nested text in their images in order to annotate experiment settings, conditions and results.

* Corresponding author at: Department of Pathology, Yale University School of Medicine, 300 Cedar Street, New Haven, CT 06510, USA. Fax: +1 203 785 3644.

E-mail addresses: xus1@ornl.gov (S. Xu), michael.krauthammer@yale.edu (M. Krauthammer).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.09.006

1.2. Related work

1.2.1. Image text detection algorithms

First, we are going to briefly look at prior work on image processing algorithms for image text detection, which is concerned with separating image text elements from other elements in an image. Ohya et al. [2] presented an algorithm for text detection from scene images. In their work, they first detect character components according to gray-level differences and then match the results to standard character patterns captured in a database. Their method is very robust to the font, size and intensity variation in the image texts, but is not able to deal with color and orientation changes. To address the text detection problem for color images, Zhong et al. [3] introduced a connected component-based method for locating texts in a complex color image. Their method analyzes the color histogram of the RGB space to detect text regions. Jung [4] introduced a neural network based approach for identifying text in color images. To attack the text detection problem for texts with different orientations and other distortions, Messelodi and Modena [5] described the use of low level image features such as density and contrast to detect image texts, with the ability to deal with skew in the image text. Hasan and Karam [6] also proposed a morphological approach for image text detection, which is robust to the presence of noise, text orientation, skew and curvature.

There is a body of work using advanced texture and graph segmentation methods to detect text in images. For example, Jain and Karu [7] introduce a method for learning texture discrimination masks for image text detection. Jain and Zhong [8] used a learning based approach to detect image text through image texture analysis. Wu et al. [9] introduced a system for image text detection and recognition, which adopts a multi-scale texture segmentation scheme. In their method, a collection of second-order Gaussian derivatives are used to detect candidate text regions, followed by a *K*-means clustering process and a multi-resolutional stroke generation, filtering and aggregation process to further refine the detected text region. Felzenszwalb and Huttenlocher [10] proposed a graph-based image segmentation algorithm for efficiently separating textual elements from graphical elements in an image. Their algorithm can automatically adapt itself to the image structure variation. Liu et al. [11] proposed a novel method for text detection and segmentation through using stroke filters for text polarity assessment in analyzing features in local image regions.

There also exists a growing collection of work on text detection from videos or motion images, which are closely related to the image text detection problem studied in this paper. For example, Li et al. [12] used a hybrid neural network and projection profile analysis based approach to detect and track text regions in a video. Antani et al. [13] applied a variety of text detection methods and then fused the individual text detection results together to achieve a robust text detection for videos. Kim et al. [14] introduced a support vector machine based approach for image text detection in videos. Lyu et al. [15] proposed a coarse-to-fine localization scheme for detecting texts in multilingual videos. Recently, Qian et al. [16] proposed a discrete cosines transform coefficients based method for text detection in compressed videos. Despite the many commonalities between the video text and image text detection problems, one of the main differences between them is that frame images in a video demonstrate temporal coherence, which offer much useful information for text detection. Such clues are not present in still images, and hence make the image text detection problem more challenging than its counterpart in videos.

1.2.2. Biomedical image processing algorithms and systems

Our study is related to other projects in biomedical image processing. For example, Shatkay et al. [17] used image features for text categorization. Tulipano et al. [18] studied the use of natural language processing to index and retrieve molecular images. Qian and Murphy [19] described an algorithmic system for accessing fluorescence microscopy images via image classification and segmentation.

In our own prior work [1], we discussed a novel approach for biomedical image search based on OCR. We have shown that the approach offers additional advantages compared to searching over image captions alone, notably the retrieval of additional and relevant images. The current study is closely linked to that project, discussing the algorithmic details for detecting image text regions.

2. Approach

2.1. Overview

An overview of our method is shown in Fig. 1. An input image (i.e. an image from a biomedical publication) undergoes detection of layout lines and panel boundaries, which are excluded from the image to increase text detection robustness. We implement the algorithm proposed by Busch et al. [20] for detecting these layout elements. The image is then converted to black and white, and subjected to an edge detection algorithm. The resulting edge image is then subjected to a pivoting text region detection (PTD) algorithm for extraction of text regions. PTD is repeated several times, in order to divide detected text regions into text sub-regions. If no more text regions are detected, the algorithm exits. Our algorithm is based on traditional histogram analysis-based text region detection, which takes edge images as input. We extend the traditional approach as follows: We perform a pivoting procedure while applying the histogram analysis, and repeat the procedure until no more text (sub)regions are detected.

2.2. Traditional histogram analysis-based text region detection

One of the most popular and well known text region detection methods is through analyzing the vertical and horizontal projection histograms of an image. More concretely, given an input image, we first detect the edge pixels in the image. Then a vertical and a horizontal projection histogram are derived. It is assumed that text regions generally exhibit higher density of edge pixels than non-text regions. The vertical and horizontal histograms will thus show the highest density of edge pixels in text areas. A density threshold defines the exact dimensions of the text area along the vertical and horizontal histogram. The elements of this basic procedure are discussed in more detail in the next section.

One distinct feature of many biomedical images is that they often employ a distributed and nested text layout. Figs. 3a and 4a show two typical examples, where text is distributed across many different image regions. Also, text regions often display some degree off nestedness. For example, the numbers along the x axis in Fig. 4b can be grouped in one large text area, or-more correctly-into separate (inner) text areas surrounding each individual number (Fig. 4d). The traditional histogram-based analysis technique does not cope well with distributed and nested text layout. To address this problem, we introduce a new iterative pivoting histogram analysis procedure for text region detection.

2.3. Pivoting text region detection (PTD)

We introduce a pivoting step into the classical histogram-based text detection algorithm in order to account for the distributed nature of biomedical image text. The pivoting procedures subdivides image regions into its text subcomponents, instead of identifying large text blocks. Our procedure is realized through analyzing the histograms of the input image region following the vertical and