Image and Vision Computing 38 (2015) 33–51

Contents lists available at ScienceDirect

Image and Vision Computing

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

Document image binarization using local features and Gaussian mixture modeling☆

Nikolaos Mitianoudis *, Nikolaos Papamarkos

Image Processing and Multimedia Laboratory, Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece

ARTICLE INFO

Article history:
Received 26 October 2013
Received in revised form 19 January 2015
Accepted 8 April 2015
Available online 29 April 2015

Keywords:
Binarization
Handwritten documents
Historic documents
Classification
Background estimation

ABSTRACT

In this paper, we address the document image binarization problem with a three-stage procedure. First, possible stains and general document background information are removed from the image through a background removal stage. The remaining misclassified background and character pixels are then separated using a Local Co-occurrence Mapping, local contrast and a two-state Gaussian Mixture Model. Finally, some isolated misclassified components are removed by a morphology operator. The proposed scheme offers robust and fast performance, especially for both handwritten and printed documents, which compares favorably with other binarization methods.

1. Introduction

Document images commonly arise from historical documents, books or printed documents that are digitized using a scanning device. The advancement of imaging devices, such as scanners and digital cameras, has widely facilitated the digitization of paper-printed material, including historical documents and books. Many libraries throughout the world, such as the British Library in London, UK,[1] have digitized books, manuscripts and other printed material from their collection, which are available online as images. We can extract the text information from these document images using Optical Character Recognition (OCR) techniques. Nevertheless, to enhance the performance of OCR algorithms, a number of preprocessing steps are systematically applied, including page skew detection, artifact and noise removal, document page layout analysis and document image binarization [1–4]. In this paper, we address the problem of background removal and document image binarization.

Scanned documents often contain undesired textual noise, such as specks, dots, black borders, lines, and hole-punch marks. *Background estimation* and removal is a preparatory step that enhances the quality of the document images and is beneficial for binarization techniques [5–8]. For example, historic document images often suffer from different types of degradation that render document image binarization and character recognition very challenging tasks. In summary, the main objective of background removal techniques is to remove all these degradations from a document image and enhance the discrimination of characters from the page background.

After the original document images have been enhanced, the output of most document processing systems is a bi-level image containing characters and background. Image binarization can then be performed either on a global or a local basis. Conventional binarization techniques of gray-scale documents were initially based on global thresholding algorithms (clustering approaches) [9], which have proved to be efficient in converting simple gray-scale images into a binary form but are inappropriate for complex documents, and degraded documents. For this purpose, the local binarization techniques of Niblack [10], Sauvola [11] and Bernsen [12] have been extensively used by the document image processing community. There are numerous specialized binarization techniques for document images (see [13] for a more detailed review). Here, we will outline several important binarization methods that have appeared so far.

In [1], Papamarkos proposed a neuro-fuzzy technique for binarization and gray-level (or color) reduction of mixed-type documents. Badekas and Papamarkos [13] proposed a binarization technique that combines the results of multiple binarization algorithms using a Kohonen Self-Organizing Map (KSOM) neural network. In [14], the binarization results of many independent techniques were initially produced and then combined with a Kohonen Self-Organizing Map (KSOM). Badekas et al. [15] also introduced a binarization technique, specialized for color documents, where the resulting “binary” image contains the detected text regions with black characters in white background leaving the remaining original color parts of the document intact. In [16], Makridis and Papamarkos introduced a two-stage approach to image binarization. The first stage included a background

☆ This paper has been recommended for acceptance by Seong-Whan Lee.
* Corresponding author.
E-mail address: nmitiano@ee.duth.gr (N. Mitianoudis).
[1] http://www.bl.uk/aboutus/stratpolprog/digi/digitisation/.

http://dx.doi.org/10.1016/j.imavis.2015.04.003
0262-8856/© 2015 Elsevier B.V. All rights reserved.

removal technique that was based on fixed-size median filtering of the document image. Once the background was removed, the second stage aimed at creating 2D clusters of neighboring pixels of similar intensity, i.e., document characters and background. Binarization was then performed by identifying 2 clusters (text-background) using the multithresholding technique of Reddi et al. [17].

Gatos et al. [18] (GPP method) estimated the document background by an adaptive threshold which labels each pixel as either text or background. To estimate the background surface, they used Sauvola's binarization algorithm to roughly extract the text pixels and calculated the background surface from them by interpolation of neighboring background pixels intensities. For the other pixels, background surface is set to the gray level of the original image. Ntirogiannis et al. [19] proposed a modular system for handwritten document binarization. Background is initially estimated via an inpainting procedure starting from the Niblack binarization output. The background estimate is then normalized to smooth great variations and is used as an input to Otsu's global thresholding which removes most unwanted noise but also some faint characters. Therefore, the local binarization algorithm of Niblack is also used, but initialised using the stroke width information, extracted by skeletonization of Otsu's output, window size and contrast information. The two binarization outputs are combined at connected component level.

In [20], Su et al. demonstrated the use of local contrast image thresholding in estimating the text stroke width more accurately. In [6], Lu et al. performed background estimation using a modified version of 1D iterative polynomial smoothing to compensate for several degradation types. Text-stroke edges are then identified via Otsu's global thresholding on L1-norm horizontal and vertical edge detection. Document text pixels are extracted, since they are surrounded by text stoke edges and feature lower intensity levels.

Hedjam et al. [7] used grid-based modeling and impainting techniques to recover text pixels starting from an under-binarization result using Sauvola's technique. The proposed technique featured smooth and continuous strokes, due to its spatially adaptive estimation of the text pixels' statistical features. Moghaddam and Cheriet [8] presented an adaptive form of Otsu's thresholding for binarization. Based on a rough binarization result, they produce an estimated background and a stroke gray level map using a multi-scale framework. This estimated background is further refined using the AdOtsu method, which is an adaptive, parameterless form of Otsu's thresholding, which is generalized to a multiscale setup. Finally, skeleton-based post-processing is employed to remove possible artifacts and sub-strokes.

Valizadeh and Kabir [21] devised a novel feature space consisting of the structural contrast and the intensity value of each pixel. Structural contrast relates text stroke width, pixels' intensities and their relationships with their neighbors at stroke width distances. This results in a 2D image representation where text and background pixels are separable. Clustering is performed by partitioning the feature space into small regions. Then, using the result of another binarization algorithm with at least 50% successful labeling (Niblack), each region is classified either as background or text, according to the prevailing number of text or background pixels in the region. The reverse procedure produces the document binary image.

Howe [2] performed binarization by minimizing a global energy functional inspired by Markov Random Fields, where a) the image Laplacian edge map is employed to distinguish between ink and background in the energy data fidelity term and b) ink discontinuities are enforced in the binarization result by incorporating a Canny edge detector into the smoothness term. Howe also introduced a procedure to automate the optimal parameter selection for his algorithm.

Ramirez-Ortegon et al. [22] introduced the concept of transition pixel, i.e., calculating intensity differences over a small neighborhood, which can then be employed by common gray-level thresholding algorithms to produce a binarization result (transition method). This was further refined in [23], where an unsupervised thresholding was proposed for unimodal histograms, assuming Gaussian priors for the distribution of character and background neighborhoods. In [4], the method was enriched with a mechanism to remove binary artifacts after binarization. An auxiliary image is calculated via minimum-error-rate thresholding. The connected components of the auxiliary and the original binary image are compared in terms of an intersection ratio to remove possible binarization artifacts. In [24], Ramirez-Ortegon et al. explored possible effects of inaccurate estimations of the transition proportion on the estimated thresholds. In [25], Ramirez-Ortegon et al. proposed the use of skewed log-normal, instead of symmetrical Gaussian, priors [23] for the background and character clusters.

Lelore and Bouchara [3] introduced the FAIR binarization algorithm, where they ran the S-FAIR (simplified) algorithm for two threshold values: one giving a noiseless binarization output but with important edges missing and another containing all character edges but with some misclassification noise. The S-FAIR algorithm first performs text localization using the Canny algorithm. A Gaussian Mixture Model is then used to classify pixels around edges to belong either to the text or the background image or to a third class where pixels cannot be attributed with certainty to text or background. The FAIR algorithm merges the two outputs with a “max” rule. Finally, a post-filtering process classifies unknown pixels using a variety of rules. The most important feature is an iterative procedure where the text labeled regions grow into the unknown using morphological dilation and the previous EM algorithm is used to define the final class of these regions. Final unknown areas are connected morphologically and labeled according to neighboring pixels.

In this paper, the authors extend the previous work of Makridis and Papamarkos [16] toward a more automated three-stage document image binarization system. In the first stage, the background removal technique in [16] is enhanced by automating the window size selection for the median filter and improving the threshold selection between the document image and the background estimate. In the second stage, the proposed local neighborhood representation is redesigned to also include local contrast information to enhance the presence of character outlines. Binarization is then performed by separating two clusters of document characters and background artifacts that were not removed in the first stage of background removal. Clustering is performed using Mixtures of Gaussians (MoG). The Gaussian with lowest value mean corresponds to the character cluster. The local neighborhood representation share a similar concept with those introduced by Valizadeh and Kabir [21] and Ramirez-Ortegon et al. [22], however, the proposed multidimensional representation is different to the 1D representations discussed in [21,22]. Contrast information for binarization was also used by Su et al. [20], however, in this work contrast is incorporated into a local intensity representation forming a joint, rather than an isolated feature. Similarly, Gaussian modeling for binarization has been employed before by Hedjam et al. [7] and Ramirez-Ortegon et al. [23], but here it is applied on the novel LCM representation. Moreover, MoG-based clustering is a common clustering technique in pattern recognition, thus it is the application that is novel here. In the final post-processing stage, small-size 8-connected clusters are removed to eliminate possible binarization noise.

The paper is organized as follows: Section 2 sets the essential notation and outlines the system. Section 3 describes the background removal process in detail; Section 4 describes the binarization stage using GMM clustering; Section 5 explains the post-processing step; Section 6 presents the evaluation results of the proposed methodology and finally Section 7 concludes this paper.

2. System description

Let $\mathbf{I}(x, y)$ be the initial color document image of size $3 \times M \times N$, where x, y denote integer samples across the horizontal and vertical axes. The desired output of a document image binarization algorithm is a bi-level $M \times N$ image $I_{BN}(x, y)$ that attributes the value 255