A voting approach for image binarization of text-based documents

Giorgiana Violeta Vlăsceanu
University Politehnica of Bucharest
Romania
giorgiana.vlasceanu@cs.pub.ro

Costin-Anton Boiangiu
University Politehnica of Bucharest
Romania
costin.boiangiu@cs.pub.ro

ABSTRACT

The increasingly popular practice of digitising scanned documents often requires a preprocessing step of image enhancement, in order to increase the accuracy of the text recognition. In this paper various thresholding techniques for text-based images are combined through a voting mechanism. The scope of this paper is to offer the best result for binarization preprocessing.

KEYWORDS

image thresholding, voting technologies, Otsu, Niblack, Sauvola, Wolf, Nick, Bradley-Roth

1 INTRODUCTION

The focus of the current paper is proposing methods of qualitative evaluation for existing implementations of thresholding algorithms and to assess how we can improve choosing the most suitable output given a fixed set of algorithms.

2 PROPOSED METHOD

The proposed voting system takes into consideration multiple elections in order to come up with a certain set of proper thresholds on the given image. The **first election** of the voting algorithm is based on eliminating trivial candidates that don't represent a viable threshold. Since we started with the assumption that our input images are photos taken from books/documents with text-only information, we can impose that whatever image we will apply our algorithms on, the ratio between foreground and background pixels cannot be beyond above a certain ratio.

The **second election** of the voting algorithm takes a similar approach, but it applies it locally on windows of the approximate size of three text rows' height ensuring that, although global thresholding is successful, locally there have been detected some abnormalities. Although at this moment, we may use fixed thresholds like in the first step, this might most likely fail if we have for example page margins.

The **third election** is a tournament-based step designed in 2 steps: once on all binarizations and once on groups of binarizations; where intend to eliminate a ratio of the players.

The last step, which is not an election per se, takes the remaining images, builds a probability matrix, as in the previous step, consider it as a grayscale image where the probability in the [0%-100%] range stretched to [0-255] gray-shades range, and constructs the final binarized image by thresholding in the middle (127).

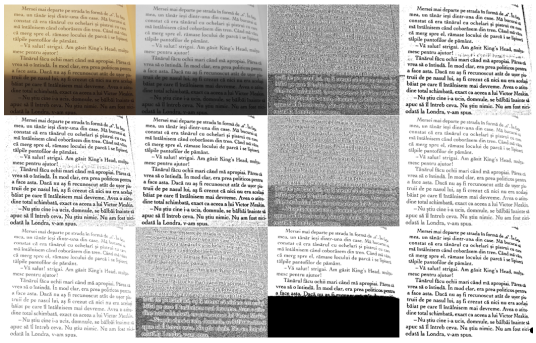

Figure 1: From left to right, top to bottom: original image document, grayscale conversion using CIE-Y luminance component, local average, adaptive based on offset average, adaptive based on offset Gaussian, Bradley-Roth, Niblack, Nick, Sauvola, Wolf, Otsu, Proposed voting approach.

3 IMPLEMENTATION AND RESULTS

To evaluate the presented approach, a thorough comparison was performed against some of the most popular thresholding techniques: Otsu[2] for the globally-based approach and Niblack, Sauvola [1], Wolf, Nick for the local-based approaches. The tests revealed that the presented voting approach always ranks amongst the best methods if the input image document is text-based. For documents that contain unusual writing, illustrations, decorators, or diagrams, the presented research may not offer optimal results since the voting selection is based on rejection tests using the average font-filling statistics.

4 CONCLUSIONS

This paper presented a series of techniques for image thresholding to separate the foreground from the background. The proposed approach presents a binarization system that runs several methods to generate viable candidates for the problem, then performs a series of validations tests and voting-based approaches in a tournament-like selections to generate the most suitable candidate.

Future work may be oriented toward better rejection statistics at both individual level and tournament level and a more educated shuffle operation.

REFERENCES

[1] M Pietikainen J. Sauvola. 2000. Adaptive document image binarization, Pattern Recognition.
[2] Andrey Samorodo O. A. Samorodova. 2016. Fast implementation of the Niblack binarization algorithm for microscope image segmentation. In *Pattern Recognition and Image Analysis.*