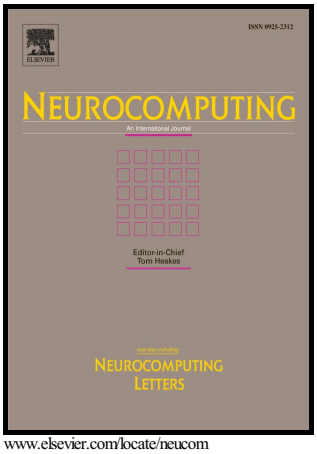
Broken and Degraded Document Images Binarization

Yiping Chen, Liansheng Wang



www.elsevier.com/locate/neucom

PII: S0925-2312(16)31602-2
DOI: http://dx.doi.org/10.1016/j.neucom.2016.12.058
Reference: NEUCOM17889

To appear in: *Neurocomputing*

Received date: 17 February 2015
Revised date: 2 September 2016
Accepted date: 26 December 2016

Cite this article as: Yiping Chen and Liansheng Wang, Broken and Degraded Document Images Binarization, *Neurocomputing*, http://dx.doi.org/10.1016/j.neucom.2016.12.058

Broken and Degraded Document Images Binarization

Yiping Chen[a,b], Liansheng Wang*[a]

[a]*Department of Computer Science, School of Information Science and Engineering, Xiamen University, Xiamen, China.*
[b]*Institute of Space Electronic Information Technology, School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, P. R. China.*

Abstract

Document image binarization refers to the conversion of a document image into a binary image. For broken and severely degraded document images, binarization is a very challenging process. Unlike the traditional methods that separate the foreground from the background, this paper presents a new framework for the binarization of broken and degraded document images and restoring the quality of the document images. In our approach, the non-local means method is extended and used to remove noises from the input document image in the step of pre-process. Then the proposed method binarizes the document image which takes advantage of the quick adaptive thresholding proposed by Pierre D. Wellner. To get more pleasing binarization results, the binarized document image is post-processed finally. There are three measures in the post-process step: de-speckle, preserve stroke connectivity and improve quality of text regions. Experimental results show significant improvement in the binarization of the broken and degraded document images collected from various sources including degraded and broken books, magazines and document files.

Keywords: binarization, thresholding, document image

1. Introduction

Historical and ancient document collections available in libraries throughout the world are of great cultural and scientific importance. The transformation of such documents into digital form is essential for maintaining the quality of the originals while provide scholars with full access to that information [1]. However, the document images may be broken and degraded