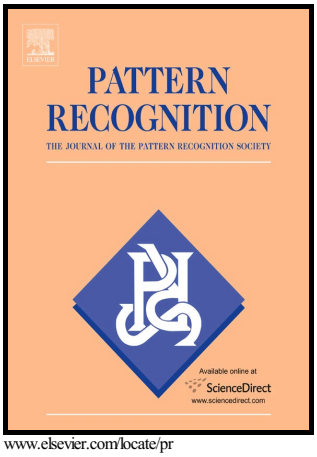
An Approach for Detecting and Cleaning of Struck-out Handwritten Text

Bidyut B. Chaudhuri, Chandranath Adak



www.elsevier.com/locate/pr

PII: S0031-3203(16)30190-X
DOI: http://dx.doi.org/10.1016/j.patcog.2016.07.032
Reference: PR5818

To appear in: *Pattern Recognition*

Received date: 18 January 2016
Revised date: 22 July 2016
Accepted date: 22 July 2016

Cite this article as: Bidyut B. Chaudhuri and Chandranath Adak, An Approach for Detecting and Cleaning of Struck-out Handwritten Text, *Pattern Recognition* http://dx.doi.org/10.1016/j.patcog.2016.07.032

An Approach for Detecting and Cleaning of Struck-out Handwritten Text

Bidyut B. Chaudhuri[a,1], Chandranath Adak[a,b,1]

[a] *CVPR Unit, Indian Statistical Institute, India-700108*
[b] *School of ICT, Griffith University, Australia-4222*

Abstract

This paper deals with the identification and processing of struck-out texts in unconstrained offline handwritten document images. If run on the OCR engine, such texts will produce nonsense character-string outputs. Here we present a combined (a) pattern classification and (b) graph-based method for identifying such texts. In case of (a), a feature-based two-class (normal vs. struck-out text) SVM classifier is used to detect moderate-sized struck-out components. In case of (b), skeleton of the text component is considered as a graph and the strike-out stroke is identified using a constrained shortest path algorithm. To identify zigzag or wavy struck-outs, all paths are found and some properties of zigzag and wavy line are utilized. Some other types of strike-out stroke are also detected by modifying the above method. The large sized multi-word and multi-line struck-outs are segmented into smaller components and treated as above. The detected struck-out texts can then be blocked from entering the OCR engine. In another kind of application involving historical documents, page images along with their annotated ground-truth are to be generated. In this case the strike-out strokes can be deleted from the words and then fed to the OCR engine. For this purpose an inpainting-based cleaning approach is employed. We worked on 500 pages of documents and obtained an overall F-Measure of 91.56% (91.06%) in English (Bengali) script for struck-out text detection. Also, for strike-out stroke identification and deletion, the F-Measures obtained were 89.65% (89.31%) and 91.16% (89.29%), respectively.

Keywords: Crossed-out text, Document cleaning, Handwritten OCR, Inpainting, Strike-out stroke, Struck-out text processing.

Email addresses: bbcisical@gmail.com (Bidyut B. Chaudhuri), adak32@gmail.com (Chandranath Adak)

[1] Both the authors contributed equally to this work.