International Journal on Document Analysis and Recognition (IJDAR) (2023) 26:335–346
https://doi.org/10.1007/s10032-023-00438-7

SPECIAL ISSUE PAPER

Line extraction in handwritten documents via instance segmentation

Adeela Islam[1] · Tayaba Anjum[1] · Nazar Khan[1]

Received: 14 December 2022 / Revised: 16 March 2023 / Accepted: 18 April 2023 / Published online: 21 May 2023

Abstract
Extraction of text lines from handwritten document images is important for downstream text recognition tasks. It is challenging since handwritten documents do not follow strict rules. Significant variations in line, word, and character spacing and line skews are acceptable as long as the text remains legible. Traditional rule-based methods that work well for printed documents do not carry over to the handwritten domain. In this work, lines are treated as objects to leverage the power of deep learning-based object detection and segmentation frameworks. A key benefit of learnable models is that lines can be implicitly defined through annotations of training images which allows unwanted textual content to be ignored when required. A deep instance segmentation model trained in end-to-end fashion without any dataset-specific pre- or post-processing achieves 0.858 pixel IU and 0.899 line IU scores averaged over 9 different datasets comprising a wide variety of handwritten scripts, layouts, page backgrounds, line orientations, and interline spacings. It achieves state-of-the-art results on DIVA-HisDB, VML-AHTE, and READ-BAD datasets and almost state-of-the-art results on Digital Peter, ICDAR2015-HTR, ICDAR2017, and Bozen datasets. We also introduce a new, annotated dataset for Urdu script. Our model trained only on Urdu generalizes to multiple other scripts, indicating that it learns a script-invariant representation of text lines. All code, pre-trained models, and the new Urdu dataset can be accessed at https://github.com/AdeelaIslam/HLExt-via-IS.

Keywords Text line extraction · Urdu · Multiscript · Handwritten · Instance segmentation · Deep learning

1 Introduction

Writing has been an effective way of human communication for ages. Technological advancements are progressing to make text recognition as easy as possible. Text line extraction serves as the starting point of multiple document digitization tasks such as text recognition [2], spotting [3], manuscripts alignment [4], and writer recognition [5]. This task becomes even more challenging in images of handwritten documents with skewed as well as curved lines in varying writing styles, pen types, inks, and page backgrounds.

Obtaining a single definition of what constitutes a text line is not trivial. Figure 1 demonstrates that definitions used for Latin-based scripts and/or printed text do not carry over to non-Latin scripts such as Urdu and Arabic in handwritten form. However, the situation is similar to object detection and segmentation problems where instead of defining an object or segment, deep learning-based models learn from multiple examples of manually annotated and segmented objects. Therefore, we treat lines as objects in order to leverage the power of existing deep learning-based object detection and segmentation frameworks. Existing attempts at line segmentation predominantly use fully convolutional networks (FCNs) [6]. The Mask R-CNN framework [7] improves semantic segmentation by decoupling classification and segmentation problems.

For low-resource scripts such as Urdu and Persian, there is a lack of methods as well as datasets. The script that is closest to them is Arabic, for which there exists a handwritten line segmentation method [8] which uses a Mask R-CNN to classify and segment individual image patches into line versus non-line pixels. We show that decomposition into patches is counterproductive and restrictive. Instead, instance segmentation frameworks such as a Mask R-CNN are fully capable of learning to segment complete line objects in an end-to-end fashion.

✉ Nazar Khan
nazarkhan@pucit.edu.pk

Adeela Islam
adeela.islam@pucit.edu.pk

Tayaba Anjum
tayaba.anjum@pucit.edu.pk

[1] Department of Computer Science, University of the Punjab, Lahore 54000, Punjab, Pakistan