Sistem Optical Character Recognition Untuk Huruf Arab Pegon = Optical Character Recognition System for Printed Pegon Manuscripts

Muhammad Hanif Fahreza, author

Abstrak

Aksara Pegon adalah sistem penulisan berbasis Arab yang digunakan untuk menulis bahasa Jawa, Sunda, Madura, dan Indonesia. Karena berbagai alasan, aksara ini telah diturunkan ke ranah kolektor naskah sejarah dan pesantren, sehingga perlu dilestarikan. Salah satu metode pelestarian ini adalah melalui digitalisasi; lebih tepatnya dengan mentranskripsikan isi dari naskah-naskah yang ada ke dalam bentuk teks machine encoded, dimana proses tersebut jika dilakukan secara otomatis disebut juga sebagai OCR, atau Pengenalan Karakter Optik. Sampai saat ini belum ada literatur yang dipublikasikan mengenai sistem OCR untuk aksara ini. Oleh karena itu, penelitian ini bertujuan untuk menjembatani kesenjangan tersebut dengan menyediakan OCR untuk subset tertentu dari naskah Pegon, yaitu naskah Pegon yang dicetak. Penelitian ini memperkenalkan dataset yang disintesis dan yang dianotasi untuk pengenalan teks Pegon cetak. Dataset-dataset ini kemudian digunakan untuk mengevaluasi sistem OCR Arab konvensional yang sudah ada pada domain Pegon, baik versi asli maupun yang dimodifikasi, serta sistem berbasis teknik deep learning yang lebih baru dalam literatur. Hasilnya menunjukkan bahwa teknik deep learning mengungguli teknik konvensional, di mana teknik konvensional gagal mendeteksi teks Pegon sama sekali, sementara sistem yang diusulkan, khususnya menggunakan YOLOv5 untuk segmentasi baris dan arsitektur CTC-CRNN untuk pengenalan teks baris, mencapai nilai F1 sebesar 0,94 untuk segmentasi dan CER 0,03 untuk pengenalan teks.
.....The Pegon script is an Arabic-based writing system intended for writing the Javanese,
Sundanese, and Indonesian languages. Due to various reasons, this script has been
relegated to the domain of historical manuscript collectors and private Islamic boarding
schools or pesantren, presenting a need for preservation. One of these methods of
preservation is through digitization; more specifically, by transcribing the content of
these existing manuscripts into machine-encoded text, the automated process of which is
referred to as OCR. There has been heretofore no published literature on OCR systems
for this specific script. Hence, this research aims to bridge that gap by providing a
foray into the OCR of a specific subset of Pegon manuscripts, namely of printed Pegon
manuscripts. This research evaluates existing and modified versions of conventional
Arabic OCR systems on the domain of Pegon, as well as the more recent deep learning
techniques in the literature, along with introducing new datasets for use in developing
with said deep learning techniques. The results show the outperformance of these deep
learning techniques over the conventional techniques and with which components of a
Pegon OCR system is proposed.