IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2019, accepted September 24, 2019, date of publication October 7, 2019, date of current version October 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945825

Detecting Mathematical Expressions in Scientific Document Images Using a U-Net Trained on a Diverse Dataset

WATARU OHYAMA[1], (Member, IEEE), MASAKAZU SUZUKI[2], AND SEIICHI UCHIDA[3], (Member, IEEE)

[1]Graduate School of Engineering, Saitama Institute of Technology, Fukaya-shi 3690293, Japan
[2]Faculty of Mathematics, Kyushu University, Fukuoka 8190395, Japan
[3]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 8190395, Japan

Corresponding author: Wataru Ohyama (ohyama@sit.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant JP17H06100.

ABSTRACT A detection method for mathematical expressions in scientific document images is proposed. Inspired by the promising performance of U-Net, a convolutional network architecture originally proposed for the semantic segmentation of biomedical images, the proposed method uses image conversion by a U-Net framework. The proposed method does not use any information from mathematical and linguistic grammar so that it can be a supplemental bypass in the conventional mathematical optical character recognition (OCR) process pipeline. The evaluation experiments confirmed that (1) the performance of mathematical symbol and expression detection by the proposed method is superior to that of InftyReader, which is state-of-the-art software for mathematical OCR; (2) the coverage of the training dataset to the variation of document style is important; and (3) retraining with small additional training samples will be effective to improve the performance. An additional contribution is the release of a dataset for benchmarking the OCR for scientific documents.

INDEX TERMS Character recognition, neural networks, object detection.

I. INTRODUCTION

The performance and effectiveness of document retrieval systems heavily depend on both the amount and quality of registered document content. Although born-digital documents have become more common recently, a large number of printed documents remain. To input such printed documents into retrieval systems, optical character recognition (OCR) techniques have been used for digitizing documents for a long time. Continuous research and development over the last five decades have achieved OCR techniques that are sufficiently mature for such a purpose.

Although OCR techniques demonstrate good performance for digitizing ordinary text in documents, there is still scope for improvement in the recognition accuracy of mathematical expressions (MEs). MEs are essential information containers, particularly for scientific articles and textbooks. The accurate recognition of MEs is strongly expected because it has a wide variety of applications, for instance, correct retrieval, automatic proofing of MEs, and learning support for blind or handicapped people. ME recognition has been considered and developed as an independent module outside of ordinary OCR because of the distinctive properties of MEs, where spatial structures and spatial relationships between symbols contain mathematical information.

Zanibbi and Blostein [1] stated that there are four key problems in ME recognition: ME detection, symbol extraction and recognition, layout analysis and mathematical content interpretation. These four problems are closely related to each other. In particular, ME detection has a large influence on other tasks. There are two types of MEs: displayed and in-line, as shown in Figure 1. The detection processes for each displayed (offset from text lines) and in-line (embedded in text lines) ME are usually implemented separately. When an in-line ME is not detected, the expression is passed to the recognition module for ordinary characters even though it should be passed the recognition module for mathematical symbols. This scenario commonly occurs, and the undetected ME may cause recognition errors that cannot be easily corrected in subsequent postprocessing modules.

Because of the high performance of deep neural networks (DNNs), particularly convolutional neural

The associate editor coordinating the review of this manuscript and approving it for publication was Habib Ullah.