Originally published in: 15th IAPR International Workshop on Document Analysis Systems (DAS 2022).
– La Rochelle: La Rochelle University, 2022. Pp. 32-35.

Robust Extraction of Marked-Up Text Sections from Scientific Document Printouts

Mark-Christoph Müller[0000−0001−5639−7682]

Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
mark-christoph.mueller@h-its.org

Abstract. We present a simple tool for extracting text and markup information from printouts of (not only) scientific documents. While the heavy-lifting OCR is done by off-the-shelf TESSERACT, our focus is on detection, extraction, and basic *categorization* of color-highlighted text sections, as well as on providing a framework for downstream processing of extraction results. The tool can be useful for document analysis tasks that must, or benefit from being able to, use printed paper.

Keywords: Document Images · Information Extraction · OCR · Multimodality · Natural Language Processing

1 Introduction

Despite the shift towards PDF and XML, **printed paper** is still crucial for scientific document use.[1] It is the medium of choice for *active reading*, supporting straightforward markup with highlighter pens, which is commonly done during manual excerption from scientific literature. However, as soon as the highlighted text is supposed to undergo further *computational* processing, paper ceases to be practical. Biomedical database curation [1] is a case in point: Here, human domain experts often use paper printouts to mark up relevant sections in scientific documents, but for the subsequent database insertion (often done by other people), the data has to be re-keyed manually, which is both inefficient and error-prone.

We present a simple OCR-based document analysis tool which combines the advantages of working with paper hard-copies and the efficiency of automatic text recognition and extraction. In essence, the tool mainly integrates an OCR component (off-the-shelf TESSERACT, see below), a simple image processing module, and an XML-based multi-level annotation processing framework from natural language processing (NLP). Thus, our focus is on providing robust **core extraction functionality** based on proven state-of-the-art components, rather than on optimizing individual modules. Also, by using an NLP data representation

[1] This work was done as part of the project DeepCurate, which is funded by the German Federal Ministry of Education and Research (BMBF) (No. 031L0204) and the Klaus Tschira Foundation, Heidelberg, Germany.

framework including an API, we establish straightforward technical connectivity between extraction results and downstream processing (see Section 4). Code and data are available at https://github.com/nlpAThits/docimg2mmax.

2 System Overview

The tool significantly extends and improves our previous work in [4], where earlier versions of some of the current functionality were used. Basically, the tool reads a scanned document, consisting of one image per page, recognizes and extracts the text content, then (optionally) analyses the image for color-highlighted sections, and creates special word-level annotations for highlighted content.

We use the MMAX2[2] [5] multi-level annotation processing framework for representation and further processing of OCR and extraction results. MMAX2 supports visualization and manual annotation of the extracted data (see below), and also provides a Python API [3]. In a nut shell, data in MMAX2 is stored in the form of so-called MARKABLES, which aggregate arbitrary attribute-value pairs and associate these with underlying, immutable text data (in this case with the OCR result).

OCR is performed with TESSERACT (tested with version 4.1.1), which is only loosely integrated and called via Python sub-processes. TESSERACT can output its results in hOCR format[3], which includes highly detailed recognition information. The generation of hOCR output is always activated, while other parameters (`--oem`, `--psm`, `--dpi`, and `--tessdata-dir`) are directly passed through. This way, a high degree of transparency and flexibility is maintained. After recognition, the hOCR file is analysed, and the recognized text as well as bounding box and confidence information for line, word, and character elements is stored in MARKABLES on different annotation levels. Optionally, the tool can also create an HTML file with an SVG-based overlay of the original image, which visualizes the extracted marked-up text. Markup detection and extraction works by analyzing the page image, identifying colored areas, and mapping these to previously extracted words, based on the latters' bounding boxes. Highlighting can appear either *horizontally* on the desired text, or, for larger sections that span several lines, *vertically*, e.g. on page margins (see Figure 1). The detection of colored image areas takes advantage of the fact that, in an RGB image, *non-colored* pixels have highly similar values in their three channels, while whenever at least one channel value differs above a certain threshold (we use an absolute value of 10) from the others, the pixel actually has a discernible color.

3 Examples

We demonstrate the tool on a black-and-white printout of an open-access scientific paper [2] which has been marked up using different colors and then scanned

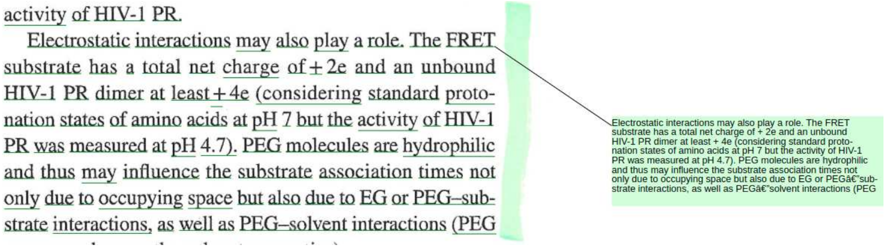
[2] https://github.com/nlpAThits/MMAX2

[3] http://kba.github.io/hocr-spec/1.2/

in 300 dpi. Figure 2 shows two example extraction results. In each example, the left image is a part of the scanned page image, and the right image shows the rendering of the extracted full text in MMAX2. Highlighted words are rendered with a yellow background. Note the OCR accuracy (courtesy of TESSERACT), which at least for standard text is almost perfect. Boxes in the left images are drawn automatically around highlighted words. For each highlighted word, two properties are determined, viz. the *percentage of the word area* that is actually highlighted, and the *dominant highlighting color*. A threshold on the first property is used (here: 10%) to discard words that are only marginally touched by coloring. The second property is intended to capture a kind of highlighting *category* by allowing to cluster words that were highlighted *in the same color*. It is implemented by just selecting, from the colored part of each word's bounding box, the most frequent RGB triple. Table 1 shows the respective properties for one word each from the four colored regions in Figure 2. Visualization of the dominant colors is for illustration only; actual clustering / categorization will have to be done by analysing the ratio of the three color channel values.

Word	*"reaction"*	*"peptide"*	*"substrate"*	*"crowding"*
% HL	36%	69%	89%	85%
Dominant color	245:255:244	253:224:246	241:255:255	203:254:213

Table 1. Highlighted words with automatically extracted dominant color.



activity of HIV-1 PR.

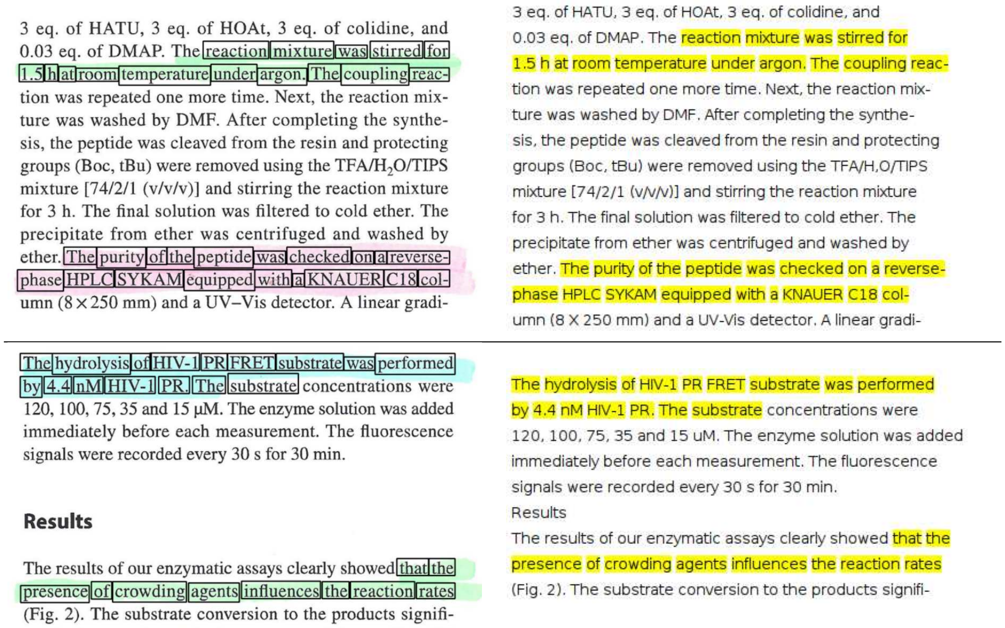
Electrostatic interactions may also play a role. The FRET substrate has a total net charge of + 2e and an unbound HIV-1 PR dimer at least + 4e (considering standard protonation states of amino acids at pH 7 but the activity of HIV-1 PR was measured at pH 4.7). PEG molecules are hydrophilic and thus may influence the substrate association times not only due to occupying space but also due to EG or PEG–substrate interactions, as well as PEG–solvent interactions (PEG

Electrostatic interactions may also play a role. The FRET
substrate has a total net charge of + 2e and an unbound
HIV-1 PR dimer at least + 4e (considering standard proto-
nation states of amino acids at pH 7 but the activity of HIV-1
PR was measured at pH 4.7). PEG molecules are hydrophilic
and thus may influence the substrate association times not
only due to occupying space but also due to EG or PEGâ€"sub-
strate interactions, as well as PEGâ€"solvent interactions (PEG

Fig. 1. Detail of HTML file with extracted *vertical* markup.

4 Summary & Outlook

The presented text extraction and markup detection tool is deliberately designed to be simple and reduced to core functionality. Nevertheless, our rather superficial evaluation showed that both out-of-the-box OCR and markup extraction quality is very good, provided that 1) the image quality is good (clean black-and-white printout) and 2) appropriate highlighter colors are used. In an actual application scenario, these factors can easily be controlled for. Next, we are going to evaluate the applicability of the tool in a literature-based biomedical database curation scenario. Database curation from documents should be able to benefit strongly from powerful and flexible text search, including e.g. handling of synonyms. Once a document has been processed with our tool, these functionalities



3 eq. of HATU, 3 eq. of HOAt, 3 eq. of colidine, and 0.03 eq. of DMAP. The reaction mixture was stirred for 1.5 h at room temperature under argon. The coupling reaction was repeated one more time. Next, the reaction mixture was washed by DMF. After completing the synthesis, the peptide was cleaved from the resin and protecting groups (Boc, tBu) were removed using the TFA/H_2O/TIPS mixture [74/2/1 (v/v/v)] and stirring the reaction mixture for 3 h. The final solution was filtered to cold ether. The precipitate from ether was centrifuged and washed by ether. The purity of the peptide was checked on a reverse-phase HPLC SYKAM equipped with a KNAUER C18 column (8 × 250 mm) and a UV–Vis detector. A linear gradi-

The hydrolysis of HIV-1 PR FRET substrate was performed by 4.4 nM HIV-1 PR. The substrate concentrations were 120, 100, 75, 35 and 15 μM. The enzyme solution was added immediately before each measurement. The fluorescence signals were recorded every 30 s for 30 min.

Results

The results of our enzymatic assays clearly showed that the presence of crowding agents influences the reaction rates (Fig. 2). The substrate conversion to the products signifi-

3 eq. of HATU, 3 eq. of HOAt, 3 eq. of colidine, and 0.03 eq. of DMAP. The reaction mixture was stirred for 1.5 h at room temperature under argon. The coupling reaction was repeated one more time. Next, the reaction mixture was washed by DMF. After completing the synthesis, the peptide was cleaved from the resin and protecting groups (Boc, tBu) were removed using the TFA/H,O/TIPS mixture [74/2/1 (v/v/v)] and stirring the reaction mixture for 3 h. The final solution was filtered to cold ether. The precipitate from ether was centrifuged and washed by ether. The purity of the peptide was checked on a reverse-phase HPLC SYKAM equipped with a KNAUER C18 column (8 X 250 mm) and a UV-Vis detector. A linear gradi-

The hydrolysis of HIV-1 PR FRET substrate was performed by 4.4 nM HIV-1 PR. The substrate concentrations were 120, 100, 75, 35 and 15 uM. The enzyme solution was added immediately before each measurement. The fluorescence signals were recorded every 30 s for 30 min.
Results
The results of our enzymatic assays clearly showed that the presence of crowding agents influences the reaction rates (Fig. 2). The substrate conversion to the products signifi-

Fig. 2. Images with color highlighting (left) and extracted text (right).

are available with little extra effort on the basis of the MMAX2 format and the Python API.

References

1. International Society for Biocuration: Biocuration: Distilling data into knowledge. PLOS Biology **16**(4), 1–8 (2018). https://doi.org/10.1371/journal.pbio.2002846
2. Maximova, K., Wojtczak, J., Trylska, J.: Enzymatic activity of human immunodeficiency virus type 1 protease in crowded solutions. European Biophysics Journal **48**(7), 685–689 (2019). https://doi.org/10.1007/s00249-019-01392-1
3. Müller, M.C.: pyMMAX2: Deep access to MMAX2 projects from python. In: Proceedings of the 14th Linguistic Annotation Workshop. pp. 167–173. Association for Computational Linguistics, Barcelona, Spain (Dec 2020), https://aclanthology.org/2020.law-1.16
4. Müller, M.C., Ghosh, S., Wittig, U., Rey, M.: Word-level alignment of paper documents with their electronic full-text counterparts. In: Demner-Fushman, D., Cohen, K.B., Ananiadou, S., Tsujii, J. (eds.) Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021. pp. 168–179. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.bionlp-1.19
5. Müller, M.C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (eds.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006)