A Hybrid Information Extraction Approach using Transfer Learning on Richly-Structured Documents

Arnab Ghosh Chowdhury[1], Nils Schut[2] and Martin Atzmueller[1]

[1]Osnabrück University, Semantic Information Systems (SIS) Group, Osnabrück, Germany
[2]Polymer Science Park, Zwolle, The Netherlands

Abstract

Richly-structured documents such as PDFs provide a rich source of information, where - however - its extraction is often challenging due to the complex structures. Computer vision, optical character recognition (OCR) and deep learning offer significant opportunities in the field of information extraction from PDF articles. However, it is extremely challenging to create a unified framework to extract information from different types of PDF documents due to their diverse visual appearance. In this paper, we propose a hybrid information extraction approach for documents with complex structures. In particular, it features a pipeline which uses OCR for plain textual information extraction and transfer learning for table detection from documents with such rich and complex structure. Our application context is given by technical (product) datasheets, in particular plastic product technical data sheets for service provisioning. We discuss first experimental results and outline several challenges in this context.

Keywords

Information Extraction, Transfer Learning, Optical Character Recognition (OCR), Table Detection

1. Introduction

In the age of digitalization, information is often provided in digital form, however, often without semantic structuring or annotation. In this context, information is often provided in the form of richly-structured documents, for which the structure itself imposes constraints on the semantic interpretation of its contents, such as provided by complex tables in technical documents. However, these are often only reflected in the layout of the documents without being accessible to methods for automatic interpretation or (semantic) information extraction being applied on such documents.

In this paper, we propose a hybrid information extraction approach using transfer learning on richly-structured documents. This approach features a pipeline which uses OCR for plain textual information extraction and transfer learning for table detection, aiming to exploit the rich but complex structure of richly-structured documents for extracting the relevant information. With this approach, we can tackle the issues discussed above, in order to extract rich information from complex (technical) documents, where we combine both methods for extracting complementary information elements which are then integrated subsequently.

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany
arnab.ghosh.chowdhury@uos.de (A. Ghosh Chowdhury); N.Schut@polymersciencepark.nl (N. Schut); martin.atzmueller@uni-osnabrueck.de (M. Atzmueller)

Our application context is given by the project Di-Plast[1] which aims to improve processes for a more stable recycled plastics material (rPM) supply and quality using artificial intelligence methods and data science approaches. The utilization of rPM is at the moment below its potential due to primarily low uptake of rPM because of lack of information about quality and quantity of the material. To alleviate such challenges, the software-based tools developed in Di-Plast are divided in four separate disciplines. One of the disciplines is *Matching Product Requirements* which mainly addresses how to improve the uptake of available rPM quantity by introducing the *Matrix Tool* [1, 2] – a knowledge-based system for this purpose. For this, information about plastics materials needs to be provided, where one option is to extract this from plastic product technical data sheets automatically.

Plastic product technical data sheets offer high quality material information commonly in PDF format. In general, data extraction is quite complex due to diverse layout and visual appearance of PDF documents. Different plastic product manufacturers follow different types of document templates to provide the relevant information. An information extraction pipeline is essential to integrate such material information into a comprehensive database that can then be leveraged by the stakeholders in the plastic recycling industry. This paper focuses on a hybrid information extraction pipeline approach on richly-structured PDF articles that utilizes the transfer learning technique for table detection in the research area of document layout analysis. Our contributions are summarized as follows:

1. We propose a hybrid approach for information extraction, utilizing OCR for plain textual information extraction and a transfer learning based table detection technique to support effective tabular data extraction.
2. We present and discuss our experimentation, evaluating our proposed approach using transfer learning on our applied dataset, and review current challenges with respect to our proposed approach and its context.

The rest of the paper is organized as follows: Section 2 discusses related work in the area of document layout analysis and transfer learning approaches. Section 3 presents our proposed hybrid information extraction approach combining plain textual information extraction with transfer learning based table detection. Next, Section 4 presents the results of our experimentation: We provide first results of our proposed approach, and also discuss some of the current limitations regarding the use of transfer learning for table detection on our applied dataset. Finally, Section 5 concludes the paper with a summary and outline directions for future work.

2. Related Work

Document layout analysis is an important approach in document image analysis and recognition to identify and perceive the logical and physical structure of PDF documents [3]. Rich textual information is inferred from such documents along with its semantic contents by such analysis approaches. On the other hand, OCR[2] is a best practice to extract relevant information from document images. Below, we discuss related work in the general scope of document layout analysis, information extraction and transfer learning on document layout analysis.

[1] https://www.nweurope.eu/projects/project-search/di-plast-digital-circular-economy-for-the-plastics-industry/
[2] https://cordis.europa.eu/project/id/IST-1999-20021

2.1. Computer Vision based Document Layout Analysis

We consider computer vision based document layout analysis as a preliminary step towards retrieving tabular information from PDF articles such as product properties, values and testing methods from the tables of product technical data sheets in our example domain of the plastic recycling industry. Furthermore, a data driven approach is applied to use OCR for extracting other textual information such as *product description*, *processing techniques*, *storage information*. The fundamental objective of computer vision based object detection technique is to locate and to classify existing objects in an image and labeling the objects with rectangular bounding boxes to exhibit the confidences of its existence.

The generic object detection techniques can be categorized into two types. The first type follows a pipeline where region proposals are generated at first and then each proposal is classified into different object categories. The second type considers object detection as a regression or a classification problem by incorporating a unified framework to identify categories and locations directly. The region proposal based techniques comprise R-CNN (Region Based Convolutional Neural Networks), Fast R-CNN, Faster R-CNN, Mask R-CNN, FPN (Feature Pyramid Network) and other models. On the other hand, the regression or classification based techniques include MultiBox, AttentionNet, YOLO (You Only Look Once), YOLOv2, SSD (Single Shot MultiBox Detector) and other models. The correlations between this two pipelines are associated by the anchors introduced in Faster R-CNN [4].

In general, object detection techniques play a significant role in document layout analysis. A method is adapted to visually segment important regions of scientific articles by Faster R-CNN model for document layout detection [5]. A research on PubLayNet dataset containing over 1 million PDF articles is carried out considering Faster R-CNN and Mask R-CNN models to identify different document layout objects such as Title, Text, List, Figure and Table on biomedical articles [6]. An image based table detection and structure recognition research based on Faster R-CNN model with different settings is conducted on TableBank dataset [7]. To empower the research on document layout analysis, an empirical research on DocBank dataset is performed on four baseline models such as BERT, RoBERTa, LayoutLM and Faster R-CNN with fine-grained token level annotations for document layout analysis [8]. Another computer vision based research is carried out for automatic document layout analysis and content extraction to obtain rich information from historical Japanese documents with complex layouts by considering Faster R-CNN, Mask R-CNN and RetinaNet models [9].

Several other extensive researches are conducted by state-of-the-art computer vision based algorithms on historical American digitized newspapers [10] and scanned pages from contemporary magazines and technical articles [11]. To infer the complete hierarchical structure of digitized documents, a system named Docparser is developed to parse the complete document structure which includes text elements, nested figures, tables, and table cell structures [12]. Furthermore a data-driven system is proposed mostly to detect and extract figures and tables in PDF documents [13]. On the other hand, a multimodal, fully convolutional network is presented to extract semantic structures from document images by considering an Encoder-Decoder architecture [14]. Another unified framework named VSR is proposed for multimodal layout analysis which integrates vision, semantics and relations, but suffers to generalize as it needs the positions and contents of texts in the document [15].

2.2. Transfer Learning based Document Layout Analysis

Data annotation on document images is labor intensive and time consuming task in our dataset. A research is carried out for document layout detection and OCR where a pre-trained model on a dataset can be fine tuned on other dataset [16]. When the annotated training dataset is relatively small, Few-shot object detection is an alternative approach in document layout analysis to use a pre-trained model on a large source dataset and to fine tune the model on a relatively very small target dataset [17].

3. Method and Experiment

Technical data sheets are generally represented as PDF articles and include product names, document titles and subtitles, different other information about the products, property information in tabular format and the logo of the manufacturers, etc. For extracting textual information from such PDF articles with complex structures, several open-source tools are available e. g., PyPDF2, PDFMiner. But it is quite difficult to parse PDF articles by such tools maintaining the order of word sequences properly as well as the proper document page orientation and also to obtain tabular data with the appropriate semantic table structure. We follow a hybrid method combining state-of-the-art computer vision based deep neural networks to analyze the layout of document images for table location detection by transfer learning and OCR based plain textual information extraction to extract other textual information from such technical data sheets.

To extract data in our proposed hybrid method, we use *Pytesseract*[3], a wrapper for Google's *Tesseract-OCR* engine which can recognize more than 100 languages. In general, it is a cost-intensive effort to properly annotate a large number of different segments on our document images following the PubLayNet research work [6] into various categories; this considers, for example, the logo of the manufacturer, title, text and tables in our dataset, etc. Therefore, we consider a hybrid method to reduce the manual annotation effort, and to perform manual data annotation only for table detection thus performing transfer learning, while only requiring a relatively small dataset in this way.

With the rapid development of deep learning in computer vision data-driven image-based approaches for document layout analysis [8] are widely adopted. TableBank exhibits an image-based table detection and recognition research work with weak supervision from Microsoft Word and Latex documents and builds multiple strong baselines using state-of-the-art deep neural network models. We train our model using transfer learning for table detection from a base model which is pre-trained on the TableBank dataset [6]. Furthermore, we use *Detectron2*[4], a PyTorch-based Facebook AI Research's library that provides the state-of-the-art detection and segmentation algorithms. We consider technical data sheets on low density polyethylene (LDPE) resins from a multinational chemical manufacturer as our Di-Plast dataset for this research work. In the following, we provide some anonymized/synthesized examples of the respective technical data sheets, where we recreate the important structures for exemplification, providing illustrative example documents in this way. Image-1 of Figure 1, for example, refers to such a sample page of an exemplary plastic product technical data sheet.

[3] https://github.com/madmaze/pytesseract/

[4] https://github.com/facebookresearch/detectron2

3.1. Plain Textual Information Extraction

Initially we split all pages of PDF articles and convert them to images using the *pdf2image*[5] tool. During pre-processing in a plain textual information extraction process, we first apply several image processing operations such as converting a color image into a gray scale image, removing the logo of manufacturers and binarization on the images using an adaptive threshold via *OpenCV*[6]. We use OCR techniques to extract textual data from the images via *Pytesseract* and store them in a text file after applying some post-processing. Sometimes the logo of some manufacturers contain text which is nonessential during textual data extraction using OCR. In order to remove such logos from the images, we use template matching to identify the location of such a template image (e. g., the manufacturer logo) in the larger images. When textual data is extracted using OCR techniques, we use a post-processing method to remove some nonessential data or stop words for subsequent analysis such as *page number*, *recipient tracking number*, *request number*. Later, we extract textual information between two subtitles of the technical data sheets such as *Product Description* and *Regulatory Status* from the relevant text file by regular expression operations. The subtitle *Product Description* of a technical data sheet is associated with the respective extracted textual data as a key-value format of a dictionary.

3.2. Table Detection

Tabular data extraction comprises various tasks such as table detection, table structure recognition and applying OCR. Our applied Di-Plast dataset contains, for example, unbordered tables and semi-bordered tables, but no fully bordered tables. An image processing based method can then be considered for semi-bordered table detection using *OpenCV*, where the contours of horizontal or vertical lines of a semi-bordered table can be detected and the coordinates, width and height of the contours are inserted into a *Pandas*[7] dataframe. Then, the four corners of the table can be detected by subsequently analyzing the dataframe. Image-2 of Figure 1, for example, illustrates such a semi-bordered table detection approach using the *OpenCV* tool.

However, there are considerable challenges if we have to consider all possible types of table templates from various types of technical datasheets. It is extremely hard to detect tables in this approach if new table templates become available. Then, one possible option involves deep learning methods for table detection. Although a deep learning model can overfit when applying very small datasets, we consider Transfer Learning to detect tables on our dataset [18]. For this process, we manually annotated 294 tables among 167 document images (from the Di-Plast dataset, see below). We draw bounding boxes of these tables using the *LabelImg*[8] tool, and subsequently save the manually annotated corpora in PASCAL VOC XML annotation format [19]. Afterwards, we convert the annotated corpora to COCO JSON annotation format [20], as it is quite convenient to use such a custom COCO dataset with the *Detectron2* library, e. g., for evaluation purposes. We consider a pre-trained Faster R-CNN model which is considered as the baseline in TableBank [7] and fine tune the deep neural networks on our Di-Plast dataset. The Faster R-CNN model is a state-of-the-art in computer vision based object detection [21].

[5] https://pypi.org/project/pdf2image/
[6] https://opencv.org/
[7] https://pandas.pydata.org/
[8] https://github.com/tzutalin/labelImg

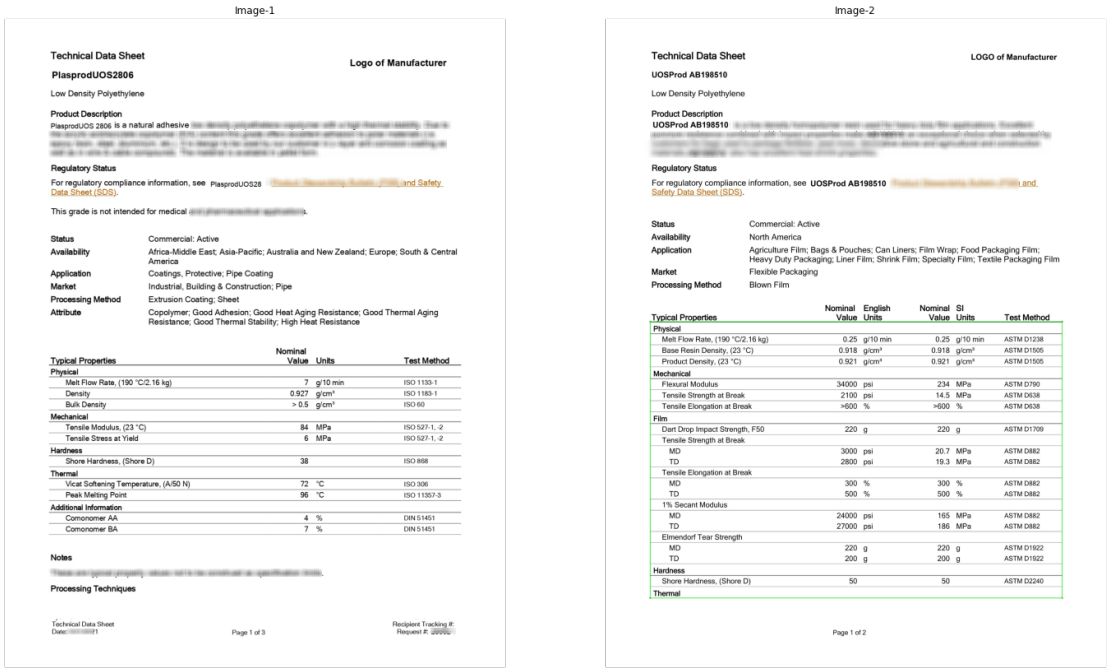

Figure 1: Sample document image and Table detection using the *OpenCV* tool

The other state-of-the-art R-CNN and Fast R-CNN models use selective search to discover the region proposals, although selective search is a slow, time consuming process which influences the performance of the network [7, 22, 23].

Faster R-CNN presents a Region Proposal Network (RPN) which shares full-image convolutional features with the detection network. Therefore, it provides a cost-efficient way to improve object detection accuracy [7, 21]. Faster R-CNN includes two steps. In the first step a RPN proposes candidate object bounding boxes and in second step the features are extracted using RoIPool (Region of Interest Pool) from each candidate box to perform classification and bounding-box regression. RoIPool is a standard operation to extract a small feature map such as 7×7 pixels from each RoI (region of interest) [24].

In table detection, we primarily split our Di-Plast table detection dataset by a nearly 4:1 ratio regarding the number of tables for training and validation. During transformation, we resize the original images into 1333 x 800 pixels along with corresponding ground-truth bounding box coordinates. Afterwards we consider a pre-trained *faster_rcnn_R_101_FPN_3x* model weight and corresponding configuration file[9] from TableBank for transfer learning. We freeze the stem and one residual stage of the backbone ResNet network for fine-tuning on our dataset and adapt few changes such as 2 sub-processes for data loading, 4 training images per iteration, 1 foreground class for table detection, 256 regions of interest (RoIs) per training/mini-batch according to our research work. We consider 50,000 iterations to train our training dataset.

[9]https://github.com/Layout-Parser/layout-parser/blob/master/src/layoutparser/models/detectron2/catalog.py

Table 1
Overview of Di-Plast dataset distribution

Dataset	Total
PDF articles	116
PDF pages in such 116 PDF articles	284
Document images	284
Document images contain tables	167
Document images do not contain any table	117
Di-Plast dataset for table detection	167
Tables in such 167 document images	294

4. Results

We perform our research applying our hybrid information extraction pipeline on 116 PDF articles, which contain 284 pages. Those 284 pages are converted into document images, for which 167 of those contain tables and the other 117 document images do not contain any table. As a result, we consider 294 tables among those 167 document images for table detection. In the following, we refer to the 167 document images as our Di-Plast dataset for table detection. Table 1 summarizes the characteristics as discussed above.

For our experimentation on transfer learning based table detection, we apply a machine equipped with a *Nvidia Quadro RTX6000* graphics card and 24 GB of GPU memory. The table detection experiment is carried out on a *conda environment* with CUDA version 10.2, and PyTorch Version 1.8.0.

4.1. Plain Textual Information Extraction

During the plain textual information extraction process, we extract textual data using OCR from the document images. We perform a layout-structure-driven approach, where we only consider the textual information between two document subtitles in our Di-Plast dataset which follows one specific kind of document layout. We then extract tabular data by preserving inter-word spaces from document images in this approach, while not focusing on complex table structure of plastic product technical data sheets – which is the target of the table detection process. So far, the textual information between two subtitles in our PDF articles maintains single column text layout on each document image. If different column layout for example, two columns or three columns text layout would appear in such PDF articles, then several page orientation operations can be considered during respective pre-processing steps before applying OCR.

4.2. Table Detection Process

The table understanding problem in document layout analysis can be split generally into three categories- table location detection, table structure recognition and table interpretation through semantic interpretation and functional analysis [25]. The table detection problem can be generally classified into a bounding box regression problem and a classification problem.

Table 2
Experimentation for table detection: train, validation, test sets regarding the Di-Plast dataset

Dataset	Training	Validation	Test	Total
Number of document images	130	30	7	167
Number of tables in document images	228	53	13	294

Table 2 presents the details/characteristics of the applied training, validation and test datasets for table detection using transfer learning. The evaluation relies on predicting the tabular regions against the ground-truth rectangular bounding boxes on an image of a pre-defined object class *Table* along with a given confidence score. The confidence score whose value generally is bounded by 0 and 1 shows how confident the detector is about a respective prediction. During the bounding box regression problem, a perfect prediction is determined when the bounding box coordinates of the ground-truth and the predicted rectangular boxes are equal. This is evaluated by the *Intersection over Union* (IoU) metric, which is a measurement based on the Jaccard Index. A perfect bounding box prediction is given when IoU=1. If IoU=0, then the predicted and ground-truth bounding boxes do not intercept each other. The evaluation metric Average Precision (AP) considers respective IoU thresholds for computation and averaging, counting a *positive* when a value above the threshold is observed, otherwise a *negative*. Since strict thresholds for IoU could induce biases in estimation, it is also possible to consider a set of thresholds, averaging over all the (intermediate) results. At IoU=[0.50:0.05:0.95], for example, we calculate the average among all the computed AP results where 10 equal spans of IoU thresholds (t = [0.5, 0.55, ..., 0.95], t is IoU threshold) are considered on 100 detections (maxDets=100) per image. Two other metrics AP_{50} and AP_{75} consider IoU thresholds 0.5 and 0.75 respectively. The metrics AP_s, AP_m and AP_l evaluate small ground-truth objects (with an area less than 32 x 32 pixels), medium ground-truth objects (with an area between 32 x32 pixels and 96 x 96 pixels) and large ground-truth objects (with an area greater than 96 x 96 pixels) respectively, cf. [26].

As the tables in our dataset are larger than 96 x 96 pixels, AP_l is only considered during evaluation of our model. We evaluate our model on the preprocessed Di-Plast dataset for single class object detection – i. e., for a *Table* – and set the minimum testing threshold score to 0.75 for this model. These scores are considered to balance obtaining high recall while not having much low precision detection that slows down inference post-processing steps such as Non-maximum Suppression (NMS)[10]. Table 3 summarizes the evaluation results of our model on the Di-Plast validation dataset for bounding box regression with the above mentioned minimum testing threshold score. We also evaluated the pre-trained TableBank model on the Di-Plast validation dataset in comparison to a direct use of the pre-trained model for bounding box regression with the above mentioned minimum testing threshold score. Table 4 summarizes the evaluation results where the pre-trained TableBank *faster_rcnn_R_101_FPN_3x* model weight and the corresponding configuration file are considered. AP is also entitled as Mean Average Precision (mAP) as it is calculated the average over all categories. Similarly AR and Mean Average Recall (mAR) are entitled interchangeably [11]. Furthermore, another set of evaluation metrics AR_1,

[10] https://detectron2.readthedocs.io/en/latest/modules/config.html
[11] https://cocodataset.org/#detection-eval

Table 3
Summarized results of our proposed model – Di-Plast dataset, bounding box regression

AP	AP_{50}	AP_{75}	AP_l
89.988	100.000	100.000	89.988

Table 4
Summarized results of pre-trained TableBank model – Di-Plast dataset, bounding box regression

AP	AP_{50}	AP_{75}	AP_l
32.939	56.436	41.018	32.939

Table 5
Results of our proposed model – Di-Plast dataset, bounding box regression

Metric	IoU	area	maxDets	Value
AP	[0.50:0.05:0.95]	all	100	0.900
AP	0.50	all	100	1.000
AP	0.75	all	100	1.000
AP	[0.50:0.05:0.95]	small	100	-1.000
AP	[0.50:0.05:0.95]	medium	100	-1.000
AP	[0.50:0.05:0.95]	large	100	0.900
AR	[0.50:0.05:0.95]	all	1	0.542
AR	[0.50:0.05:0.95]	all	10	0.909
AR	[0.50:0.05:0.95]	all	100	0.909
AR	[0.50:0.05:0.95]	small	100	-1.000
AR	[0.50:0.05:0.95]	medium	100	-1.000
AR	[0.50:0.05:0.95]	large	100	0.909

AR_{10}, and AR_{100} evaluates the Average Recall (AR) specified by a fixed amount of detection per image such as 1, 10 and 100 respectively and calculate the average over all classes and IoUs. To measure the recall values, the IoUs are same as in $AP_{[0.5:0.05:0.95]}$ [26]. Table 5 summarizes first evaluation results of our model on the Di-Plast dataset for bounding box regression. Average Precision (AP) and Average Recall (AR) values are in the range from 0.0 to 1.0 – when the boundary boxes of the objects are predicted. For reported Average Precision (AP) and Average Recall (AR) values equal to -1.000, then in this case the metric cannot be computed for small objects (an area less than 32 x 32 pixels) and medium objects (an area between 32 x 32 pixels and 96 x 96 pixels)[12]. Therefore no predictions are performed for small and medium objects, as the area of each table in our Di-Plast dataset is larger than 96 x 96 pixels.

Three typical table detection errors are observed such as partial-detection, un-detection and mis-detection. Only some part of the tables is identified and some information is missing in partial-detection. However some tables in the PDF articles are not identified properly in

[12]https://detectron2.readthedocs.io/en/latest/modules/evaluation.html?highlight=COCOEvaluator#detectron2.evaluation.COCOEvaluator

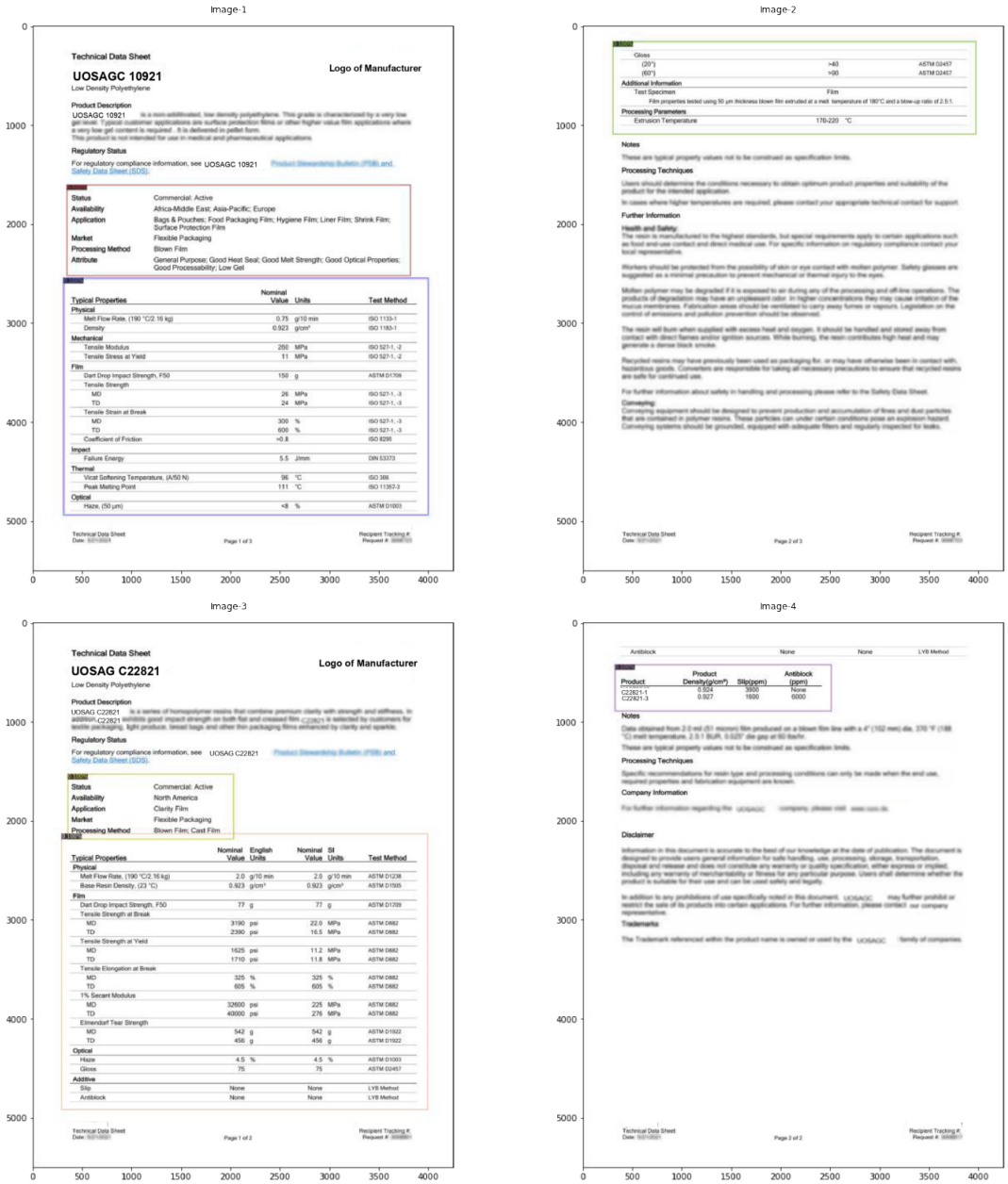

Figure 2: Exemplary instances illustrating table detection inference results of our proposed model.

un-detection. Other components such as text blocks, figures in the PDF articles are occasionally identified as tables in mis-detection [7]. To exemplify our analysis and the respective predictions, we show some exemplary instances as depicted in Figure 2. Here, we indicate exemplary instances of table detection results, illustrating cases as observed on our Di-Plast test dataset.

Regarding current limitations, we observe that different types of tables in technical data sheets have different visual appearances. Here, in general it is rather difficult to rely on transfer learning techniques completely in order to obtain good table detection models with diverse and small amount of training dataset [7], motivating advanced – e. g., hybrid approaches. Furthermore, sometimes a table splits into two consecutive pages in our dataset. Then, specific adaptations are necessary to handle this situation, which we aim to explore in future work.

5. Conclusions

In this paper, we proposed a hybrid information extraction pipeline approach on richly-structured plastic product technical data sheets that includes plain textual information extraction using OCR and deep learning based table detection adopting transfer learning technique. In general, manual data annotation for table detection on such data sheets is time and cost effective task. Moreover, building a deep learning model for table detection from scratch on very small domain-specific dataset can induce an overfitting problem.

We explored the pre-trained model for table detection by utilizing a transfer learning method, fine tuning the model on a rather small dataset to avoid the overfitting problem [18]. Furthermore, we evaluated our table detection model and reviewed generic as well as transfer learning challenges for table detection. In our experimentation, we obtained first promising results.

For future work, we aim to explore further refined architectures as well as to extend our approach using more annotated data as well as integrating knowledge-based approaches. Overall, convolutional neural networks (CNNs) are the primary design paradigm in image recognition problems, as well as for the table detection problem. Here, the use of Transformer needs to be explored in computer vision based document layout analysis domain to leverage the attention mechanism. The Vision Transformer (ViT), for example, is introduced for image classification tasks by alleviating the reliance on CNNs [27]. We aim to explore the use of state-of-the-art Transformer for table detection and table structure recognition to extract tabular data from our dataset. The use of rule-based text annotation and extraction [28], e. g., using the Apache UIMA Ruta framework[13] is another interesting direction to be reviewed for the plain textual information extraction process, cf. [29, 30, 31]. Furthermore, different types of technical data sheets often exhibit different visual appearances. We aim to include a more diverse set of technical data sheets with diverse templates in our dataset. Here, hybrid approaches integrating rule-based approaches including meta-learning are an interesting direction to consider [32, 33].

Acknowledgments

This work has been funded by the Interreg North-West Europe program (Interreg NWE), project Di-Plast - Digital Circular Economy for the Plastics Industry (NWE729).

[13] https://uima.apache.org/ruta.html

References

[1] J. van den Hoogen, S. Bloemheuvel, M. Atzmueller, The Di-Plast Data Science Toolkit – Enabling a Smart Data-Driven Digital Circular Economy for the Plastics Industry, in: Proc. Dutch-Belgian Database Day, Jheronimus Academy of Data Science, 's-Hertogenbosch, Netherlands, 2019, p. (Extended Abstract).

[2] S. Bloemheuvel, J. van den Hoogen, M. Atzmueller, Complex Network Modeling of Supply and Demand Data: An Application Case in the Plastics Recycling Industry (Abstract), Presented at the INSNA Sunbelt XL International Social Network Conference (2020).

[3] S. Marinai, Learning algorithms for document layout analysis, in: Handbook of Statistics, volume 31, Elsevier, 2013, pp. 400–419.

[4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: A review, IEEE transactions on neural networks and learning systems 30 (2019) 3212–3232.

[5] C. X. Soto, C. X. Soto, Visual Detection with Context for Document Layout Analysis, Technical Report, Brookhaven National Lab.(BNL), Upton, NY (United States), 2019.

[6] X. Zhong, J. Tang, A. Jimeno Yepes, Publaynet: Largest dataset ever for document layout analysis, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1015–1022.

[7] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, Z. Li, Tablebank: Table benchmark for image-based table detection and recognition, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1918–1925.

[8] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, M. Zhou, Docbank: A benchmark dataset for document layout analysis, arXiv preprint arXiv:2006.01038 (2020).

[9] Z. Shen, K. Zhang, M. Dell, A large dataset of historical japanese documents with complex layouts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 548–549.

[10] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, D. S. Weld, The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america, arXiv preprint arXiv:2005.01583 (2020).

[11] C. Clausner, A. Antonacopoulos, S. Pletschacher, Icdar2019 competition on recognition of documents with complex layouts - rdcl2019, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1521–1526.

[12] J. Rausch, O. Martinez, F. Bissig, C. Zhang, S. Feuerriegel, Docparser: Hierarchical document structure parsing from renderings, in: Proc. AAAI Conference on Artificial Intelligence (AAAI-21)(virtual), 2021, pp. 1–10.

[13] M. Hansen, A. Pomp, K. Erki, T. Meisen, Data-driven recognition and extraction of pdf document elements, Technologies 7 (2019) 65.

[14] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, C. L. Giles, Learning to extract semantic structure from documents using multimodal fully convolutional neural networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4342–4351.

[15] P. Zhang, C. Li, L. Qiao, Z. Cheng, S. Pu, Y. Niu, F. Wu, Vsr: A unified framework for document layout analysis combining vision, semantics and relations, arXiv preprint arXiv:2105.06220 (2021).

[16] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, W. Li, Layoutparser: A unified toolkit for deep learning based document image analysis, arXiv preprint arXiv:2103.15348 (2021).
[17] P. Singh, S. Varadarajan, A. N. Singh, M. M. Srivastava, Multi-domain document layout understanding using few-shot object detection, in: International Conference on Image Analysis and Recognition, Springer, 2020, pp. 89–99.
[18] W. Zhao, Research on the deep learning of the small sample data based on transfer learning, in: AIP Conference Proceedings, volume 1864, AIP Publishing LLC, 2017, p. 020018.
[19] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, International Journal of Computer Vision 111 (2015) 98–136.
[20] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. arXiv:1405.0312.
[21] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497 (2015).
[22] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[23] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[24] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[25] M. Göbel, T. Hassan, E. Oro, G. Orsi, Icdar 2013 table competition, in: 2013 12th International Conference on Document Analysis and Recognition, IEEE, 2013, pp. 1449–1453.
[26] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, E. A. da Silva, A comparative analysis of object detection metrics with a companion open-source toolkit, Electronics 10 (2021) 279.
[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[28] B. Waltl, G. Bonczek, F. Matthes, Rule-based information extraction: Advantages, limitations, and perspectives, Jusletter IT (02 2018) (2018).
[29] M. Atzmueller, P. Kluegl, F. Puppe, Rule-Based Information Extraction for Structured Data Acquisition using TextMarker, in: Proc. LWA 2008 (Knowledge Discovery and Machine Learning Track), University of Wuerzburg, 2008, pp. 1–7.
[30] P. Kluegl, M. Atzmueller, Integrating the Rule-Based IE Component TextMarker into UIMA, in: Proc. LWA 2008 (Information Retrieval Track), University of Wuerzburg, 2008, pp. 73–77.
[31] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, F. Puppe, Uima ruta workbench: rule-based text annotation, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, 2014, pp. 29–33.
[32] M. Atzmueller, G. J. Nalepa, A Textual Subgroup Mining Approach for Rapid ARD+ Model Capture, in: Proc. 22nd International Florida Artificial Intelligence Research Society Conference (FLAIRS), AAAI Press, Palo Alto, CA, USA, 2009, pp. 414–419.
[33] P. Kluegl, M. Atzmueller, F. Puppe, Meta-Level Information Extraction, in: The 32nd Annual Conference on Artificial Intelligence, Springer, Berlin, 2009. (233–240).