SCI-3000: A Dataset for Figure, Table and Caption Extraction from Scientific PDFs

Filip Darmanović(✉), Allan Hanbury, and Markus Zlabinger

TU Wien, Vienna, Austria
filip.darmanovic.96@gmail.com,
{allan.hanbury,markus.zlabinger}@tuwien.ac.at

Abstract. Extracting figures and similar visual elements from PDFs of scientific publications is important but non-trivial, and progress is impeded by a lack of datasets for evaluation and machine learning. In this work, we describe and publish the *SCI-3000 dataset*, containing 3 000 PDFs of scientific publications (34 791 pages) with annotations of figures, tables, and corresponding captions, from the fields of *computer science*, *biomedicine*, *chemistry*, *physics*, and *technology*. We demonstrate the use of the dataset to benchmark two figure, table, and caption extraction approaches from recent literature: one rule-based and one deep learning-based.

Keywords: Figure Extraction · Table Extraction · Caption Extraction

1 Introduction

Scientific papers are generally published electronically in PDF format. Extracting information from PDFs and making it machine-actionable has proven to be a challenge, sought to be addressed by research in the field of *Page Object Detection* (POD), also referred to as *Semantic Document Segmentation*. The reason behind this challenge is that vector graphics, various symbols, tables, and other miscellaneous elements like page decorations get represented by rudimentary vector drawing commands in a PDF. This makes it difficult to extract individual blocks of text, figures, tables, etc.

Use cases for extracting these elements are numerous, especially in academia. In fields like *biomedicine* and *computer science*, interest in mining figures from previous publications is notably high [19,30]. Examples range from various figure search engines [16,18], to extracting semantic information from graphs [29], to using curated databases for training machine learning models [1,2]. Going beyond the benefits to the respective research communities, having textual descriptions of figures in the form of captions provides input data for training cross-media machine learning systems, which use different forms of the same data to extract deeper semantic meaning, for example, neural networks which

G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14187, pp. 234–251, 2023.
https://doi.org/10.1007/978-3-031-41676-7_14

learn to describe images with natural language [23]. Furthermore, Clark and Divvala [5] show that the number of figures per paper page and the average caption length have been rising steadily over the past few decades, suggesting that the amount of information presented visually has been on the increase relative to plain text.

Although there exist several tools that can take apart a PDF file with varying degrees of success [24], the task of figure, table, and caption extraction is an area with much potential for improvement. For the sake of brevity, we refer to this task simply as *figure extraction* unless explicitly noted otherwise, as in [5,30]. The previously-mentioned internal structure of the PDF poses a challenge for most tools available today, as they are usually not able to discern an entire graphical element, but instead output its individual pieces, like the background, text and so on. Extracted separately, these elements are far less valuable than the entire semantic unit they belong to. Including the corresponding caption further increases the difficulty of this task. While captions contain essential information for understanding figures and tables they describe, the number of document layouts and designs possible makes their extraction difficult. The few tools that extract both captions and graphical elements from scientific publications are, in most cases, usable only on works from specific research fields [19]. This has created the need for more advanced approaches, motivating the recent increase of research in the field of POD [20]. Nonetheless, researchers have been vocal regarding the lack of standardized metrics and datasets for evaluation and machine learning. Most extraction tools from the literature have either been tested on unpublished validation sets, or datasets that are not specifically tailored for the discipline, for example, by including non-scientific publications. Other validation datasets currently available have different issues, e.g., only containing images of pages instead of full PDFs, requiring the user to piece together the dataset from multiple sources, or focusing on only one scientific field or element type. Therefore, addressing the lack of standardized metrics and datasets is a critical research topic.

With that in mind, this paper has three main contributions:

1. A novel dataset, SCI-3000, built by annotating figures, tables, and captions in 3000 documents (34,791 pages) from the fields of *computer science*, *biomedicine*, *chemistry*, *physics*, and *technology*.
2. A suite of tools for evaluating figure, table, and caption detection, as well as annotation of such elements.
3. A SCI-3000-based evaluation of two figure-extraction methods from recent literature; a rule-based approach (PDFFigures2 [5]), and a deep learning-based approach (DeepFigures [30]).

While previous research efforts have predominantly focused on *computer science* and *biomedicine*, they have produced methods that do not perform well on other fields [25]. By including five research disciplines, we make our dataset more general. SCI-3000 also includes the original PDFs of publications. This is in contrast to many other datasets [30] [5] that require the user to manually

acquire them because the underlying licenses prevent redistribution. We publish SCI-3000[1] under CC-BY 4.0.

Finally, our evaluation of existing figure-extraction approaches demonstrates their acceptable effectiveness for tasks where perfect recall or precision is not required. Still, it shows that there is room for improvement, especially regarding caption extraction.

2 Available Annotated Datasets

One characteristic of previous work on figure extraction and on POD in general, is the focus on single scientific disciplines. Arguably, the most focused-on domains are *computer science* (CS) and *biomedicine*.

The former was the most represented discipline in our literature research, with 12 papers either using a predominantly CS-based dataset in their evaluation phase, or focusing on building one. The two most prominent open datasets in this field are the CS-150 [6], containing 150 papers sampled from three CS conferences, and CS-Large [5], with 350 CS papers published after 1999. The ICDAR2013 [8] dataset facilitated multiple challenges regarding table detection and interpretation in PDFs. It was later extended with data on graphs in [14]. A well-used pair of datasets from this group are the ICDAR2016 and ICDAR2017 [7] challenge validation sets, sampled from the CS-focused repository CiteSeer. These datasets were used by Saha et al. [27] and Li et al. [20], before Younas et al. [36] pointed out a lack of quality in the annotations and posted an amended version. These datasets have the disadvantage of only containing rasterized versions (i.e., each page is available as an image) of papers, which means that approaches taking advantage of PDF structure cannot use them. Younas et al. [36] also noted that a dataset with more types of page objects, e.g., captions, is needed to push the field forward. CiteSeer appears to be a popular choice for source material, as many other publications sourced their datasets from it [4,29,35], even though these were never made public. Three papers from Kuzi et al. [16–18] sourced their dataset from the ACL Anthology, which is a repository consisting of work from the areas of *Natural Language Processing* and *Computational Linguistics*. Finally, Chiu et al. [3] sampled 30 papers from two CS conferences: ACM UIST and IEEE ICME. Their test dataset was also not made public.

The second research field in terms of representation was *biomedicine*. PubMed and repositories like Biomedcentral are the main sources for building PDF extraction datasets [22,28,32,34]. One popular dataset that came up during our literature review was the ImageCLEF 2016 Medical dataset [11], used by Tsutsui and Crandall [33] and Yu et al. [37]; however, this is a collection of already extracted images from medical publications. The most recent, and largest dataset in this category is PubLayNet [38] [13], which includes figures and tables along with other typical document elements. It does not, however, include relationships between those elements.

[1] DOI: 10.5281/zenodo.6564971

We found two papers focusing on both *computer science* and *biomedicine*. One is [19], using the previously mentioned CS-150 dataset, and the other is [30]. For the latter dataset, several aspects are missing in terms of usability. While around 5 million annotations were released publicly on Github[2], the licenses of the underlying publications hosted on arXiv do not allow for redistribution without explicit permission from each individual author. PubMed, which they used to source PDFs from the biomedical domain, does have an open publishing arrangement with the authors[3], but the scope of this repository is focused only on biomedicine and other sub-fields of life science. This limitation means that the dataset must be pieced together from three sources. While there is no doubt that these challenges can be overcome, they definitely present hurdles for re-use in future efforts.

3 Evaluation Methodology

Correctly assessing if and how two sets of annotations differ is an essential part of our work. When evaluating existing figure extraction approaches, we need to analyze if their output matches the ground truth. In the crowd-sourced annotation stage (Sect. 4), we need to know if two people agree in their annotation of the same page. Both of these use cases can be served by a single automated annotation assessment system. Furthermore, we argue that using the same methodology in both stages is a requirement for the consistency of our work.

To make sure we implement the correct evaluation strategy, we examined related research on figure extraction from scientific publications. Most researchers described the performance of their extraction approaches through metrics like *accuracy*, *recall*, *F1 measure*, and *precision*. To apply these metrics to bounding boxes, an adaptation of the Jaccard index to 2D space was often used, called Intersection Over Union (IOU) [30]. The IOU is computed by dividing the intersection surface of two bounding boxes by the surface area of their union. Authors of [5,6,20,30] used an IOU of 0.8 as the minimum threshold when deciding if a predicted bounding box matched the ground truth.

Going beyond these similarities however, the information is so scarce, that recreating evaluation setups from most papers becomes impossible. This lack of clarity and standardized evaluation sets was also observed by Choudhury et al. [4]. Even the most influential papers in the broader field of *object detection* like [26] vaguely reference other work instead of giving a detailed description of their evaluation setup. This makes it hard to know how exactly that referenced work was applied when benchmarking new systems. Essentially, we had to resolve three main ambiguities during the implementation of our automated annotation evaluation system:

1. Mapping of bounding boxes between annotation sets.
2. Handling of misclassifications between Figures, Tables and Captions.

[2] https://github.com/allenai/deepfigures-open, accessed on 15.09.2021
[3] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist, accessed on 15.09.2021

3. Evaluation of relations between captions and elements they describe (refer to).

The first ambiguity was addressed by Liu and Haralick [21] by modeling the task as an *optimal assignment problem*. This formulation entails two distinct finite sets, K and L. These represent the two sets of bounding boxes we are comparing. Let there also be a cost function for associating a $k \in K$ with an $l \in L$ denoted with $q(k, l)$ (Euclidean distance between the centers of two bounding boxes in our case). The goal is to find an optimal assignment $a : K \rightarrow L$, such that the sum of costs for all one-to-one mappings is the smallest possible. If the cost is a rational valued function, like in our case, the optimal solution(s) can be found in $O(N^3)$ by applying the Hungarian algorithm [15]. The only change that has to be made to the original problem formulation is to allow K and L to have different sizes, since the prediction and ground truth sets do not necessarily have the same cardinality.

Where our approach differs from [21], and by extension from [12], is the way we handle classes and misclassification errors. In these two papers, detection (localization) and classification errors are handled separately. For example, if a predicted bounding box matches the ground truth in the IOU metric, but its class is wrong, some points are still given. In our case, however, we run the Hungarian algorithm for each class separately, meaning that a correctly detected but misclassified element would incur both a false positive for the predicted class and a false negative for the ground truth class. While our approach makes the evaluation more strict, it simplifies the result, as each prediction can either be entirely correct or incorrect.

In contrast, a more lenient approach is taken when assessing the correctness of relation assignments between captions and tables or figures. More specifically, we run the Hungarian algorithm for all classes together and match each bounding box in one annotation set to its closest corresponding annotation in the other set (if one exists). For every caption-figure/table pair, we then check if the reference relation exists between their respective closest elements in the other annotation set. A true positive is recorded if the corresponding pair of elements is linked in the same manner. When assigning the closest corresponding element, misclassification or IOU do not play a role. Only the proximity between the center points of bounding boxes is considered (Fig. 1). This design decision was made in order to make evaluating the assignment of relations between elements less dependent on the precision of drawn bounding boxes and their predicted classes.

We have made our implementation of the entire evaluation pipeline available as a python package[4] We encourage other researchers to contribute parsers and exporters for a variety of tool outputs to it.

[4] https://pypi.org/project/sci-annot-eval

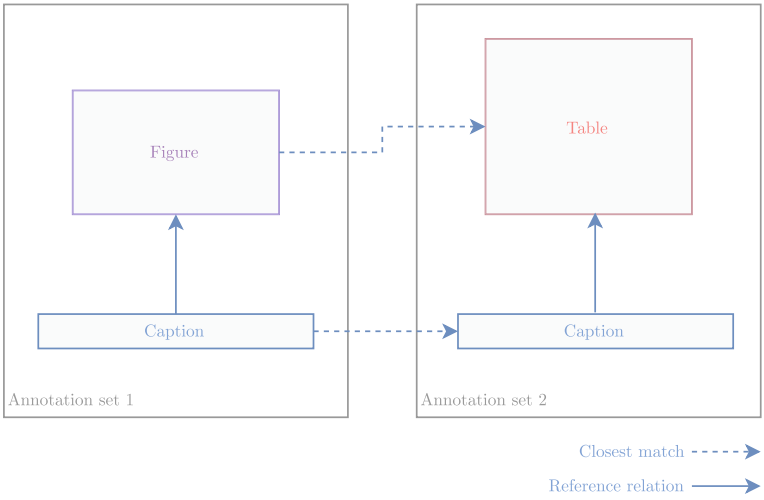

Fig. 1. Example of a matching reference relation across annotation sets. Even though the shape and class of the referenced element is not the same on both sides, we match both bounding boxes by proximity and conclude that the reference relation has been assigned in the same manner.

4 Building the SCI-3000 Dataset

Before starting with the acquisition of PDFs and the crowd-sourced annotation, we set four properties that the dataset must fulfill:

1. The included scientific works must be published with a license that allows redistribution.
2. The dataset should be relevant and useful as a training and benchmarking aid to other researchers in the area of figure extraction.
3. The dataset should facilitate focus on scientific fields that have previously been under-represented as targets of figure extraction research but have a need for such systems.
4. Each scientific field considered in the dataset should have sufficient documents to act as an individual validation dataset, able to produce performance metrics comparable based on statistical significance.

4.1 Data Source and Sampling

To obtain PDFs eligible for redistribution, we turned to the DOAJ[5], a meta-repository hosting millions of open-access publications under the CC BY-SA 4.0 license. DOAJ gives access to the metadata of all indexed work, including journal language, year of publication, and field of research according to the Library of

[5] https://doaj.org/

Congress Classification Scheme (LCC)[6]. The following decisions were made when downloading the papers: (i) to ensure that all papers are in English, we omitted papers from journals that are indicated as publishing papers in languages other than English; (ii) for papers that list multiple fields of research, we took the first field listed as the main one; and (iii) we used the fields at the second level of the LCC hierarchy and mapped all classifications to this level (except for papers that only provided classifications at the top level).

We included the research fields of *computer science* and *biomedicine* due to the existing extensive related work on POD in these areas.

For the further fields, the results of a meta-study [31] on the number and sizes of figures and captions in scientific publications across different research areas helped us identify research fields with an above-average number of per-page figures and caption lengths. We also identified fields in which at least some initial work on POD has been done.

Praczyk et al. [25] focused on the automatic extraction of figures from the field of *high-energy physics*. When referencing *physics* in the previously-mentioned meta-study [31], the authors found that the field has an above-average number of figures (0.8 compared to 0.7), charts (5.7 versus 3.6), as well as caption length (468 characters versus 411), which further reinforces the field as a relevant target of figure-caption extraction approaches.

Choudhury et al. [4] describe an end-to-end figure-caption extraction and search engine system for chemistry. In the meta-study [31], the authors found a slightly above-average number of graphs per paper (3.7 compared to 3.6), as well as caption length (416 versus 411 characters), though the number of images is significantly lower than the mean (0.3 compared to 0.7).

Kuzi et al. [18] explored the use of their system FigExplorer in supporting mechanical failure diagnosis. Referencing the meta-study [31], the field of mechanical engineering has more than double the average number of charts per paper and almost five times more images than the mean; however, the average caption length lies significantly under the mean (119.8 characters compared to 411). Looking at other similar fields, we noticed the same trend, even more pronounced. Therefore, we decided to generalize by including the entire first-level classification of *technology* (T) from the LCC in our dataset.

To summarize, we have identified five research fields for which figure extraction is critical: *computer science*, *medicine*, *physics*, *chemistry*, and *technology*. We equally split the entire corpus into these five research fields and sample 3000 documents, containing 34,791 pages in total. Rasterized versions of these pages were created using version 22.02.0 of Poppler[7], using the default media box cropping. We limited the maximum number of pages per paper to 20 to prevent a sampling bias towards longer publications, which would reduce the variety of visual styles in the dataset. Note that some of the sampled papers are cross-discipline, belonging to more than just one of the five selected research fields.

[6] https://www.loc.gov/catdir/cpso/lcco/

[7] https://poppler.freedesktop.org/, accessed on 24.04.2023

4.2 Data Annotation

We go into detail on how we used Amazon Mechanical Turk (AMT) to crowdsource the annotations of the 34,791 pages. The annotation tool used in this process is available on GitHub[8]

Task Specification. Our task is defined as follows: Bounding boxes have to be drawn around *figures*, *tables* and their corresponding *captions* in rasterized document pages. In addition to determining their location and size, the element to which each caption refers needs to be established. A reference relation has the cardinality of 1:1, meaning that captions refer to a single element and vice-versa, although ones without a reference relation are also allowed. While this should not come up in the context of an entire document, focusing on one page at a time makes elements without a reference possible if they refer to each other across pages [19], like a figure whose caption is on the following page. Another case where this might happen is if a table or figure is broken into multiple parts to fit on one page. Although the problem could be solved by assigning multiple bounding boxes to one element, it makes the system needlessly complicated, so in our formulation of the task, the caption always references its closest part of figure or table it describes, while all others are considered separate elements without a reference.

Submission Review Policy. To ensure that submissions are accepted and rejected consistently and to asses the quality of our dataset, we designed a clear and transparent submission review process. We describe this process and explain how it was used to build a pool of workers for our task.

Our previous experience with AMT has shown that picking a few top workers to annotate the entire corpus is more efficient than opening the task for everyone and manually reviewing erroneous submissions. To rank workers, we built a scoring system by giving a worker one point each time a submission was manually verified as correct by us and subtracting five if it was rejected. This grading disparity is motivated by the fact that around half of the pages have no elements to annotate. Making the reward and penalty equal would make the score a less meaningful indicator of the quality of work. In terms of review criteria, we aim to be fair and only reject submissions that are intentionally wrong. For example, assigning random bounding boxes, skipping clearly visible elements, and submissions that violate our instructions. In cases where less severe mistakes are made, like imprecise bounding boxes, we simply correct the submission. In such cases, the worker is still compensated after 72 h without rewarding or discounting points.

[8] DOI: 10.5281/zenodo.7878627

To build a pool of qualified workers, we submit pages to AMT in batches of 100 and wait until they are all worked through. At this stage, each page is annotated by only one person and subsequently reviewed by us. The batching is done to avoid a small number of workers speeding through all of the possible assignments hoping they would get the payment without review. Since it would be infeasible for us to manually review all thirty-four thousand pages, once the qualified worker pool reaches around 15 members, we start the main annotation phase by posting more tasks and letting two annotators label each page. We set a threshold using AMT's *Qualification* feature so that only workers with over a certain number of points could see and work on them.

Manual disagreement resolution would only be needed when two submissions for the same page have not passed the automated evaluation procedure. It works by first cropping the whitespace in every bounding box and then applying the evaluation framework described in Sect. 3, with an IOU threshold set to 95%. We have released our system for administrating annotation by AMT on GitHub[9]

Task Pricing. To help us determine a fair compensation amount, we turned to observational studies of the crowdsourcing marketplace. Two in-depth studies by Hara et al. [9,10] used an opt-in browser plugin to collect metadata for 3.8 million task instances from AMT, including the compensation. They found that, once unpaid work like searching for tasks was accounted for, the mean and median hourly wages were \$3.31/h and \$1.77/h, respectively. With this way of calculating wages, only 4% of workers earned more than the U.S. federal minimum wage of \$7.25/h. Ignoring the unpaid work, the median and mean wages rise to \$3.18/h and \$6.19/h, respectively. Splitting the earnings by task type, the authors found that the task of image transcription, which is closely related to our work, is by far the lowest-paid task type on the platform, with a median wage of \$1.13/h, while at the same time having the most instances compared to other types.

With the above-mentioned findings in mind, we settle on a price around the U.S. federal minimum wage of \$7.25/h, which we believe is fair considering that the task does not require any special qualifications. The workers are paid per annotated page.

After settling on an hourly wage, we measure the median time needed to annotate a single page and use the result to infer the final compensation amount per page. We enlisted three volunteers that have never done this task or used our tool to annotate 40 pages each. The average annotation time measured in this experiment was between 20 and 35 s. Therefore, we set the payment per page to \$0.04.

Annotation Results. The crowd-sourced annotation process was started by building a pool of qualified workers. The threshold for the worker score was set to 25 and stayed in that range during the entire run. This distilled about

[9] DOI: 10.5281/zenodo.7878638

Table 1. Annotation statistics by research field.

Research Field	Page Count	Figures	Tables	Captions	Empty Pages
Chemistry	7,664	3,880	1,226	5,092	3,981
Computer Science	7,796	4,277	1,784	5,999	4,244
Medicine	6,144	1,752	1,431	3,147	3,872
Physics	5,520	4,025	683	4,703	2,576
Technology	7,667	4,448	1,613	6,027	3,793
Total	**34,791**	**18,382**	**6,737**	**24,968**	**18,466**

15 workers out of the 164 from the phase where each submission was manually reviewed from our side. As the second stage, where each page was annotated by two workers, took us three weeks, a few more runs of single-worker annotations were performed to increase the size of the worker pool. At the end of the process, we had 241 workers and 62,100 submissions but only 20 workers were responsible for more than 90% of them.

77.8% of all annotations in our dataset were the case where a page was annotated by two workers and their submissions passed our automatic evaluation procedure (IOU between the annotations greater than 95%). Since those submissions were nearly identical, a random one was picked as the final annotation in our dataset. The second group of annotations, at 16.4%, resulted from either in-house annotation or corrections to submissions from the initial annotation phase (one worker per page). The final 5.7% are disagreements between workers that we manually resolved by either picking the correct submission or amending annotation mistakes. When taking into consideration only pages that were annotated by at least two workers, we derive an inter-annotator agreement of 93.1%.

Throughout the experiment, workers would send requests to overturn our rejections of their submissions. We handled each of these on a case-by-case basis and always made sure to explain what was wrong with the submission and why we rejected it. In total, our rejection rate was less than 1%, most of which were submissions from the initial pool-building stage.

When analyzing the working time, our initial estimates were correct, as the average time per task was just over 22 s.

A per-field breakdown of the annotated objects is shown in Table 1. The entire annotated dataset contains 18,382 figures, 6,737 tables, and 24,968 captions. Roughly every second page has contains one annotation.

All annotations in the published dataset have had white space surroundings cropped to make them as precise as possible. Full details on all aspects of the dataset are available in the thesis[10] on which this paper is based.

[10] https://doi.org/10.34726/hss.2022.94800

5 Evaluation

We test the performance of two existing approaches for figure, table, and caption extraction from scientific publications on the SCI-3000 dataset.

5.1 Evaluated Approaches

Approaches in this research field of POD can be classified on a spectrum between rule-based systems on one end and machine-learning-based ones on the other. Between them are systems using both approaches in various proportions. The best representative of the rule-based group in terms of impact is PDFFigures 2.0 [5]. This system by Clark and Divvala (including the first version [6]) was referenced by a significant number of papers in the field [4,16,18,19,29–31,33,36].

As a representative approach for figure extraction using machine learning, we selected DeepFigures [30]. This system was trained on the largest dataset for our task currently available, containing over a million papers and over 5.5 million labels. Additionally, the dataset contains works from several fields, including *biomedicine*, *computer science*, *biology*, and *physics*. This should make the model more robust than PDFFigures 2.0, which was fine-tuned only on *computer science* papers. A possible drawback of DeepFigures in the context of our comparison is that it uses PDFFigures 2.0 for detecting and assigning captions to graphical elements, meaning that both systems share the same approach for this sub-task. The authors justify this decision by the way of reduced performance when the model is trained to also identify captions, although they describe a different design that could produce a neural network capable of performing both sub-tasks equally well.

5.2 Experiment Setup

To compute predictions from the selected systems, their respective source codes were downloaded from GitHub[11][12].

During the benchmarks, both systems had difficulties with some documents because of special PDF features or encodings. We skipped thirteen PDFs for Deepfigures and four for PDFFigures 2.0. A lack of output was considered as an empty prediction, meaning that a false negative prediction is counted for each ground truth annotation in skipped documents. We ran both systems on a machine with 8 AMD EPYC 7542 cores, 8 GB of memory, and around 300 GB of storage. PDFFigures 2.0 took three hours to complete, and Deepfigures needed more than 72. However, our aim is not to compare runtimes, and therefore, we have not used any optimizations that could improve these results.

For the actual evaluation process, we use the strategy described in Sect. 3 to get True Positive (TP), False Negative (FN), and False Positive (FP) per-page

[11] https://github.com/allenai/pdffigures2, accessed on 15.05.2022
[12] https://github.com/allenai/deepfigures-open, accessed on 15.05.2022

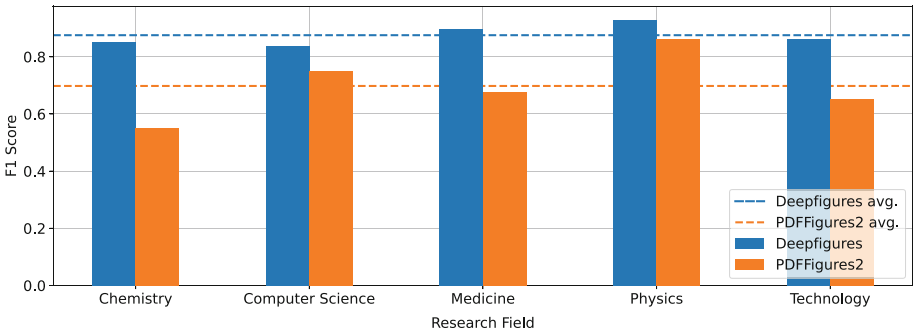

Fig. 2. F1 Score comparison between PDFigures 2.0 and DeepFigures for the task of figure detection. The bars represent average F1 scores grouped by research fields, while the dotted lines represent overall averages.

counts for four element types: *figures*, *tables*, *captions* and *references between captions and their corresponding elements* at an IOU of 0.8. From these counts, we derive the key evaluation metrics: *Precision*, *Recall*, and the *F1* score.

The metrics are calculated per *element type* and *research field*. We use the macro-averaging strategy for any averages displayed in the next section by first computing the mean inside each element group and then using the results to derive the overall average. Because there is a caption for almost every graphical element and a reference relation between each caption and a figure/table, computing averages over all element types in one step (micro-averaging) would skew the performance metrics towards these two larger groups. The same reasoning is applied to research fields, as the amount of graphical elements varies between them.

5.3 Results

An in-depth breakdown of performance metrics is provided in Table 2. We guide the reader through these results in a visual manner, starting with figure-specific performances.

As shown in Fig. 2, there is a substantial performance difference between PDFFigures 2.0 and DeepFigures regarding figure detection. The former reaches an F1 score of 0.68, while the latter does better, with a score of 0.79. The hand-tuned nature of PDFFigures 2.0 can also be seen in the difference in its performance across different research fields. The system seems to struggle with publications in the fields of *chemistry* and *technology*, while *physics* publications seem to be a better extraction target than *computer science*: the field for which the system was optimized. On the other hand, DeepFigures shows a similar F1 score across all research fields.

Moving to table detection (Fig. 3), a similar discrepancy can be seen between the two systems, as both reach almost the same scores for extracting figures, demonstrating the similarities between those two tasks. This time, however,

Table 2. Evaluation results for PDFFigures2 and DeepFigures

		PDFFigures2			DeepFigures		
Research Field	El. Type	F1	Prec	Rec	F1	Prec	Rec
Average	Average	0.68	0.75	0.62	0.79	0.84	0.76
	Caption	0.52	0.51	0.53	0.52	0.56	0.48
	Figure	0.70	0.83	0.60	0.88	0.95	0.81
	References	0.81	0.88	0.75	0.95	0.95	0.95
	Table	0.69	0.79	0.61	0.83	0.89	0.77
Chemistry	Average	0.60	0.68	0.55	0.80	0.85	0.76
	Caption	0.50	0.46	0.54	0.51	0.57	0.47
	Figure	0.55	0.74	0.44	0.85	0.97	0.76
	References	0.69	0.75	0.64	0.94	0.92	0.96
	Table	0.67	0.77	0.60	0.88	0.93	0.84
Computer Science	Average	0.71	0.79	0.65	0.75	0.82	0.70
	Caption	0.54	0.55	0.53	0.52	0.58	0.47
	Figure	0.75	0.87	0.66	0.84	0.93	0.76
	References	0.87	0.94	0.81	0.94	0.95	0.92
	Table	0.70	0.81	0.61	0.72	0.81	0.65
Medicine	Average	0.67	0.77	0.61	0.82	0.86	0.78
	Caption	0.53	0.53	0.53	0.54	0.58	0.51
	Figure	0.68	0.83	0.57	0.90	0.95	0.85
	References	0.81	0.90	0.73	0.97	0.98	0.97
	Table	0.68	0.81	0.59	0.85	0.93	0.79
Physics	Average	0.76	0.81	0.72	0.84	0.87	0.81
	Caption	0.60	0.59	0.60	0.60	0.63	0.58
	Figure	0.86	0.93	0.80	0.93	0.97	0.90
	References	0.92	0.96	0.88	0.98	0.98	0.97
	Table	0.67	0.77	0.59	0.85	0.91	0.80
Technology	Average	0.64	0.71	0.59	0.76	0.80	0.73
	Caption	0.43	0.40	0.46	0.42	0.46	0.39
	Figure	0.65	0.80	0.55	0.86	0.95	0.79
	References	0.77	0.84	0.72	0.94	0.93	0.94
	Table	0.71	0.79	0.64	0.83	0.87	0.79

PDFFigures 2.0 reaches consistent results across research fields. On the other hand, DeepFigures underperforms on publications from *computer science*.

For the task of detecting correct references between captions and tables/figures, both systems performed better than in the previous two (Fig. 4). DeepFigures achieves a precision and recall of 0.95, indicating that even when the underlying bounding boxes do not perfectly overlap with the ground truth,

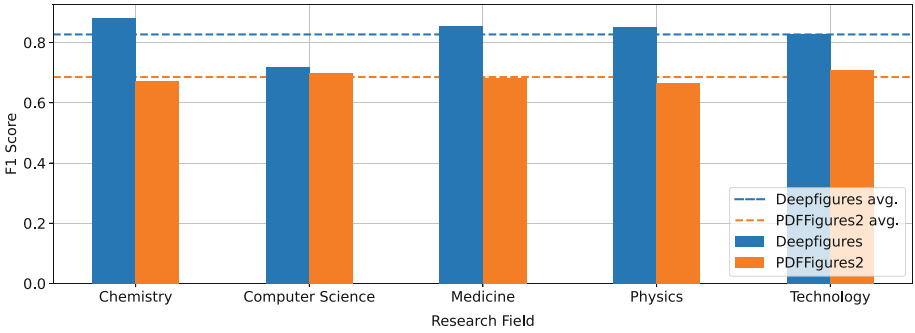

Fig. 3. F1 Score comparison between PDFigures 2.0 and DeepFigures in the task of table detection. The bars represent average F1 scores grouped by research fields, while the dotted lines represent overall averages.

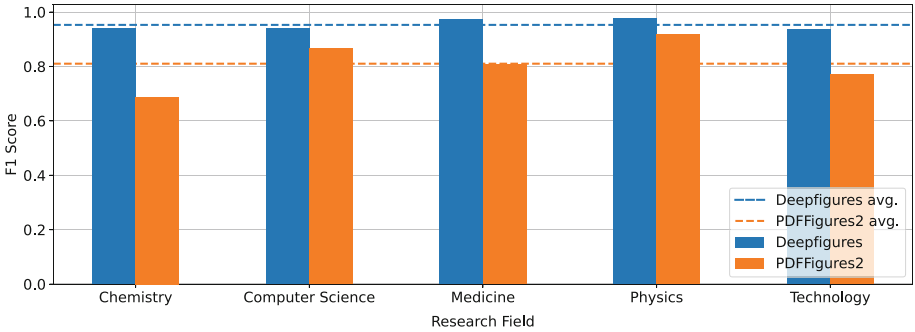

Fig. 4. F1 Score comparison between PDFigures 2.0 and DeepFigures for the task of reference assignment between captions and tables or figures. The bars represent average F1 scores grouped by research fields, while the dotted lines represent overall averages.

the system has a good idea of which elements reference each other. PDFFigures 2.0 achieves an average precision of 0.87 and an average recall of 0.75. The system shows similar performance across research fields for the figure extraction task, suggesting that its reduced performance in reference matching arises from its inability to consistently detect figures.

For caption detection, PDFFigures 2.0 and DeepFigures reach an F1 score of around 0.5. We skip a direct comparison between the systems for this sub-task, as DeepFigures relies on the output of PDFFigures 2.0, making their performance nearly identical. The reason for this performance drop compared to other sub-tasks is that PDFFigures 2.0 often produces caption bounding boxes that do not enclose the entire caption (see example in Fig. 5). This difference is small when considering the absolute area of the boxes; however, the small size of captions makes the relative difference significant enough, that the prediction does not pass an IOU threshold of 0.8. This problem was cited by the authors of DeepFigures [30] as one of the main hurdles in the neural network training process and was

(a) Imprecise bounding box - Produced by PDFFigures 2.0.

(b) Correct bounding box.

Fig. 5. Example of an imprecise caption bounding box produced by PDFFigures 2.0, compared with our annotation.

the reason why they decided to use PDFFigures 2.0 as the underlying caption extraction mechanism. The authors of PDFFigures 2.0 even reported this as an issue in the evaluation phase and used OCR-ed text as a fallback matching technique [5]. The problem could be fixed by snapping the bounding boxes to a grid in order to make them less sensitive to changes, but that introduces another variable to the evaluation process. In our case, the F1 metric for the reference detection task shows that PDFFigures 2.0 is effective at detecting captions but ineffective at precisely defining their bounding boxes.

6 Conclusion

We addressed one of the most prevalent problems currently plaguing research on figure, table, and caption extraction from scientific PDFs: the lack of a large, cross-discipline, and easily-accessible dataset. We published SCI-3000: a novel dataset of annotated scientific publications from five research areas: *computer science*, *biomedicine*, *chemistry*, *physics*, and *technology*. Two state-of-the-art figure, table, and caption extraction methods were evaluated on our dataset, using an evaluation protocol we made publicly available as a python library.

The SCI-3000 dataset not only surpasses most of its predecessors in size and scope by incorporating new scientific fields, but also provides source publications in PDF format, made possible by the permissive licensing of the sourced PDF articles. This characteristic makes the dataset viable for extension and republication, for example, by adding new annotations for elements like equations and paragraphs. An alternative future research path would be to make the available annotations more specific, for example, by classifying figures into different types such as graphs, light-photography, or biomedical images.

Our evaluation of state-of-the-art methods showed that there is still room for improvement, especially for the task of caption detection. Therefore, developing more effective extraction and caption detection methodologies is another viable path for future research.

References

1. Ahmed, Z., Zeeshan, S., Dandekar, T.: Mining biomedical images towards valuable information retrieval in biomedical and life sciences. Database **2016**, baw118 (2016). https://doi.org/10.1093/database/baw118
2. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artif. Intell. Rev. **54**(1), 137–178 (2021)
3. Chiu, P., Chen, F., Denoue, L.: Picture detection in document page images. In: Proceedings of the 10th ACM Symposium on Document Engineering, pp. 211–214. DocEng 2010, Association for Computing Machinery, New York (2010). https://doi.org/10.1145/1860559.1860605
4. Choudhury, S.R., et al.: A figure search engine architecture for a chemistry digital library. In: Proceedings of the 13th ACM/IEEE-CS joint Conference on Digital libraries, pp. 369–370. JCDL 2013, Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2467696.2467757
5. Clark, C., Divvala, S.: PDFFigures 2.0: Mining figures from research papers. In: 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), pp. 143–152 (2016)
6. Clark, C.A., Divvala, S.: Looking beyond text: extracting figures, tables and captions from computer science papers. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015), https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10092
7. Gao, L., Yi, X., Jiang, Z., Hao, L., Tang, Z.: ICDAR2017 competition on page object detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1417–1422 (2017). https://doi.org/10.1109/ICDAR.2017.231, ISSN: 2379-2140
8. Göbel, M., Hassan, T., Oro, E., Orsi, G.: ICDAR 2013 table competition. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1449–1453 (2013). https://doi.org/10.1109/ICDAR.2013.292, ISSN: 2379-2140
9. Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., Bigham, J.P.: A data-driven analysis of workers' earnings on amazon mechanical turk. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–14. Association for Computing Machinery, New York (2018), https://doi.org/10.1145/3173574.3174023
10. Hara, K., et al.: Worker demographics and earnings on amazon mechanical turk: an exploratory analysis. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6. CHI EA 2019, Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3290607.3312970
11. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (2016)
12. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 340–353. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_25
13. Jimeno Yepes, A., Zhong, P., Burdick, D.: ICDAR 2021 competition on scientific literature parsing. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12824, pp. 605–617. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86337-1_40

14. Kavasidis, I., et al.: A saliency-based convolutional neural network for table and chart detection in digitized documents. In: Ricci, E., Rota Bulò, S., Snoek, C., Lanz, O., Messelodi, S., Sebe, N. (eds.) ICIAP 2019. LNCS, vol. 11752, pp. 292–302. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30645-8_27
15. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logistics Q. **2**(1–2), 83–97 (1955)
16. Kuzi, S., Zhai, C.X.: Figure retrieval from collections of research articles. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11437, pp. 696–710. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_45
17. Kuzi, S., Zhai, C.X.: A study of distributed representations for figures of research articles. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12656, pp. 284–297. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72113-8_19
18. Kuzi, S., Zhai, C., Tian, Y., Tang, H.: FigExplorer: a system for retrieval and exploration of figures from collections of research articles. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2133–2136. SIGIR 2020, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3397271.3401400
19. Li, P., Jiang, X., Shatkay, H.: Figure and caption extraction from biomedical documents. Bioinformatics **35**(21), 4381–4388 (2019)
20. Li, X.H., Yin, F., Liu, C.L.: Page object detection from PDF document images by deep structured prediction and supervised clustering. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3627–3632 (2018). https://doi.org/10.1109/ICPR.2018.8546073, ISSN: 1051-4651
21. Liu, G., Haralick, R.M.: Optimal matching problem in detection and recognition performance evaluation. Pattern Recogn. **35**(10), 2125–2139 (2002)
22. Lopez, L.D., Yu, J., Arighi, C.N., Huang, H., Shatkay, H., Wu, C.: An automatic system for extracting figures and captions in biomedical PDF documents. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine, pp. 578–581 (2011). https://doi.org/10.1109/BIBM.2011.26
23. Peng, Y.X., et al.: Cross-media analysis and reasoning: advances and directions. Front. Inf. Technol. Electron. Eng. **18**(1), 44–57 (2017). https://doi.org/10.1631/FITEE.1601787
24. Pitale, S., Sharma, T.: Information extraction tools for portable document format. Int. J. Comput. Technol. Appl. **2**, 2047–2051 (2012)
25. Praczyk, P.A., Nogueras-Iso, J.: Automatic extraction of figures from scientific publications in high-energy physics. Inf. Technol. Libr. **32**(4), 25–52 (2013)
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91, ISSN: 1063-6919
27. Saha, R., Mondal, A., Jawahar, C.V.: Graphical object detection in document images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 51–58 (2019). https://doi.org/10.1109/ICDAR.2019.00018, ISSN: 2379-2140
28. Shao, M., Futrelle, R.P.: Recognition and classification of figures in PDF documents. In: Liu, W., Lladós, J. (eds.) GREC 2005. LNCS, vol. 3926, pp. 231–242. Springer, Heidelberg (2006). https://doi.org/10.1007/11767978_21

29. Siegel, N., Horvitz, Z., Levin, R., Divvala, S., Farhadi, A.: FigureSeer: parsing result-figures in research papers. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 664–680. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_41
30. Siegel, N., Lourie, N., Power, R., Ammar, W.: Extracting scientific figures with distantly supervised neural networks. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 223–232. JCDL 2018, Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3197026.3197040
31. Sohmen, L., Charbonnier, J., Blümel, I., Wartena, Ch., Heller, L.: Figures in scientific open access publications. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J. (eds.) TPDL 2018. LNCS, vol. 11057, pp. 220–226. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_19
32. Stahl, C.G., Young, S.R., Herrmannova, D., Patton, R.M., Wells, J.C.: DeepPDF: a deep learning approach to extracting text from PDFs. In: Proceedings of the 7th International Workshop on Mining Scientific Publications (2018), https://www.osti.gov/biblio/1460210
33. Tsutsui, S., Crandall, D.J.: A data driven approach for compound figure separation using convolutional neural networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 533–540 (2017). https://doi.org/10.1109/ICDAR.2017.93, ISSN: 2379-2140
34. Yang, S.T., et al.: Identifying the central figure of a scientific paper. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1063–1070 (2019). https://doi.org/10.1109/ICDAR.2019.00173, ISSN: 2379-2140
35. Yi, X., Gao, L., Liao, Y., Zhang, X., Liu, R., Jiang, Z.: CNN based page object detection in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 230–235 (2017). https://doi.org/10.1109/ICDAR.2017.46, ISSN: 2379-2140
36. Younas, J., et al.: Fi-Fo detector: figure and formula detection using deformable networks. Appl. Sci. **10**(18), 6460 (2020)
37. Yu, Y., Lin, H., Meng, J., Wei, X., Zhao, Z.: Assembling deep neural networks for medical compound figure detection. Information **8**(2), 48 (2017)
38. Zhong, X., Tang, J., Jimeno Yepes, A.: PubLayNet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015–1022 (2019). https://doi.org/10.1109/ICDAR.2019.00166, ISSN: 2379-2140