A Form Understanding Approach to Printed and Structured Engineering Documentation

Gabriel L. Santos*, Vanessa T. Silva*, Laura A. Dalmolin*, Ricardo N. Rodrigues*,
Paulo L. J. Drews-Jr*, Nelson L. Duarte Filho*
*Centro de Ciências Computacionais - C3
Universidade Federal do Rio Grande, Rio Grande, RS, Brazil
Email: gabriellavoura13@gmail.com

***Abstract*—A significant amount of companies still depends on printed documents, such as healthcare reports, engineering specifications, or historical documents. Those documents are diverse in terms of layout and content, thereby it requires different approaches for each document structure, which makes information extraction a costly and inefficient task. We classify documents into three categories, non-structured, semi-structured, and structured documents. The last one being the focus of the present work. We propose a pattern recognition method for structured documents with an anchoring relationship between question-answer objects through a system of hypotheses and a probability distribution in order to identify which predefined model the document belongs to. Therefore, acting as a system for both identification and content extraction to structured documents. The method has promising results for pattern recognition from all document models, with 78% to 97% objects extracted correctly.**

I. Introduction

In the last years, the importance of digitizing, sharing, retrieving and storing large amounts of data has become evident after the continuous advancements in IT and computer science in general [1].

However, a significant amount of the industry still depends on printed documents, such as healthcare reports, engineering specifications, or historical documents [2]–[4]. Frequently, those documents are diverse in terms of layout and content, which makes automatic information extraction a costly and inefficient task [5]–[7].

Usually, methods for automatic information extraction of printed documents requires additional steps, such as pre-processing the document and transforming the text from an image to searchable text. The latter is achieved through optical character recognition (OCR).

After the mentioned steps, algorithms and techniques can be implemented to attribute semantic meaning and relate the entities identified in the sparse text acquired from the OCR tool. Such approach is called form understanding [8].

It is possible to apply form understanding in non-structured, semi-structured, and structured documents. The last one being the focus of the present work. Describing the structured document's logical structure enables to predetermine a set of instructions to specify how to extract the desired information from data through a rule-based approach [9] [10].

Since the rules of a traditional rule-based system are written manually, the approach tends to have good performance. However, rule-based system present scalability issues because the processing of rules becomes time expensive with an increase in difficulty due to manual description of heuristics [10].

To overcome the above problems, we propose an information extraction method for structured documents with an anchoring relationship between question-answer objects through a system of hypotheses and a probability in order to identify which pre-defined model the document belongs to. Thereby, acting as a system for both identification and content extraction to structured documents.

As a case study, our two datasets were mainly composed of engineering documents from the shipbuilding and offshore industry.

Contributions of the work are as follows: a rule-based approach to cope with printed forms and structured documents together with a probability hypothesis testing, a quantitative evaluation of the feature extraction process for each class of the dataset, it sheds some light on the lack of expressivity and traceability transparency in deep learning approaches, and fair results on extracting information from documents (between 78% and 97%, depending on the document class).

II. Related Work

In order to understand further details on the mentioned subjects, the reading of surveys such as [10]–[13], or [14] are recommended to the reader. In this section some of the relevant work concerning the proposed method will be reviewed.

Information extraction has been a widely studied field among researchers in the last few decades [10]. Among various approaches to the subject, the rule-based technique emerges. It consists in a set of predefined instructions that specify how to extract key information from data [15]. Due to its operating principle, such approach allows an abundance of possible uses at the most variable sets of data and content types, including recipe ingredients [16], [17], healthcare [18]–[20], law documents [21]–[23], stock trends [24], research assistance [25], etc.

Unlike electronic documents, there is an additional difficulty in extracting information from physical documents, and paper documentation. That difficulty is related to the non-existence of digital content, being a previous text extraction step, through optical character recognition (OCR) tools, needed. In other words, an extra OCR step increases the complexity of the

solution, as it is clear in works like [26] which proposes *Odinson*, a framework that targets the extraction of pieces of information based on rules using an inverted index for sentences and document metadata. The framework defines a format for data ingestion in Javascript Object Notation (JSON), using their own custom preprocessing pipeline for segmentation, tokenization, sequence tagging, and parsing, in a way that each identified sentence is indexed individually.

An information extraction method for multi-page printed documents, which works through a set of available classes, is proposed in [27]. The classes are created based on layout and content similarity between documents. Their description is built through a user interface using sample documents submitted to the processing stage, which deliver blocks of objects extracted from the text as an output.

In [4] the design and evaluation of a new scanned medical document management system (SCAN) is presented. Such system uses OCR to transcribe documents to formats compatible with HIS (Hospital Information System) database. SCAN's goal is to use the scanned documents alongside the ones previously available in the database, in order to help healthcare providers retrieve and manipulate patient data within a reasonable time.

An approach using machine learning and deep learning concepts is proposed in [28]. The mentioned work created DocStruct, a model that uses a deep CNN-based model to extract features from semantic contents, layout information, and visual images. DocStruct contains feature extraction modules, such as the feature fusion module and the relation prediction module, and it was verified with both MedForm and FUNSD benchmarks.

In theory, machine learning and deep learning techniques tend to be superior when it comes to information extraction due to their scalability. However, such technique operates as black boxes when it comes to decision traceability transparency. In contrast, rule-based approaches follow a mostly declarative approach leading to expressive and transparent models [29]. Since the implementation of rules in a rule-based-approach is manual, human knowledge is directly transferred to them, assuring their quality.

III. Materials and Methods

A. Dataset

A private dataset, composed of 11 different types of databooks, was used in the experiments. Databooks are sets of technical engineering documents related to normative, construction and assembly processes of the naval and offshore industry, that were digitized and stored in PDF format. The dataset has 5720 pages, adding up to 5GB, and contains documents from two different natures, divided into class 1 and class 2.

The class 1 consists of documents related to a technical normative of an engineering project and the class 2 consists of documents related to materials used during the execution of the previously mentioned projects.

In Fig 1 is shown an example of a document from class 1 and a document from class 2.

1) Class 1: - Documents named class 1 are structured documents that describe normative techniques related to construction and assembly processes of the naval offshore industry. The documents have their content displayed as forms. Each one of their pages was placed in different layout models: A1, A2 and A3. The A1 model represents a cover page in portrait orientation, and the A2 and A3 layout models represent content pages in the same orientation. Each page has in average 28 fields to be filled in with information such as the project's name, the document's number, related documents and date. A fourth layout, named 'undefined', which is the union of all the other layouts, excluding the fields that are common to multiple layouts, is used when the system fails to identify the correct layout to a certain page. Class 1 documents are PDF files that consist of a cover page (A1) and content pages that either fit into a predefined layout (A2, A3) or are set as 'undefined'. The dataset related to Class 1 documents has 2GB of data and 2104 pages, in total, that are allocated among the four layout models.

2) Class 2: - Class 2 documents represent structured documents related to industrial and assembly. The documents contain information concerning specific inspections that occur during the naval construction process. There are 7 types of possible documents and in general their content is based on forms and tables that describe header information and materials used to perform the inspection, as well as tables containing results and reference papers. The documents also contain a footer with information about the inspection chief. Such documents are generally longer than those on Class 1 and each PDF contains only one specific layout. There are 7 types of documents, a total of 3616 pages or 3GB of databooks.

B. OCR and TesseractOCR

Optical Character Recognition (OCR) is a process that allows extracting editable text from documents digitized as images. Thus, it is possible to alter and process documents automatically through computing systems [30].

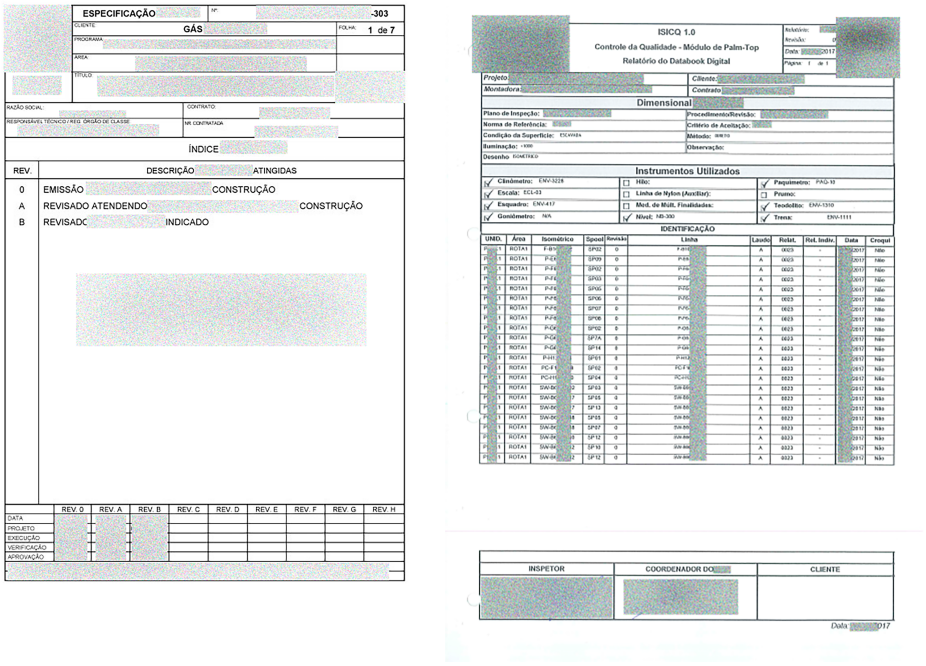
TesseractOCR is a powerful open-source OCR engine that supports numerous languages. It was originally developed by Hewlett Packard (HP), later acquired and maintained by Google. TesseractOCR is currently available on GitHub [30] and is a good alternative among the ones on the market [31].

C. An overview of the method

In this section, we discuss an overview of the proposed method and in the following topics a more detailed discussion of some of the key elements that were major contributors to the process of form understanding.

We propose a new rule-based method for document classification and extraction of semantic elements from the document. We designed our approach based on the method proposed by [27].

First, we pre-processed the documents to increase the quality of the images that compose them and facilitate data



(a) Class 1 document. (b) Class 2 document.

Fig. 1. Example of documents present in dataset.

extraction. As pre-processing, we applied noise reduction, binarization, skew detection, and reduction. Then, TesseractOCR was used to transform image text into editable text.

Next is the first step of form understanding, which is responsible for classifying documents through a probabilistic approach that regards layout similarity between document and models previously defined by users. Models are schemes that contain a set of elements describing the layout and the content of each part of the document.

For data extraction, we used an anchoring system in which elements of the document are identified as text, questions, or answers. Both text and question types are elements whose target value and position are known. Therefore, Levenshtein's Distance is used to identify words of interest.

On the other hand, the coordinates are known in answer-type elements, but their content is not. Thus, attributing a semantic relation to the information demands anchoring the answers to question-type elements. In that way, a hierarchical link is established between the elements, configuring the question-answer bond implemented in the algorithm.

Finally, the system provides an output file that describes the document in JSON format. Each page of the document is represented by as a set of blocks. Each block consist of a list of objects that describe the entities found in the document. The entities have information related to their types, such as unique identification, extracted words, the bounding box as a tuple, the information whether it has a link to another document or not, and a confidence value between 0 and 1. The confidence level is related to a block hypothesis, is discussed in more detail next.

The representation of both, form understanding method and the descriptive format of models were developed based on the document description proposal on [8]. All form understating steps are described in the Fig 2.

1) Document Classification: Document classification can be executed in two different ways, depending on the dataset's class.

For class 1, document classification's algorithm extracts blocks from the page using block hypothesis probability along with block probability, being the predefined models used as a reference.

The model that extracts the most number of blocks is considered correct and, as a reference, is applied to the following metrics. Since the correct model is used in the function responsible for drawing the matching fields and their IDs upon the document, a visual analysis of the results is delivered as well, as it shown in Figures 4b and 5b.

In contrast, identification of the document's title is the basis of model classification for class 2 documents. Title information is extracted using the Regex tool through layout analysis with the coordinates of possible title positions mapped. Therefore, the algorithm can identify which model does the title belong to. Thus, element extraction is made based on the selected model's rules.

2) Block Hypothesis: The system does not know how many words each block has, since it varies from document to document, we decided to group sequences of close words together and call it block hypothesis.

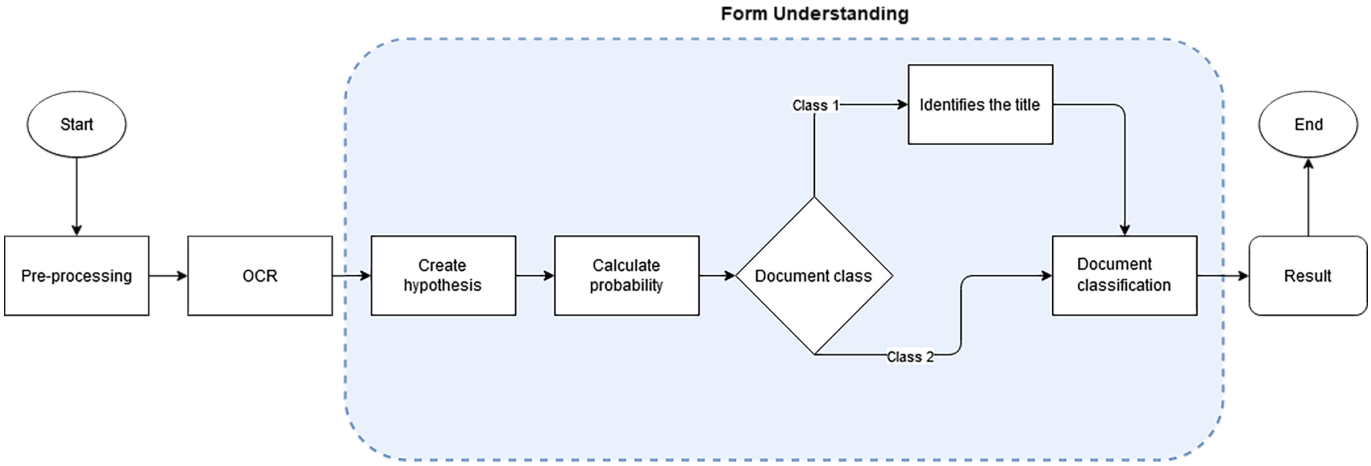


Fig. 2. Process flow of the form understanding method

Formally, block hypothesis consists of a set of up to five words in the same line. We extracted them in a way that a single word is considered a block hypothesis by itself, as well as a set containing the word and up to four posterior words, are too.

Such process is demonstrated by Equation 1, where t_i represent a word from a document containing a total of T words. The set of all hypothesis is composed by the union of sequences up to 4 words long:

$$h = \bigcup_{i=1}^{T} \{(t_i), (t_i, t_{i-1}), (t_i, t_{i-1}, t_{i-2}), (t_i, t_{i-1}, t_{i-2}, t_{i-3})\} \tag{1}$$

Then, all hypothesis are converted into a set of features h_{feat} following the tuple 2:

$$h_{feat} = \langle\, x, y, w, h, t \,\rangle \tag{2}$$

Wih t as the word and the set of information, $\langle x, y, w, h \rangle$ as coordinates of the x, y, and size (width and height), representing the word's bounding box.

3) Hypothesis probability: is calculated for each block belonging to the predefined models, as in Equation 3. p_{bb} and p_t represent respectively the probability of the bounding box and the probability of the text identified on the hypothesis.

Hypothesis' probability value is used to define whether the hypothesis belongs to a block in the model or not. Thereby, it attributes a similarity score between the document being processed and the model in question.

$$p_h = p_t * p_{bb} \tag{3}$$

a) Probability of the Bounding box: There are three different ways to calculate the probability of the bounding box.

When the block has the type '$region$' (it is a question) a verification is made regarding hypothesis' coordinates h_{bb} and block's coordinates q_{bb}. Such analysis uses uniform distribution, as shown in Equation 4.

$$p_{bb} = \begin{cases} 1, & \text{if } h_{bb} \subset q_{bb} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

For a block with the type '$anchored - region$' (it is an "answer"), just as the previous approach, a verification is made regarding the hypothesis' coordinates and block's coordinates, as shown in Equation 5. However, since we anchored the answers to the question, the block coordinates are given as $a_{bb} = anc_{bb} + q_{bb}$, where anc_{bb} represents the answer's bounding box coordinates.

$$p_{bb} = \begin{cases} 1, & \text{if } h_{bb} \subset a_{bb} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

b) Probability of the text:: The text identified by the hypothesis has its probability calculated using Levenshtein's Distance. Among the possible words found, we chose the one that presents the best distance value, as shown in Equation 6.

$$p_t = 1 - max\ (lev\ _{distance}) \tag{6}$$

4) Block probability:: is responsible for defining which model, among all models tested through hypothesis probability, is the most similar to the processed document. To do so, were create a matrix $P \in R^{H \times B}$, where H represents the amount of hypothesis and B represents the number of blocks found in the model. The matrix P receives p_h values.

The block probability approach aims to find the greater value of hypothesis probability to each block individually. The algorithm remains searching for a better hypothesis until the probability value exceeds a certain threshold value.

In cases where two or more hypotheses present the same probability value, an analysis based upon the word's length, t_{size}, is executed by choosing the word with the highest length value. The mentioned procedure is demonstrated by Equation 7.

$$P_{ij} = max\ (\ p_{h_{ij}}\ (\ max\ (t_{size}\ )\ )\ ) \tag{7}$$

Each block is then stored with information such as id, type, bouding box (x, y, w, h), text, links and conf. $Links$ are used to reference the question anchor for each answer and $conf$ represents the probability attributed to the block in question.

IV. Experiments and Results

In this session, the experiments performed are explained, as well as the results obtained by the proposed method.

All of the experiments were performed using documents provided by a shipbuilding and offshore industry. Since the documents contain sensitive data, their content and images were withheld so that the required confidentiality is maintained.

We search for entities found over the text to evaluate the results obtained by the method in the JSON output. All documents were subjected to every stage in the pipeline, starting with pre-processing and ending up with the relevant information extracted and structured as described by the identified template. Each class document was evaluated individually.

For class 1, the Fig 4a show an example of document before the method. In Fig 4b a logic description drawn by the method is demonstrated, where question-type objects are in red, answer-type objects are in blue and the anchoring connections between two or more objects are in green. Finally, in Fig 4c only objects identified by the method are exhibited.

In Fig 5a, an example of class 2 document is demonstrated. In Fig 5b, just like in the other class, a logic description drawn by the method is demonstrated, where question-type objects are in red, answer-type objects are in blue and the anchoring connections between two or more objects are in green. In Fig 5c only objects identified by the method are exhibited.

The documents contain a lot of sensitive data, therefore it is difficult to provide examples of question-type elements and answer-type elements, but Fig 3 shows a result sample of non-sensitive data detected by our method. The word "data", date in Portuguese, is a question-type entity, and the date itself is an answer-type entity. Title of entities are in red, blue represents bounding-boxes and green the connection between question and answer.

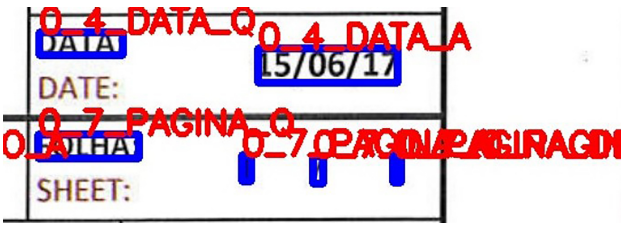


Fig. 3. Sample of data filed question-type and answer-type entities detected by the method.

A. Experiments

To validate and obtain comparison metrics, it was necessary to count how many entities of interest the documents had. The mentioned process was performed through a validation algorithm that compared the expected number of elements and the number of elements found by the information extraction system. For that, the entities were separated in three classes, which are text, questions and answers. The errors obtained due to the OCR tool were not considered. Such errors can be described as every word that could not be found because they simply did not exist or because their edition value is smaller than than the expected word for that element.

Each document in both class 1 and class 2 have tables with a varying number of lines, so we calculated an average number of table lines per specific page for each document model.

Each page takes on average 13.5 seconds to be executed, therefore 21 hours and 27 minutes of experiments in total.

As a baseline, the run time to perform the page description to JSON format through an annotation tool was 15 minutes, while the description made manually by a volunteer was 4 hours.

B. Results

Table I presents the results related to class 1 documents. The A1 and A2 models extracted from 85% to 98% of the text-type objects, being text and questions values superior to answer values.

Class 1 documents present tables with a varying number of lines, which configures a semi-structured document. The majority of answer-type objects' results are related to the mentioned tables, so the method delivers inferior results compared with fixed fields, such as text and question.

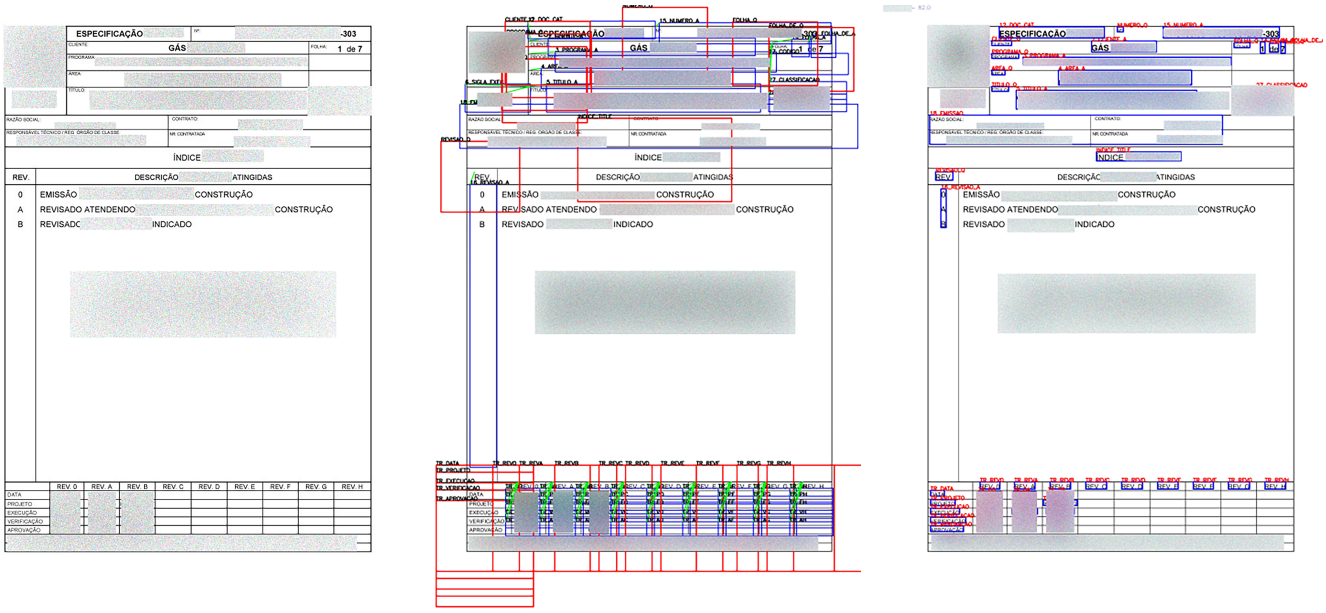
A3 documents achieved superior results due to their scarcity, less than 5% of the dataset. On further experiments, we hope to obtain more A3 documents and reevaluate the extraction method.

No results for the 'undefined' type were found because the system found no document that it could not extract a specific number of blocks n. That condition had been defined as a trigger to attribute a document to the 'undefined' category. We empirically defined the n number after experiments in the dataset during the construction of template models.

TABLE I
Values of success in extraction process of class 1 documents.

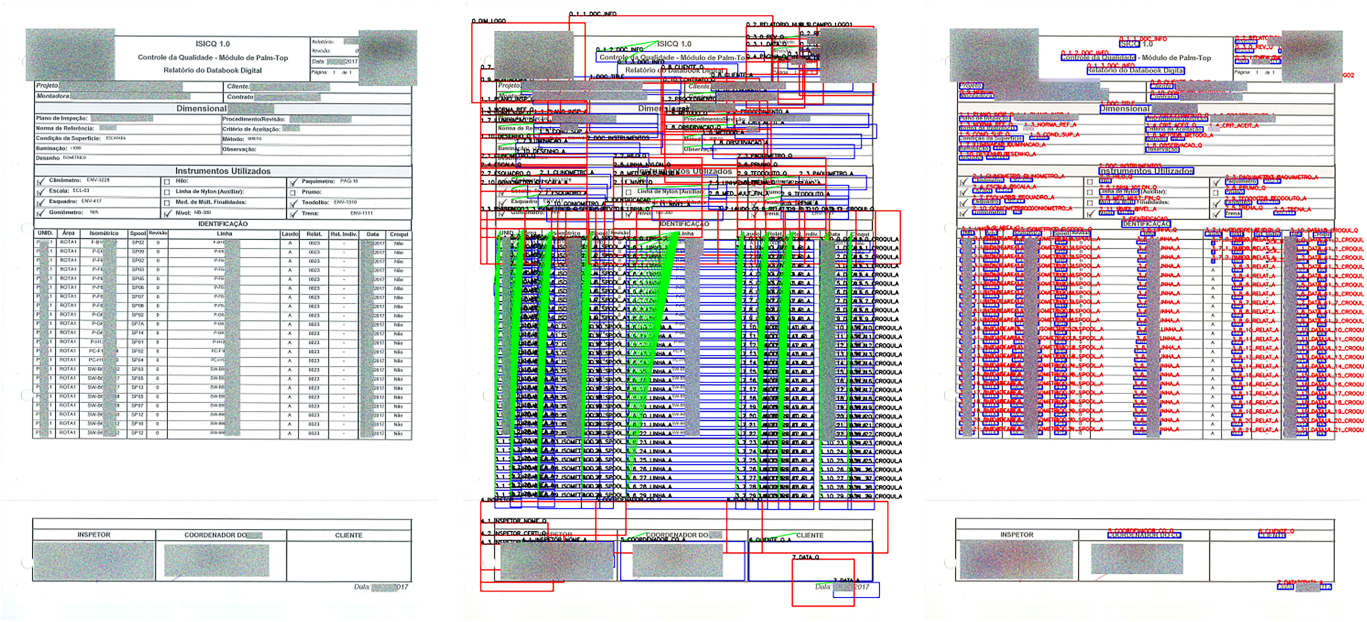
Document Type	Text	Questions	Answers
A1	0.98	0.97	0.95
A2	0.97	0.85	0.85
A3	1.0	1.0	1.0

Results of the form understanding method for class 2 point that from 78% to 97% of text-type objects were identified, as can be seen in table II. The model of type 01 presented the greater values on text and question categories, while models 04 and 06 achieved better results for the answer category.



(a) Original. (b) With the model drawn. (c) Result.

Fig. 4. Comparison between steps in processing pipeline of class 1 documents.



(a) Original. (b) With the model drawn. (c) Result.

Fig. 5. Comparison between steps in processing pipeline of class 2 documents.

As shown in the latter dataset, fixed objects presented greater extraction values than the answer object. Question-type, for instance, presented extraction values of above 91% in all models.

Text-type objects, despite their fixed layout, presented inferior values because logos present in the documents fit in the mentioned category.

The method is able to identify one of the available logos, however the second consists only of visual elements. Since there is no text content and the method's proposal is to extract text information, extracting information from that logo is above the scope of the method.

Evidently, in general, all models achieved promising extraction results in both classes, demonstrating the good performance of the model in reference to the proposed task.

TABLE II
VALUES OF SUCCESS IN EXTRACTION PROCESS OF CLASS 2 DOCUMENTS.

Document Type	Text	Questions	Answers
01	0.97	0.96	0.86
02	0.87	0.94	0.83
03	0.87	0.93	0.80
04	0.88	0.91	0.90
05	0.91	0.91	0.79
06	0.88	0.91	0.90
07	0.88	0.93	0.78

V. CONCLUSION

In this paper, we proposed a method for information extraction through block hypothesis and an anchored linked system between question-answer objects. We used probability and uniform distribution to choose the right model for the document among the layout models pre-defined. Then, the extraction process was based on the template description of the correct model.

As a case study, we divided documents from the shipbuilding and offshore industry into two distinct classes. The method has promising results for extracting information from all document models, with 78% to 97% objects extracted correctly.

The proposed method can help reduce the time spent on trivial analysis in those types of documents and extract a large amount of information from the dataset.

Although improvements need to be made, specifically in documents with page orientation as landscape, the method worked effectively and achieved positive results on the aimed task.

As future work, we intend to improve the information extraction in structured documents to up to 90% in all documents minimum, explore different probability distributions and methods, as well as implement a method for information extraction of semi-structured and unstructured documents based on deep learning to automatically set up the models to work for other cases and to avoid both user effort to define hundreds of templates and model pre-definition.

In the future other methods will be used to extract information from the logos, therefore they were taken into consideration during the construction of the template of each document model.

ACKNOWLEDGMENT

The authors would like to thank CNPQ - Conselho Nacional de Desenvolvimento Científico e Tecnológico and CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

REFERENCES

[1] L. Bstieler, T. Gruen, B. Akdeniz, D. Brick, S. Du, L. Guo, M. Khanlari, J. McIllroy, M. O'Hern, and G. Yalcinkaya, "Emerging research themes in innovation and new product development: insights from the 2017 pdma-unh doctoral consortium," 2018.

[2] Y. Qi, W. R. Huang, Q. Li, and J. Degange, "Deeperase: Weakly supervised ink artifact removal in document text images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3522–3530.

[3] S. Patel and D. Bhatt, "Abstractive information extraction from scanned invoices (aiesi) using end-to-end sequential approach," *arXiv preprint arXiv:2009.05728*, 2020.

[4] E. Promin and P. Suriyachai, "Improvement of scanned medical document management system," in *2019 11th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 2019, pp. 126–131.

[5] H. Cha and D. Lee, "Framework based on building information modelling for information management by linking construction documents to design objects," *Journal of Asian Architecture and Building Engineering*, vol. 17, no. 2, pp. 329–336, 2018.

[6] S. Das, P. Banerjee, B. Seraogi, H. Majumder, S. Mukkamala, R. Roy, and B. B. Chaudhuri, "Hand-written and machine-printed text classification in architecture, engineering & construction documents," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 546–551.

[7] K. Matsubayashi, A. Yamashita, H. Nonaka, and Y. Konno, "A research on document summarization and presentation system based on feature word extraction from stored informations," in *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 2018, pp. 60–63.

[8] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2. IEEE, 2019, pp. 1–6.

[9] J. André, R. Furuta, R. K. Furuta, and V. Quint, *Structured documents*. Cambridge University Press, 1989, vol. 2.

[10] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: a survey," *Scientometrics*, vol. 117, no. 3, pp. 1931–1990, 2018.

[11] P. N. Golshan, H. R. Dashti, S. Azizi, and L. Safari, "A study of recent contributions on information extraction," *arXiv preprint arXiv:1803.05667*, 2018.

[12] S. Alves, J. Costa, and J. Bernardino, "Information extraction applications for clinical trials: A survey," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2019, pp. 1–6.

[13] M. Mannai, W. B. A. Karâa, and H. H. B. Ghezala, "Information extraction approaches: A survey," in *Information and Communication Technology*. Springer, 2018, pp. 289–297.

[14] S. F. Joan and S. Valli, "A survey on text information extraction from born-digital and scene text images," *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, vol. 89, no. 1, pp. 77–101, 2019.

[15] S. Klink, A. Dengel, and T. Kieninger, "Document structure analysis based on layout and textual features," in *Proc. of International Workshop on Document Analysis Systems, DAS2000*. Citeseer, 2000, pp. 99–111.

[16] G. Popovski, S. Kochev, B. Korousic-Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction." in *ICPRAM*, 2019, pp. 915–922.

[17] N. Silva, D. Ribeiro, and L. Ferreira, "Information extraction from unstructured recipe data," in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, 2019, pp. 165–168.

[18] E. Yehia, H. Boshnak, S. AbdelGaber, A. Abdo, and D. S. Elzanfaly, "Ontology-based clinical information extraction from physician's free-text notes," *Journal of biomedical informatics*, vol. 98, p. 103276, 2019.

[19] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS one*, vol. 12, no. 6, p. e0179488, 2017.

[20] S. R. Jonnalagadda, A. K. Adupa, R. P. Garg, J. Corona-Cox, and S. J. Shah, "Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of hfpef patients for clinical trials," *Journal of cardiovascular translational research*, vol. 10, no. 3, pp. 313–321, 2017.

[21] J. Lee, J.-S. Yi, and J. Son, "Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp," *Journal of Computing in Civil Engineering*, vol. 33, no. 3, p. 04019003, 2019.

[22] M. García-Constantino, K. Atkinson, D. Bollegala, K. Chapman, F. Coenen, C. Roberts, and K. Robson, "Cliel: Context-based information extraction from commercial law documents," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 2017, pp. 79–87.

[23] F. Solihin and I. Budi, "Recording of law enforcement based on court decision document using rule-based information extraction," in

2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2018, pp. 349–354.

[24] E. T. Khaing, M. M. Thein, and M. M. Lwin, "Stock trend extraction using rule-based and syntactic feature-based relationships between named entities," in *2019 International Conference on Advanced Information Technologies (ICAIT)*. IEEE, 2019, pp. 78–83.

[25] V. Makhija and S. Ahuja, "Rule based text extraction from a bibliographic database." *DESIDOC Journal of Library & Information Technology*, vol. 38, no. 1, 2018.

[26] M. A. Valenzuela-Escárcega, G. Hahn-Powell, and D. Bell, "Odinson: A fast rule-based information extraction framework," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 2183–2191.

[27] E. Medvet, A. Bartoli, and G. Davanzo, "A probabilistic approach to printed document understanding," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 14, no. 4, pp. 335–347, 2011.

[28] Z. Wang, M. Zhan, X. Liu, and D. Liang, "Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding," *arXiv preprint arXiv:2010.11685*, 2020.

[29] B. Waltl, G. Bonczek, and F. Matthes, "Rule-based information extraction: Advantages, limitations, and perspectives," *Jusletter IT (02 2018)*, 2018.

[30] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open source ocr tool tesseract: A case study," *International Journal of Computer Applications*, vol. 55, no. 10, pp. 50–56, 2012.

[31] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, "Ocr as a service: an experimental evaluation of google docs ocr, tesseract, abbyy finereader, and transym," in *International Symposium on Visual Computing*. Springer, 2016, pp. 735–746.