# ICDAR2021 Doctoral Consortium

(in conjunction with ICDAR 2021)

Lausanne, Switzerland

Date: September 7, 2021

Location: EPFL - École Polytechnique Fédérale de Lausanne,
BC building

***Chairs: Nibal Nayef and Jean-Christophe Burie***

# INTRODUCTION

In 2011, the leadership teams of TC-10 and TC-11 jointly organized the first Doctoral Consortium in conjunction with ICDAR 2011. Its success motivated repeating the initiative in conjunction with each new edition of ICDAR: ICDAR 2013 (in Washington D.C., USA), ICDAR 2015 (in Nancy, France), ICDAR 2017 (in Kyoto, Japan) and ICDAR 2019 (in Sydney, Australia). The Doctoral Consortium at ICDAR 2021 gives continuity to this tradition, creating a unique opportunity for Ph.D. students to test their research ideas, present their current progress and future plans, and to receive constructive criticism and insights related to their future work and career perspectives. For that, a mentor (a senior researcher who is active in the field) has been assigned to each participant to provide individual feedback on the student´s Ph.D. project. In addition, students also have the opportunity to present an overview of their research plan during a special poster session.

The ICDAR 2021 Doctoral Consortium has accepted 16 students, which have been mentored by 16 senior active researchers of all nationalities working in the field of Document Image Analysis and Recognition. During the DC, which is organized this year as an hybrid (remote & onsite) event, students present research proposals through a teaser/poster session, focusing on the outline of the objectives, the methodology, the expected results, the state of the art in their area, and the current stage of their research.
During the teaser (introductory) session, each student makes a brief presentation of his/her research to the public, inviting them to attend the poster session in which the students and all interested attendees discuss projects' details. A jury of senior researchers will select best poster presentation, and the award will be announced at the end of the DC event.

# PROGRAM

| | |
|---|---|
| **2:00 – 2:15 pm** | Opening - Introduction to ICDAR Doctoral Consortium 2021<br>*Nibal Nayef & J-C Burie* |
| **2:15 – 2:45 pm** | Brief presentations (teasers) of the projects by the PhD students |
| **2:45 – 3:15 pm** | Talk "How to succeed in your Ph.D. degree"<br>***Prof. Daniel Lopresti,*** *Lehigh University, Bethlehem, USA* |
| **3:15 – 3:30 pm** | Setting-up of posters / beginning of poster session and discussions |
| ***3:30 – 4:00 pm*** | *Coffee Break / Poster session and discussions* |
| **4:00 – 5:15 pm** | Poster session and discussions |
| **5:15 – 5:30 pm** | Concluding remarks and best poster award<br>*Nibal Nayef & J-C Burie* |

# OVERVIEW OF CONTRIBUTIONS

The Ph.D. research plans of the ICDAR-DC attempt to bring innovative solutions for research problems relevant to the community of ICDAR (Document Analysis and Recognition). In this 6th edition of the DC, we had works in popular recurring topics such as: handwritten text recognition (in historical or modern document images), and layout analysis (based on semantic or structural information). Multiple works in natural language processing (NLP) topic have been presented. NLP is traditionally not a typical area of ICDAR but has recently become a topic of interest to the ICDAR community for semantic information extraction from documents. Another interesting topic is text recognition in natural scenes where it is more challenging to recognize the text than in regular document images (legal, historical …). The trend in all the presented approaches – as it is in other related research communities – is the use of deep learning-based methods. In addition to utilizing such methods, some of Ph.D. works aim to provide theoretical and/or experimental approaches towards understanding the deep neural networks for document analysis applications. Finally, we had a couple of diverse topics related to pattern analysis in general.

The 16 projects of this year may be clustered in the following research areas:

- Recognition of (historical) handwritten documents (**4**)
- Scene text recognition and its applications (**2**)
- Layout analysis (**3**)
- Natural language processing and semantic analysis (**4**)
- Theoretical deep learning (**1**)
- Miscellaneous pattern analysis topics (**2**)

| **Student** | **Ph.D. Topic** | **Country** |
|---|---|---|
| Raphaela Heil | Computerised Image Processing for Handwritten Text Recognition of Historical Manuscripts | Sweden |
| Rubèn Pérez Tito | Text based Visual Question Answering | Spain |
| Christoph Zaugg | Information Theoretical Approach to Understand Deep Neural Networks | Switzerland |
| Ibrahim Souleiman Mahamoud | Automatic and model-free learning of semantic-structural links of fields in a document | France |
| Iheb Brini | Towards an Explainable Deep Model for Archival Document Image Segmentation | Tunisia |
| Tien Nam Nguyen | Segmentation, Recognition and Indexing of characters in CHAM documents | France / Vietnam |
| Francesco Lombardi | Towards semantic understanding of scientific papers | Italy |
| Daichi Haraguchi | Font Design Analysis: Understanding Designers' Knowledge by Using Machine Learning | Japan |
| Sanket Biswas | Structural Analysis and Understanding of Complex Layouts in Document Images | India |
| Solène Tarride | Automatic recognition of historical handwritten parish records. | France |
| Sahar Arshad | Automated Summarization of Legal Judgements | Pakistan |

| Student | Ph.D. Topic | Country |
|---|---|---|
| Iqra Basharat | Information Extraction from Legal Documents Based on Natural Language Processing (NLP) | Pakistan |
| Mengbiao Zhao | Weakly-Supervised Scene Text Detection | China |
| Thomas Constum | Optical Handwritten Named Entity Recognition | France |
| Kieu-Diem Ho | Multivalent Graph Matching and Ant Colony Optimization for Pattern Recognition | France / Vietnam |
| Killian Barrere | Deep Neural Networks and Attention Mechanisms for Handwritten Text Recognition | France |

# List of Mentors

The student participants of the DC were mentored by a group of senior researchers who discussed with them their research plans. 21 mentors have thankfully accepted this role. We have assigned a mentor to each student depending on matching areas of research interests. The mentors who were not paired with students (since we had less students than mentors) have contributed to the DC event by serving in the jury committee to select the best poster presentation and/or by discussing with the students during the event to give them more feedback on their research work.

| Mentor | Affiliation |
|---|---|
| Bertrand Coüasnon | Irisa / Insa, France |
| C V Jawahar | CVIT, IIIT, Hyderabad, India |
| Dimosthenis Karatzas | CVC, Universitat Autonoma de Barcelona, Spain |
| Abdel Belaid | Université de Lorraine – LORIA, France |
| Faisal Shafait | National University of Sciences and Technology, Pakistan |
| Cheng-lin Liu | Institute of Automation of Chinese Academy of Sciences, China |
| Simone Marinai | University of Florence, Italy |
| Rafael Lins | Federal University of Pernambuco, Brazil |
| Thierry Paquet | Laboratoire LITIS, Université de Rouen, France |
| Nicholas Howe | Smith College, USA |
| Josep Llados | CVC, Universitat Autònoma de Barcelona, Spain |
| Seiichi Uchida | Kyushu University, Japan |
| Angelo Marcelli | DIEM - Universita di Salerno, Italy |
| Ioannis Pratikakis | DUTH, Democritus University of Thrace, Greece |
| Harold Mouchère | IRCCyN, Universite de Nantes, France |
| Oriol Ramos Terrades | CVC, Universitat Autonoma de Barcelona, Spain |
| Jean-Yves Ramel | Universite Francois-Rabelais de Tours, France |
| Elisa Barney-Smith | Boise State University, USA |
| Nicole Vincent | Universite de Paris, France |
| Richard Zannibi | Rochester Institute of Technology, USA |
| Shivakumara Palaiahnakote | University of Malaya, Malaysia |

# Short bio of the DC Chairs

**Nibal Nayef** is a senior research and development engineer at MyScript, France working in the field of online handwriting recognition, gesture recognition and layout analysis in online handwritten documents. She has previously worked as a post-doc researcher at the L3i Lab at the computer science department, university of La Rochelle, France. She worked there on different research topics: scene text detection, quality assessment & quality enhancement of mobile captured documents, license plate spotting & recognition, logo & information spotting, and analyzing and predicting user behavior in social networks. She received her Ph.D. degree (2012) in computer science from the Technical University of Kaiserslautern, Germany. She was a member of the IUPR (Image Understanding and Pattern Recognition) research group there, where she worked on geometric-based symbol spotting, feature grouping and textureless object recognition. She is an active participant in creating public datasets and competitions. She was the main contributor of the RRC-MLT 2017 and 2019 datasets and the associated ICDAR competitions, she participated in the organization of SmartDoc competition in ICDAR 2015, and she was the main contributor in creating the dataset SmartDoc-QA presented in ICDAR 2015. She is an active reviewer of many conferences and journals in the field of document image analysis and text recognition.

**Jean-Christophe Burie** is full Professor in computer science at La Rochelle University, France. He is currently deputy director of the L3i Lab, Head of the Joint Laboratory SAIL and vice-president of the University of La Rochelle. He is also Chair of the TC10 on Graphic Recognition of IAPR. He received his Ph.D. degree in Automatic Control Engineering and Industrial Data Processing from University of Lille, France, in 1995. He was a research fellow in the Department of Mechanical Engineering for Computer-Controlled Machinery, Osaka University, Japan from 1995 to 1997 in the framework of the Lavoisier Program of the French Foreign Office. He has been involved in the European Project EUREKA- Prometheus and has actively contributed to the ANR projects: Navidomass and Alpage. His research interests include computer vision, image processing, pattern recognition. His research topics concerns Comics analysis and indexing of, characters recognition written on ancient documents. Since 2011, he is co-leader of the e-bdtheque research program dedicated to the indexing of comics' books. He has actively participated, recently, in the organization of SmartDoc competition for ICDAR 2015, AMADI competition for ICFHR 2016 and 2018 and SSGCI competition for ICPR 2016, RRC-MLT at ICDAR 2017 and 2019. He was Co-chair of the Doctoral Consortium in 2019, General Chair of GREC workshop in 2019 and 2021 and General Co-chair of MANPU Workshop in 2016,2017,2019 and 2021.

# Research plans of the PhD students

# Computerised Image Processing for Handwritten Text Recognition of Historical Manuscripts

Raphaela Heil[0000−0002−5010−9149]

Uppsala University, 75237 Uppsala, Sweden
raphaela.heil@it.uu.se
Main Supervisor: Anders Hast
Co-Supervisors: Ekta Vats, Fredrik Wahlberg (since June 2021),
Anders Brun (until June 2021)
Starting date: 15th January 2018
Expected completion: 30th September 2022

**Abstract** The general focus of the PhD presented in this paper lies on document image analysis and handwritten text recognition for historical documents. It was originally aimed at medieval manuscripts but has recently shifted to also include the processing of 20th century stenographic records. Besides general method development, a portion of the PhD studies has been dedicated to the development of visualisation and transcription tools.

**Keywords:** Handwritten Text Recognition · Document Image Analysis · Strikethrough Removal · Stenography · Semi-Automatic Transcription.

## 1 Introduction

The main focus of my PhD is the development of new approaches in the area of document image analysis (DIA), in particular handwritten text recognition (HTR), for historical manuscripts. Originally, this only entailed documents from around the time of the middle ages. However, in the beginning of 2020, the new project 'Astrid Lindgren Code'[1] was inaugurated, which is centred around the transcription and study of the drafts that famous Swedish children's book author Astrid Lindgren (1907-2002) composed in the Swedish 'Melin' stenography system. The advent of this project has shifted my research focus to the processing of stenography. The Lindgren collection offers a number of interesting challenges, primarily due to the style of writing, stenography, which has not received much attention in recent years. A small summary of the project and some of its challenges can also be found in [3].

[1] Project website: https://www.barnboksinstitutet.se/en/forskning/astrid-lindgren-koden/

**Figure 1.** Top row: samples from the IAM database with synthetic strikethrough applied. Types of strikethrough from left to right: single line, double line, wave, scratch. Bottom row: clean ground truth.

## 2 Work Performed to Date

### 2.1 Strikethrough Removal from Handwritten Words Using CycleGANs

This project[2], which is presented in detail in [4], was primarily inspired by literary scholars in the field of genetic criticism, who are interested in the evolution of a text. Besides comparing different drafts of a text, such studies may also entail the examination of words or sentences that were struck through (cf. Figure 1) and by what they were replaced. In order to facilitate the transcription of struck-through words, this project explores options to remove the strikethrough strokes so that as much as possible of the original, clean word is retained.

Generally, significant effort is required to obtain the actual clean (ground truth) versions of struck-through words in a real corpus, for example by manually deleting the superfluous stroke pixels. In order to avoid this upfront time investment, this work focused on unpaired image-to-image translation, in particular cycle-consistent generative adversarial networks (CycleGAN)[6]. This is the first work to consider deep learning as a means to clean strikethrough and the experiments confirm that CycleGANs can be trained to produce cleaned versions of words for several types of strikethrough.

During the course of the project, and primarily due to the fact that the dataset from the main related work was not fully available, I developed an approach for generating synthetic strikethrough strokes based on information from a given word image (stroke width, intensity distribution, etc.). This method was subsequently used to create a fully synthetic strikethrough dataset, based on the IAM database [5]. Figure 1 shows four samples taken from the validation split of this dataset. For evaluation purposes, this dataset also contains the ground through for each struck-through word and can therefore be used to study paired approaches in the future. Besides the fully synthetic dataset, I also prepared a small, genuine strikethrough collection by scanning word images before and after applying strikethrough and subsequently registering them, using [7], to obtain a closely-aligned ground truth.

[2] Project website, including code and data: https://github.com/RaphaelaHeil/strikethrough-removal-cyclegans

Lastly, as part of this work I also explored the use of neural networks (mainly DenseNet121) for differentiating between clean and struck words, as well as for the classification of struck words into one of seven categories (e.g. single horizontal line, wavy stroke, etc.). The experiments confirm that neural networks can be used to address these tasks.

### 2.2 Semi-Automatic Annotation Tool

This project entails the development of a semi-automatic, interactive annotation tool. It is being implemented as a response to requests from the local community and is based on the idea of clustering word images in a feature space, as was presented in [2]. The tool allows users to explore the space of feature representations by applying a dimensionality reduction technique, mapping each datapoint into a 2D space. Users can interact with datapoints in this 2D representation, by, for example, inspecting whole clusters or individual points and annotating them one by one or in batches. Besides the aforementioned word annotation, other use cases, such as dating or writer identification, are conceivable and supported by the tool due to a flexible annotation interface. In addition to this, the tool has found applications in the visualisation of oral cancer cell images.

### 2.3 Alignment of Segmented Word Images and Transcriptions for the Astrid Lindgren Stenographic Manuscripts

This currently ongoing project focuses on the task of preparing annotated word images for the Astrid Lindgren manuscripts, by facilitating the transcriptions that are being prepared by volunteer stenographers via crowdsourcing. One of the main challenges in aligning these two sets of information is the presence of deletions, i.e. strikethrough, and additions or substitutions, typically in the form of new words written above a deletion. Due to this, the alignment not only has to be performed in the usual reading direction but also 'vertically', in the areas where several iterations of the text occupy the same space in a sentence. At the time of writing, the proof of concept is still under development. It will however entail the use of the clean vs struck classifier, mentioned in section 2.1, to aid in the alignment of struck-through words, which the transcribers are kindly tagging accordingly.

## 3 Future Work

### 3.1 Strikethrough Removal Using Paired Data

As mentioned above, CycleGANs were originally chosen for the task of strikethrough removal due to the lack of paired data. The development of an algorithm for generating strikethrough in the same project now enables me to explore deep learning approaches that are based on paired data, such as denoising autoencoders.

### 3.2 Comparison of Feature Extractors for Clustering-Based Visualisations of Text Component Images

A small amount of feature extractors have been explored as basis for meaningful clusters of word images, e.g. in [2]. In order to ensure that the annotation tool is of use to a wide range of users and use cases, this project aims for a rigorous examination of different feature extractors, both classical, such as histograms of oriented gradients [1], and modern, i.e. deep-learning-based, such as autoencoders, for a range of tasks that are of interest to humanities scholars, such as transcription, dating and writer identification.

### 3.3 Handwritten Text Recognition for the Astrid Lindgren Stenographic Manuscripts

Employing the results from section 2.3, the main focus of this project is to develop an HTR approach for the recognition and transcription of the Astrid Lindgren stenographic manuscripts.

## References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. IEEE (2005)
2. Hast, A., Mårtensson, L., Vats, E., Heil, R.: Creating an Atlas over Handwritten Script Signs. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019. CEUR Workshop Proceedings, vol. 2364, pp. 175–180. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2364/16_paper.pdf
3. Heil, R., Nauwerck, M., Hast, A.: Shorthand Secrets: Deciphering Astrid Lindgren's Stenographed Drafts with HTR Methods. In: Dosso, D., Ferilli, S., Manghi, P., Poggi, A., Serra, G., Silvello, G. (eds.) Proceedings of the 17th Italian Research Conference on Digital Libraries, Padua, Italy (virtual event due to the Covid-19 pandemic), February 18-19, 2021. CEUR Workshop Proceedings, vol. 2816, pp. 169–177. CEUR-WS.org (2021), http://ceur-ws.org/Vol-2816/short5.pdf
4. Heil, R., Vats, E., Hast, A.: Strikethrough Removal from Handwritten Words Using CycleGANs. In: 2021 International Conference on Document Analysis and Recognition (ICDAR) (2021)
5. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**(1), 39–46 (2002)
6. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (10 2017)
7. Öfverstedt, J., Lindblad, J., Sladoje, N.: Fast and robust symmetric image registration based on distances combining intensity and spatial information. IEEE Transactions on Image Processing **28**(7), 3584–3597 (2019). https://doi.org/10.1109/TIP.2019.2899947

# Text based Visual Question Answering

Rubèn Tito and Ernest Valveny

Computer Vision Center, UAB, Spain
{rperez, ernest}@cvc.uab.es

**Abstract.** Current Visual Question Answering (VQA) task has received wide attention in terms of both datasets and methods. However, they mostly focus on the visual components in the scene, while they tend to ignore the text in the images that carries essential information for scene understanding and reasoning. On the other hand, the research field of Document Analysis and Understanding (DAR) has focused on information extraction tasks of documents that are initially addressed to human comprehension blind to the purpose that can be used for. Hence, in this thesis we explore the role of text in Visual Question Answering for both Scene Text and Documents.

**Keywords:** Document Collection · Document VQA.

## 1 Introduction

Textual content in human environment conveys important high-level semantic information that is explicit and not available in any other form in the scene. Interpreting this information is essential in order to perform most everyday tasks like making a purchase, using public transportation, finding a place in the city, getting an appointment, or checking whether a store is open or not, to mention just a few. The research community on reading systems has made significant advances over the past decade [12, 8] and the current state of the art in scene text understanding allows endowing computer vision systems with basic reading capacity. However, at the moment of the beginning of this thesis, the community did not exploit this towards solving higher level problems. The task of Visual Question Answering (VQA) [6] consist in taking as input an image and a natural language question about that image to produce a natural language answer. To this end, during the last years several datasets and methods were published [6, 9, 11]. However, they mostly focus on the visual components while ignoring textual cues that can be found in real world scenarios.

On the other hand, the research field of Document Analysis and Understanding (DAR) has focused on information extraction tasks on documents that are initially addressed to human comprehension. Some of the most widely known applications involve processing office document such as text recognition, table and forms detection and recognitionor mathematical expressions. However, such tasks are designed blind to the end-purpose the extracted information will be used for. Moreover, even though documents are usually organized in collections

(historical records, purchase invoices), the research line on those have been limited to retrieval, ignoring the semantic information that provide context useful for their interpretation.

### 1.1 Work done

**Course 2018-2019:** In this context during the first year we proposed a new task and dataset named Scene-Text Visual Question Answering (ST-VQA) which consisted in VQA on natural scenes where methods were required to read in order to answer the posed questions. Along with this task we proposed a new metric named Average Normalized Levenshtein Similarity (ANLS) that in contrast to accuracy, was able to smoothly capture OCR recognition errors as well as assess the methods' reasoning capabilities. Finally, we also proposed some baselines based on heuristics (provide as answer the biggest word recognized in the image) and based on existing VQA methods like Stacked Attention Networks (SAN) and Show, Ask, Attend and Answer (SAAA) by concatenating the text features to the image features. The work was published in [2] and a competition report celebrated on this task and dataset was also published in [1].

**Course 2019-2020:** Following the idea to open new research lines on Visual Question Answering that requires methods to read, and referring to the lack of a high-level purpose of DAR mentioned in previous section, we decided to propose a new series of challenges with their corresponding datasets on documents named DocVQA, which included two different tasks and datasets. The first one Single Document VQA followed a similar setup than ST-VQA with the only difference that the images were documents instead of natural scenes. These document images were sourced from the UCSF Industry Documents Library and contained complex layout elements including tables, figures and forms. The metric we chose for this task was also the ANLS. Again, we also proposed two baselines. The first one was based on the scene-text state-of-the-art method at that moment M4C [10] and the other from the Natural Language Processing (NLP) research line, an extractive BERT [7] QA method.

The second task and dataset proposed in the scope of the DocVQA series was the Document Collection VQA (DocCVQA). In this case, the questions were posed over a whole collection of 14K documents and the methods were required not only to provide the correct answer to the given question, but also the IDs of the documents that contained the information necessary to answer such question. Thus, this task could be perceived as a retrieval-answering task given that the answers are usually a set of items sourced from different documents. The documents in this task were the US Candidate Registration forms, sourced from the Open Data portal of the Public Disclosure Commission (PDC). To evaluate the methods, we set the Mean Average Precision (MAP) to assess how good were the methods at the time of retrieving the correct document IDs. On the other hand, we set precision and recall to evaluate the answering performance.

A short report of the DocVQA Challenge series was published in [4] and the work on the Single Document VQA was published in [13].

**Course 2020-2021:** As a continuation on the DocCVQA task, we also proposed two baselines in this task. Both methods work in two stages, first they rank the document according to a confidence that indicates whether the document is relevant to answer the question and then, extract the answer from the documents marked as relevant. The first baseline is based on simple and generic existing methods. First, it ranks the documents by performing text spotting of the words in the question over the documents in the collection and then, the answer is extracted from the top ranked documents through the extractive QA BERT method. The second baseline exploits the fact that all images in this task shares the same template. So it makes use of the Amazon Textract commercial OCR that is capable to extract key-value pair relations which are afterwards dumped into a database-like data structure. Then, the questions are parsed into structured query language to retrieve the relevant documents as well as the answer. This is an ad-hoc method which can't generalize to other datasets but showcases the performance of this kind of commercial solutions on this task. Along with those baselines, we create a new metric that better evaluates the answering performance of the methods in this task. The new metric is based on the ANLS adapted to be used on a set of answers for which the order is not relevant. We named this adaptation as Average Normalized Levenshtein Similarity for Lists (ANLSL). This work is already accepted and will be presented in the ICDAR 2021 conference. The work can be found in [5].

On the other hand, after noticing the good performance on the Single Document VQA task of NLP methods that ignored visual features, we decided to extend the DocVQA series of challenges with a new task and dataset where visual features are much more relevant. The new task InfographicsVQA is set up similar to the Single Document VQA but where the questions are posed over Infographic images instead of business documents. As an evaluation metric, we use the ANLS and also propose two baselines. The first one is the same as in the Single Document VQA which is based on the method M4C [10]. The second baseline is based on LayoutLM [14], an extension of BERT that takes into account layout information. The work is currently under review, but can be found in [3].

### 1.2 Future work

I'm currently working to propose a model on DoCVQA that is capable to reason the question's answer as well as retrieve the document from which the answer has been inferred from in a single step. However, the current document collection is quite limited given the fact that all documents share the same template and therefore, in a mid-term we would like to increase the complexity of such collection by including heterogeneous documents as well as more challenging questions.

## Acknowledgment

The Ph.D. program is supported by the UAB PIF scholarship B18P0070.

## Publications

1. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D.: Icdar 2019 competition on scene text visual question answering. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1563–1570. IEEE (2019)
2. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4291–4301 (2019)
3. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. arXiv preprint arXiv:2104.12756 (2021)
4. Mathew, M., Tito, R., Karatzas, D., Manmatha, R., Jawahar, C.: Document visual question answering challenge 2020. arXiv e-prints pp. arXiv–2008 (2020)
5. Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. arXiv preprint arXiv:2104.14336 (2021)

## References

6. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. ACL (2019)
8. Gomez, R., Shi, B., Gomez, L., Neumann, L., Veit, A., Matas, J., Belongie, S., Karatzas, D.: Icdar2017 robust reading challenge on coco-text. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1435–1443. IEEE (2017)
9. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
10. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9992–10002 (2020)
11. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017)
12. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160. IEEE (2015)
13. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF WACV. pp. 2200–2209 (2021)
14. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. ACM SIGKDD (Jul 2020). https://doi.org/10.1145/3394486.3403172, http://dx.doi.org/10.1145/3394486.3403172

# Information Theoretical Approach To Understand Deep Neural Networks

Christoph Zaugg

University of Fribourg, CH-1700 Fribourg, Switzerland
DIVA, Boulevard de Pérolles 90
Supervisor: Rolf Ingold, Co-Supervisor: Andreas Fischer
Starting date: Sept 2026, Expected end date: Sept 2026
christoph.zaugg@unifr.ch

**Abstract.** In my Ph.D. thesis, I address the question, why a given Few-Shot-Learning technique reduces sample complexity. My findings will contribute to applying supervised deep learning methods in cases where labeled samples are rare. I will mainly work with synthetic data but transfer my results to classification tasks in Historical Document Analysis as a test bench. In the first section, I will outline the context of my intended research. The second section introduces two concepts from literature. A framework from information theory will serve as a basis for interpreting neural networks' learning behavior as dynamics in the information plane. A paradigm for estimating the mutual information for various encoders and decoders during the training of a DNN enables the observation of the focused dynamics. I also identify possible points of conflict and suggest how to alleviate them. In section three, I outline my research design and present how to deal with potential obstacles. Section four describes the implications and contributions of my findings to knowledge. The research schedule in section five divides my research into phases and states the objectives to achieve for given deadlines.

**Keywords:** Deep Neural Network, Classification Task, Few-Shot-Learning, Historical Document Analysis, Information Bottleneck, Mutual Information Neural Network

## 1 Introduction

In contrast to machines, human beings are good at generalizing from just a few examples because they can incorporate prior knowledge from their experience. Few-Shot Learning (FSL) addresses this gap and may serve as a hallmark of implementing intelligence to Machine Learning. In surveying FSL techniques, the authors in [1] start from the decomposition of the total error into a sum of the approximation and the estimation error. They gain a unified terminology for the state-of-the-art for FSL techniques depending on whether a specific approach incorporates prior knowledge by augmenting the data, restricting the model's hypothesis space, or altering the optimization algorithm. They convincingly argue that FSL strategies reduce sample complexity and state

that a detailed analysis of how this reduction comes about would be helpful for targeted use in the future. I want to reduce this gap of understanding applying quantitative methods from information theory. My research question aims to explain why a particular FSL technique reduces sample complexity.
Nowadays, there is plenty of data available to train deep learning models. The same goes for computing power to cope with the algorithms used and the flexible software to implement them. What is often missing is a sufficient amount of data with ground truth, as is typical for problems in Historical Document Analysis. The DIVA Research Group [2] hosts HisDB [3], a library of digitized historical documents.

## 2 Literature Review

The Mutual Information (MI) between two random variables is the amount of information obtained about one variable through observing the other. The Information Bottleneck (IB) principle characterizes a maximally compressed representation. It minimizes the MI between observation and its representation. At the same time, it is informative by keeping the MI between the representation and the target above a threshold. This principle transforms a Deep Neural Network (DNN) learning behavior into the dynamics of points corresponding to individual layers in the information plane. The horizontal and vertical axes of this plane represent the MI in the encoder, respectively decoder. For discrete random variables, the authors in [4] design a fully connected neural network with an Eiffel Tower-like architecture and carefully adjust synthetic data for estimating the various MI values. The dynamics of the points in the information plane corresponding to individual layers show two phases: A long compression with a small gradient of high variance follows a short fitting with a large gradient of low variance. The points corresponding to individual layers finally settle down at the IB curve, solving the IB principle's restricted optimization problem.
Starting from the Donsker-Varadhan theorem, the authors present in [5] the Mutual Information Neural Estimator (MINE) that allows estimating MI values for vector-valued continuous random variables. This theorem represents the MI as the supremum of the difference of two expectations w.r.t. the joint and product distributions. Taking this supremum only over a family of functions indexed by the parameters of a DNN yields a lower bound of the MI, i.e., the neural information measure associated with the DNN, acting as a statistical network. Replacing both expectations by empirical distributions associated to i.i.d. samples, the authors calculate the neural information measure by Stochastic Gradient Descent (SGD). They show that the MINE is a strongly consistent estimator of the MI, and prove a lower bound for the minimal number of samples to estimate the MI accurately with high confidence. Regularizing the MINE for VGG-16 CNN, the authors in [6] confirm that the MI converges in the information plane converges.
Even if the existence of a compression phase in the case of vector-valued continuous random variables is still an open problem [7], the IB principle allows to represent the learning behavior dynamics in the information plane. To accelerate convergence, the authors in [6] modify the MINE by regularizing it. In [8], the authors directly estimate

the gradient of the statistical network to obtain a low-variance estimator of the MI. During the time of my dissertation, I expect other authors to make further improvements in estimating the MI for vector-valued continuous random variables.
The authors of [1] explicitly state a lack of understanding why FSL techniques reduce the sample complexity. With the IB principle as a basis, and the MINE as a tool I will tackle this question to gain innovative insights.

## 3 Research Designs and Methods

To represent the learning dynamics of fully connected and convolutional DNNs in classification tasks, I will implement various modifications of the MINE. Since I expect a significant amount of computing effort, my code must be compatible with computing accelerating devices or distributed computing on cloud services. I will start with the jointly Gaussian case, where an analytical solution of the IB curve exists, and validate my implementations on this case. I will represent the learning dynamics in classification tasks when using competitive FSL techniques. During this stage, I will mainly work with synthetic data. A library of my documented examples serves me in a meta-study to identify reasons for reducing the sample complexity. In the last stage, I will transfer my findings to optimize specific classification tasks in Historical Document Analysis.
If I am designing my code for high-performance computing and working with synthetic data, the existing hardware and software should be sufficient to build my desired library. Obstacles might occur if the MINE exhibits bad convergence behavior, possibly due to numeric issues. In this case, I have to modify its implementation, either by relying to improvements of other authors or by adjusting it with novel ideas. To gain insights during the meta-study, the library comprising the documented comparisons must have sufficient diversity. Any step away from synthetic to real data is an inherent risk. Since HisDB's documents are carefully selected and well documented, this risk is also limited.

## 4 Implications and Contribution to Knowledge

The expected findings are twofold. On the one hand, there are implementations for layer-wise tracking the MI. On the other hand, there are explanations why a particular FSL technique reduces sample complexity.
As a practical implication of the expected findings a future diagnostic tool for the learning behavior of DNNs might incorporate one of my implementations with a reliable convergence behavior as a routine. If my explanations prove to be sufficiently comprehensive, then such a tool could also determine the optimal FSL technique out of several competitors.
As for progress in theory, my results could help to clarify under what conditions DNNs with layers given by vector-valued continuous random variables admit a compression phase during training.

## Research Schedule

Since I work 70% as a lecturer in mathematics and as a researcher at the ZHAW, I am doing my doctorate part-time and therefore plan a time horizon of five years.
During my probational year (Sept 2020-Sept 2021) at the University of Fribourg, I familiarized myself with PyTorch and the basics of information theory, cf. [9].

| Research Phase | Objectives | Deadline |
|---|---|---|
| Implementation | 1. Implement and validate MINE as a tool<br>2. Build library of documented examples<br>3. Publish results in workshop paper | Sept 2023 |
| Meta-Study | 1. Identify reasons for reduction of sample complexity for various FSL techniques<br>2. Publish findings in conference paper | Sept 2025 |
| Transfer | 1. Transfer results to optimize classification tasks using competing FSL techniques in Historical Document Analysis<br>2. Publish finings in a conference paper<br>3. Write Ph.D. thesis | Sept 2026 |

### References


1. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: A Survey of Few-Shot Learning, arXiv:1904.05046v3 [cs.LG] 29 May 2020.
2. DIVA Research Group, https://www.unifr.ch/inf/diva/en/.
3. DIVA-HisDB, https://diuf.unifr.ch/main/hisdoc/diva-hisdb.
4. Tishby N., Schwartz-Ziv R.: Opening the black box of Deep Neural Networks via Information, arXiv:1703.00810v3 [cs.LG], 29 Apr 2017.
5. Belghazi, M.I., Baratin A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A. Hjelm, R.D.: MINE: Mutual Information Neural Estimation. arXiv 2018, arXiv:1801.04062.
6. Jonsson, H., Cherubini, G., Eleftheriou, E.: Convergence Behavior of DNNs with Mutual-Information-Based Regularization, Entropy 2020, 22, 727
7. Geiger, B.C.: On Information Plane Analyses of Neural Network Classifiers—A Review. arXiv 2020, arXiv:2003.09671.
8. Wen, L., Zhou Y., He L., Zhou M., Z. Xu Z.: Mutual Gradient Estimator for Representation Learning, arXiv:2005.01123v1 [stat.ML], 3 May 2020, Book title. 2nd edn. Publisher, Location (1999).
9. Cover, Th. M., Thomas, J.A.: Elements of Information Theory. 2nd edn. Wiley (2006), New Jersey

# Automatic and model-free learning of semantic-structural links of fields in a document

Ibrahim Souleiman*, Mickaël Coustaty+
Aurélie Joseph*,Vincent Poulain d'Andecy*, Jean-Marc Ogier+

* Yooz
1 Rue Fleming, 17000 La Rochelle, France
Email: {ibrahim.souleimanmahamoud,aurelie.joseph,vincent.poulaindandecy}@getyooz.com

+ La Rochelle Université, L3i
Avenue Michel Crépeau, 17042 La Rochelle, France
Email: {mickael.coustaty,jean-marc.ogier}@univ-lr.fr

***Abstract*—This project is a continuation of the work initiated in the last three years between the company YOOZ and the L3i and in the joint laboratory ANR IDEAS. The general objective is to propose a generic system of information extraction from document streams in order to automatically process a heterogeneous set of documents from companies. The specific objective of this project is to use Deep Learning, expert system and incremental learning, to automatically detect the concepts of a document to make a hybrid version that can be enriched over time. Concepts can be seen in a broad sense as general semantic descriptors or specific and precise descriptors as information fields. This thesis therefore aims at applying a methodology and a theoretical model allowing the analysis of documents of any kind in order to extract precise information necessary for all levels of document processing.**

## I. Introduction

In the age of digitalization, the automatic processing of administrative documents has become both an economic and technological challenge. Whether it is small, medium or large, a company wants to spend as little time as possible processing its mail. This can be done by determining the class of a document in order to send it to the right department, for example, or simply extracting the amount of an invoice and knowing who to pay it to. However, the mass of data is such that manual processing is no longer possible because it is too costly and slow. Automatic Document Reading (ADR) is a software solution that automatically reads these scanned documents and extracts the useful information to inform the information systems and process them quickly. For the past 30 years, ITESOFT and now YOOZ have been developing information extraction and classification tools that seek to reproduce the capabilities of humans while being more efficient. However, while the documents of choice for a company were invoices, forms,financial regulations, and mail, the diversity has extended to other types of documents that are much more varied but also more ambiguous (e.g. purchase orders, estimates, deliveries, etc.). Even within the documents of choice, variability is increased by the increase and internationalization of B2B exchanges. Document analysis approaches by templates or keywords, adapted to small volumes, are no longer sufficient. Manually generated resources to identify a particular element become tedious to perform. It is then necessary to be able to generate these resources automatically or to find an alternative way to locate the information. These techniques must also be generic and multilingual with a minimum of parameters to facilitate their use. In this context, automatic classification and data extraction remain a challenge. In terms of new technologies, Deep Learning or incremental learning systems have become essential but do not offer reliable industrial solutions to the problems of real datasets (strong imbalance between learning classes, lack of classes and models during learning, etc.).

## II. PROBLEM DEFINITION

The main objective of our proposal is to provide two solutions for classification and extraction . This two solution are dependent, the extraction depends on the classifier according to the class it will try to extract or not information. Our classification problem is defined as follows . The task of prediction of document classes in an industrial environment involves input flows $s_1, s_2, s_3, ..., s_n \in S$ , where sn is the number of samples from classe n in the training set. The number of samples per class is by definition imbalanced due to the industrial context. Some document classes are very recurrent (i.e there will be thousands of invoices) while others are very rare with a very few number of samples (ten or even 1 document) as displayed in figure 1. The objective of this predictive task is to predict the class of samples S. The overall assumption is that an unknown function correlates the samples and ground truth classes of $S, i.e. Cn = f(sn)$. The goal of the learning process is to provide an approximation of this unknown function whatever the quantity of samples available by class. To better approximate this function f, we must know the are significant intraclass variation and inter-class similarity caused by different structure documents for each client brings a great deal of difficulties to classify .

We will also define the problem of information extraction. The extraction task is strongly correlated to the classification,

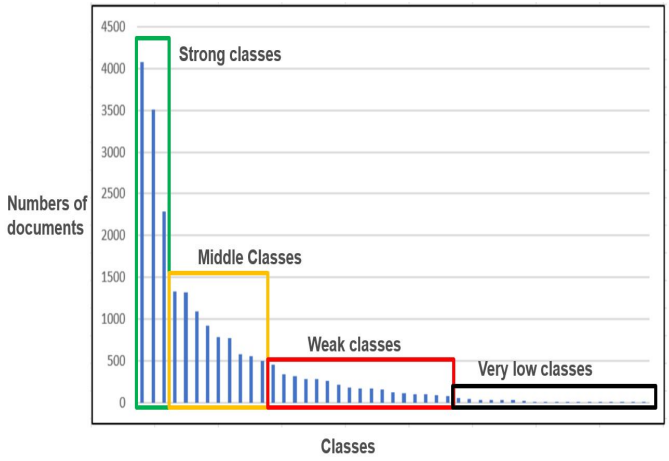



Fig. 1. Class distribution of the Yooz dataset

if our classifier makes a bad prediction and predicts us a bill then the extraction model will try to extract some information (e.g. a number in this document can be considered as a total amount). There are also cases where an invoice is predicted as other classes (i.e. GENERAL TERMS OF SALE) and the expected fields are not extracted. Hence the importance of having as much as possible a strong classifier able to correctly predict all the invoice documents and minimize the False-Positives on the invoice . We try to extract for each label (e.g. Due Date, Invoice Date, Total Amount, Invoice Number, Address, etc..) the information that corresponds to it in the input document of the model. These input documents are not all the same structure in an invoice we can for example find the Invoice Number at the top while in another document it will be in the middle. This variance is due to the template of documents because each supplier has its structure to position these words according to its needs which implies that these documents have a certain structure but are not all similar. All our labels are not explicitly present in the documents some are implicit and requires mathematical operation on certain value existing in the documents, to well find the implicit number it will be necessary to make sure that the number on which we will make the calculation are well correct. The objective of the extraction is to be able for each entity (word or n-gram) to provide the class to which it belongs and for the entities that we would not want to extract they will belong to the undefined class. You have certainly guessed that in a document we will have more words classified as undifined, it will therefore be necessary to explicitly give our model more weight to the correct classification of other tags (i.e. DUE Date, Invoice number, etc. ..) in order to provide a solution to this problem of imbalance.

## III. OBJECTIF AND CHALLENGES

The objective of this thesis can be broken down into 4 axes:

- Learn and reconstruct automatically the semanticstructural link of fields in a document. This should allow the development of a general model applicable to documents with similar semantics
- Extract data without a priori manual model (business rules entered by an expert for each client) on a large and multilingual vocabulary. The models will be learned from examples of results
- Propose methods with little or no parameterization
- Satisfy a low processing time and respect the industrial constraints on error minimization

The two main technological challenges are :

- Simplicity and automation of learning that can be achieved from naive examples, within reach of an end user.
- Continuous adaptation of classifiers as data is discovered.

## IV. RESEARCH

In this section we will describe the models we want to implement to bring solutions to the problem defined before. Document classification and information extraction for a human requires some expertise, often related to the business or domain. Document sorting often relies on the use of taxonomies (document classes) and on similarities with previously encountered documents, in order to decide which class to assign. From a machine learning point of view, this relies on the recognition of certain patterns in the documents, the position of these patterns and the presence of certain relevant words (keywords) that allow to match and classify a set of documents or a set of words. Many solutions have been proposed in the last twenty years. In figure 2 we have a global description of our End-To-End model, in a first time the classification model will have input the image and the textual content of source document. The extration model will have input the prediction of the class, the image and the text and it will extract the necessary information. The classification and extraction model may or may not use the same architecture for example they may use a combination of expert system, incremental learning and deep learning.

We will describe each model to have a global vision of how it works. If we start with the expert systems[1], they have been proposed, they tried to mimic human knowledge to reach decisions. These systems are composed of a base of facts (the documents and their classes or documents and format of the data to be extracted), a base of rules (for example, the presence of the word invoice in the title of the document can make it possible to consider it as belonging to the class invoice) and an engine of inference to simulate the mechanism of decision-making of the human being. The use of expert systems to classify documents faces several obstacles. For example, these systems impose the use of rule bases, which are not easily exhaustive and evolutive because they require the identification of all the rules for classifying documents. This becomes moreover very difficult in a context of very large volume of documents where the intra-class variability explodes because of the diversity of examples of

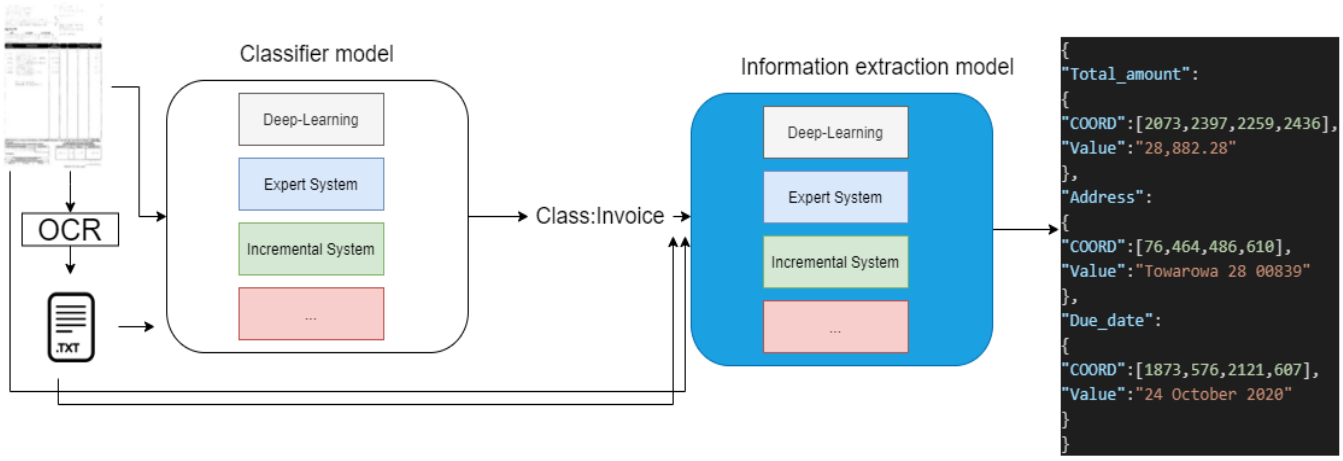



Fig. 2. End-To-End model for classfication and information extraction

the same class. This means generating a large number of rules that are difficult for a human to count. Its use in information extraction also encounters obstacles, for example if an invoice number was recognizable when it was in the header of the document and to the right of the words "Invoice Number" these rules can change for new document formats.

To avoid listing all the possible rules, deep-learning methods have been proven to automatically learn these rules. These deep-learning models have allowed us to obtain good results on well defined problems. Moreover, once the modeling is done, the inference is generally fast, which is an important notion in a high volume industrial context. Nevertheless, these techniques require a huge amount of annotated data for the learning phase and impose to know all the classes in advance. In a more detailed way, several deep architectures have been proposed in the state of the art, architectures which are interested either in the graphical part or in the textual part of the documents. The use of deep convolutional networks has been explored to extract features from an image (InceptionResNet [2], NasNet [3] et VGG [4] to name the most used). In contrast, several works have focused on the textual part of the documents. These approaches, which are generally based on recurrent neural networks, aim at learning the organization and the semantics of the words of a document. For this purpose, word embeddings are used and the most frequent are[5], GloVe [6] et Fast Text [7]. The most recent works propose to combine textual and visual information to take into account the whole content of the document. Thus, [8] uses a multimodal neural network able to learn from a lexical folding of the text extracted by character recognition [9] and visual features of the image (MobileNetV2 [10]).[11] introduces an architecture with a NasNet network [3] for the image part and a bidirectional BERT architecture for the text. There alos models that exploit multimodality to build a new representation of 2D documents[12] by replacing pixels with textual features. This allows to exploit the position of words in the document and create embedding vectors for a better understanding of the word or letter. Finally, the last major constraint is based on the high intra-class variability and the low inter-class variability, a characteristic of administrative documents. For example, we can cite a quote and an invoice that are very close visually and textually (even though they belong to two different classes), whereas invoices from different companies may be very far apart visually and textually.

Despite the good performances of these deep learning techniques, they have their limits in an industrial context. The company which is confronted with huge document flows can be treated with other types of documents (e.g. new classes of documents, new structure of documents) Incremental learning brings a solution to this problem. Incremental learning is a field of research that relies on learning algorithms able to learn from data received over time ([13],[14] ,[15],[16]. The objective is to reorganize the model at each increment in order to predict the result from the input data and the expected result. Incremental learning is expanding rapidly with the massive development of very large databases, often obtained via real-time data streams that require continuous learning and adaptation of the model.

This project is at the interface between incremental learning in order to adapt the system to the evolution of the corpus, deep learning in order to obtain a most relevant and generic characterization for large volumes of documents , and finally the expert systems which make it possible to propose rules specific to the end user's business. To our knowledge, the joint study of these three typologies of approaches, in particular in the context of processing large flows of documents, has never been carried out. This thesis project therefore aims to establish a complete state of the art on these themes, define the approaches relevant to our problem and study their combination. Finally, to evaluate and validate all of this work, we will rely on a set of data sets and the classic metrics from the

literature (Recall, Precision, F-Measure). Several databases are freely available in order to be able to pretrain the models and compare them. For the best known and largest classification dataset, which currently serves as a worldwide reference, is the RVL-CDIP . This annotated base of documents in grayscale is made up of 16 classes, each comprising 25,000 images. The base is subdivide in three sub-parts (320,000 documents for training, 40,000 for validation and for the test). But for the information extraction we do not have a dataset of documents with fields to extract complete, the little that exists in the public data consists of hundreds of documents (e.g 558 documents), so we decided to annotate data from the RVL-CDIP Invoice class in order to extract certain fields (eg Address, Invoice dates, etc.)

## V. STAKEHOLDERS IN THE PROJECT

Title of the thesis : Automatic and model-free learning of semantic-structural links of fields in a document

Student in thesis: Ibrahim Souleiman Mahamoud (70% of time in Laboratory L3i - 30% in YOOZ Company)

Director: Jean-Marc Ogier, L3i, University of La Rochelle

Scientific supervisor: Mickaël Coustaty, L3i, University of La Rochelle

Industrial supervisor : Aurélie Joseph, Yooz, Aimargues

Starting Date : 15/04/2021

Finalization date : 14/04/2024

## VI. CV

### A. EDUCATION

- PhD Extraction information from documents(04/2021-04/2024) University of la Rochelle and YOOZ company
- Post-graduate Artificial intelligence (09/2019-09/2020) École de centrale Marseille
- Master's degree Artificial intelligence and machine learning (09/2018-09/2019) University Aix-Marseille
- Bachelor's degree Database and Software (04/2015-04/2016) University of Djibouti
- High-School Diploma (09/2012-04/2013) High-School Baballa in Djibouti City

### B. PUBLICATIONS

- ICDAR-2021 — Multimodal Attention-based Learning for Imbalanced Corporate Documents Classification
- EGC-2021 (09/2019-09/2020) — Apprentissage multimodal bas´e sur des mod'eles d'attention pour la classification de documents dans un contexte d´es´equilibr´e

### C. Work Experience

- Fixed-term contracts (09/2020-04/2021) Company YOOZ — Documents Classification
- Internship Company YOOZ — Documents Classification
- Internship in Laboratary TALEP (Aix-Marseille University) (03/2019-07/2019) —Multimodal comprehension on the GuessWhat?!
- Internship Laborotary Qarma (Aix-Marseille University) (06/2018-07/2018) — Artificial Intelligence for shufimi game

## REFERENCES


[1] E. T. Ogidan, K. Dimililer, and Y. K. Ever, "Machine learning for expert systems in data analysis," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1–5.

[2] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[3] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[6] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[7] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[8] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[9] N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal deep networks for text and image-based document classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 427–443.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[11] S. Bakkali, Z. Ming, M. Coustaty, and M. Rusinol, "Visual and textual deep feature fusion for document image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 562–563.

[12] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," 2018.

[13] Z. Chen, L. Huang, and Y. L. Murphey, "Incremental learning for text document classification," in *2007 International Joint Conference on Neural Networks*, 2007, pp. 2592–2597.

[14] M.-R. Bouguelia, "Classification and active learning from dynamic dataflows with incertain labels," 2015.

[15] M. Bouguelia, Y. Belaïd, and A. Belaïd, "A stream-based semi-supervised active learning approach for document classification," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 611–615.

[16] A. B. H. Hamza, Y. Belaïd and B. B. Chaudhuri, "Incremental classification of invoice documents," 2008.

# Towards an Explainable Deep Model for Archival Document Image Segmentation

Student's name: Iheb Brini
Supervisor/s of the thesis: Najoua Essoukri Ben Amara, Rolf Ingold and Maroua Mehri
University: University of Sousse
Starting date of the Ph.D:. December 01, 2020
Expected finalization date of the Ph.D.: November 30, 2023
Email: ihebbrini.ing@gmail.com

***Abstract* — Over the past 30 years, working on archival documents has raised many open questions and increased the challenges due to the idiosyncrasies of this kind of digitized documents. Recently, numerous solutions based on deep models have been proposed to bypass these issues. However, these solutions are based on complex and non-interpretable architectures. Indeed, the increased complexity of the addressed problems often results in the creation of the so-called "black-boxes" machine learning models. This means that these models can result in some unexpected behavior in which the user is not able to debug or understand their predictions. Hence, to overcome this limitation, we are working to investigate the concepts of explainable AI (XAI) and apply them in the field of archival document image analysis (ADIA). The main objective of this Ph.D. work is to propose efficient and robust models which are at the same time transparent and trustworthy for ADIA. In particular, we focus on developing a deep framework that is at the same time efficient and explainable for segmenting historical document images.**

***Keywords — Deep learning; Explainable AI (XAI), Image segmentation; Archival document images.***

## A. Introduction

Since the last decade of the twentieth century, researchers and archivists have become increasingly interested in many challenges and open questions related to the sustainable preservation and worldwide open access to documentary heritage on the one hand, and to the exploitation and valorization of archival materials on the other hand. As a result, there is a growing need for the development of reliable Web-based access systems to the written heritage and for rapid and automatic tools for the analysis and characterization of archival materials. More specifically, a high priority was given to the development of fast, efficient and robust solutions for archival document image analysis (AIDA), and more specifically to the analysis of their structure or layout. Indeed, AIDA remains an open question due to the particularities of archival documents, such as the superposition of several layers of information (e.g., stamps, handwritten notes and noise) and the increased variability and complexity of the content and/or layout (cf. Figure 1). In addition, analyzing archival document images without prior knowledge of the document layout and content is more complex and tedious.

**Figure 1: Examples of archival document images collected by the national archives of Tunisia**

Recently, with the rise of deep architectures, research works have proposed to use deep learning (DL) representations to segment images. These DL architectures have achieved highly satisfactory performances for segmenting and recognizing objects in images. Nevertheless, researchers working in the different fields of AIDA have raised several issues related to the explicability of DL model predictions. Although some DL-based systems have already been deployed, the inherent and undeniable risk remains the abandonment of human control and monitoring in favor of these DL models.

Before deploying DL systems, it is worth noting that it is necessary to first validate their behavior, and thus establish guarantees that they will continue to perform as expected in a real environment. To contribute to this goal, tools that allow humans to verify the agreement between DL model predictions and ground truth have been explored. Thus, instead of developing and using DL models as black boxes and adapting their architecture to a variety of applications, the goal of Explainable AI (XAI) is to propose methods to "understand" and "explain" or "interpret" how these systems generate their decisions.

XAI is a subfield of artificial intelligence (AI), created to expose complex DL models to humans in a systematic and interpretable way [1]. A number of XAI-based solutions have been proposed, such as layer-wise relevance propagation (LRP) [2], local interpretable model-agnostic explanations (LIME) [3] and generalized additive model (GAM)[4]. Some of them have already shown to be useful to detect some defects in the architectures of the evaluated DL models. These solutions focus on verifying whether the decision-making behavior is based on relevant features or rather on spurious correlations or artifacts.

As shown in Figure 2, a model is fed an input image and it generates a set of predictions. In the case of semantic segmentation tasks, these predictions are in the form of output masks with a unique color to each specific class.

To explain these predictions, we use an XAI method (e.g., GradCAM) that takes into account the input image, the model parameter and the output prediction in case of an attribution-based method.

The explainer bloc is now able to generate an "explanation map" in a form of heatmap that highlights the most relevant pixels contributing to the predictions given a class of interest.

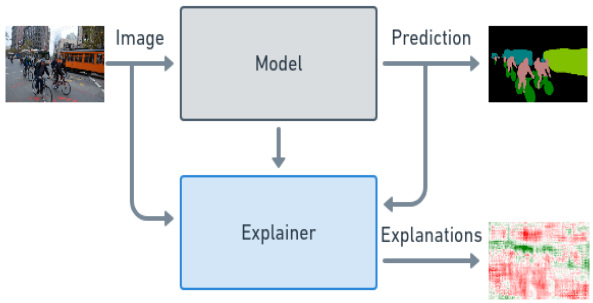



**Figure 2: Illustrative schema of using an XAI method for a semantic segmentation task**

## B. *Methodology*

In this Ph.D. work, based on different DL architectures presented in the literature, we focus on proposing novel XAI-based solutions for segmenting archival handwritten and printed document images. In detail, the work plan proposed in this Ph.D. is divided into five main steps (cf. Figure 3):

1. *Generate a synthetic dataset of historical document images;*
2. *Explore and evaluate different XAI solutions;*
3. *Propose novel solutions to improve the interpretability and explicability of DL architectures;*
4. *Design a publicly available XAI toolbox specifically for ADIA;*
5. *Introduce a novel set of objective metrics adapted to semantic segmentation in order to mathematically interpret the resulting explanations.*

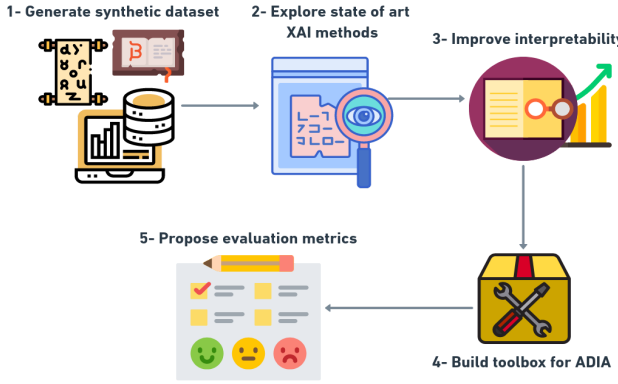



**Figure 3: Our work plan**

*1- Synthetic dataset generation*

As a first step in this Ph.D., we use the docExtractor framework for generating a synthetic dataset of historical document images [5]. docExctractor is a generic solution proposed for both synthetic dataset generation and semantic segmentation of historical document images. It has the advantage to automatically extract text-lines, figures and illustrations from historical document images that are generated synthetically. Monnier and Aubry [5] showed high performance of docExtractor as an off-the-shelf system over a wide range of datasets. For semantic segmentation purposes, a simple encoder-decoder architecture, called ResUnet, was applied. ResUnet is based on using the U-Net architecture with the ResNet-18 backbone blocks optimized with standard cross entropy loss [6]. The same U-shape of the UNet [7] was used. It is composed of three parts: encoder, bottleneck and decoder. This step is essential to have a better understanding of the resulting predictions. In later stages, we plan to assess the proposed XAI solutions using real datasets of historical documents [8].

*2- State-of-the art XAI-based solution exploration*

The second step of our work consists in exploring and evaluating different state-of-the-art XAI solutions.

*2(a)- Mathematical approach*

First, we propose to integrate different XAI techniques on the ResUnet architecture that were introduced in Monnier and Aubry' [5] framework. Then, we focus on generating two separate sets of good and poor segmented images (cf. Figure 2), and investigate them later on by computing different segmentation evaluation metrics, such as the intersection over union (IoU) or Sørensen-Dice. As suggested in the literature, the most rich location to explore on an encoder-decoder architecture is at the bottleneck stage [9]. Hence, a possible XAI solution can be explored by collecting the neuron activations at the bottleneck part of the DL network. By referring to Monnier and Aubry' [5] work, a possible investigation can be done using a mathematical paradigm, such as TCAV [10] or LIME [3]. These two methods fall into the mathematical category since TCAV is

based on concept activation vectors to explain a prediction, while LIME is based on an optimization algorithm that approximates the predictions locally with an interpretable model.

*2(b)- Gradient-based approach*

The XAI methods proposed in the literature are mainly applicable to the classification task. However, in our work, we are dealing with semantic segmentation since we are trying to assign a specific class (e.g., text or graphic) to each pixel in the image. Among these methods are, for example, the gradient-based methods that include guided backpropagation [11] and Grad-CAM [12]. To use these methods for semantic segmentation purposes, we need to make some changes to the original XAI method as described in [13].

*2(c)- Layer-wise propagation approach*

Among the most prominent XAI methods, we also cite the layer-wise relevance propagation (LRP) technique that is based on propagating the prediction backward in the DL architecture. The LRP operates by computing the magnitude of the contribution of each pixel or intermediate neuron "relevance" values. Samek et al. [2] suggested using the relevance values that propagate from the output predictions into the input layer while conserving the sum of relevance at each layer.. The LRP technique has the advantage of using much more precise information with little signal loss. There are multiple variations of LRP, one of which is SLRP [14] that improves the explanation quality of the original method. However, as stated before we need to tweak the original method in order to use it on the semantic segmentation problem as proposed in [15].

*2(d)- Explanation evaluation*

Once different state-of-the-art XAI methods have been applied on the generated synthetic dataset of historical documents, we propose to use a form of saliency maps in order to investigate the similarities and differences between the obtained explanations to the task in hand. Then, we will evaluate the resulting explainer using XAI evaluation approaches, such as image occlusion [16] that covers certain patches from the input image and measures the difference between the initial and the predicted scores of the DL model or by referring to domain experts [17,18].

*3- Proposed XAI-based solution*

As a third step of our work, we plan to propose novel DL models that are inherently interpretable by embedding interpretable blocks in their architecture. One of the promising methods in the literature is the usage of attention blocks [19]. We can use attention as a form of explainability although this method has faced some critics [20,21]. The attention technique is still a growing subject in which researchers are rushing to promote it and it sure apts for promising results [22]. Another idea is to use XAI as a form of "explain to improve" [23] in which we use the explanations to enhance the robustness and the accuracy of the model, thus improving its overall performance.

*4- XAI toolbox*

Most of the available XAI toolboxes and libraries only support classification [24-27]. Nevertheless, the neuroscope toolbox [28] tackles two tasks: classification and semantic segmentation used for scene segmentation. Hence, we plan to focus on these two tasks and create our own publicly available toolbox specifically for ADIA. This way we can use it as a form of model auditor.

*5- Objective evaluation metrics*

As stated in section (2.d), after generating explanations we should validate them to assess the original method in terms of interpretability, fidelity and class discriminability.
One approach is to design a set of objective metrics in order to mathematically interpret the resulting explanation maps. Poppi et al. [29] worked on evaluating the class activation maps (CAM) on classification tasks by computing the three following metrics:

*5(a)- Maximum coherency*

Since all important features that explain a prediction should be included in the CAM, when conditioning x on the CAM, given an input image x and a class of interest c, the CAM of x should not change. We then measure the correlation between the original CAM on the input image and the newly generated CAM.

*5(b)- Minimum complexity*

The explanation map must also be as simple and less complex as possible, which means that it must have the smallest number of pixels necessary to explain the prediction.

*5(c)- Minimum confidence drop*

In comparison to using the original input image, an ideal explanation map should produce the minimum drop in confidence.

Extending these metrics to work on semantic segmentation tasks should help us further evaluate the degree of fidelity of an XAI method applied on document layout analysis.

### C. *Future work*

As future work, we plan to investigate the DocExtractor architecture by using many synthetic datasets and by applying the main XAI-based methods proposed in the literature in order to evaluate the robustness of its predictions. Later, we will propose an inherently interpretable DL model and assess both its performance and the fairness of the explanations it provides. In later stages of the Ph.D, we are willing to combine these methods to provide our own toolbox for interpreting and explaining the decisions of DL models applied on ADIA.

## References


[1] M. Du, N. Liu and X. Hu, Techniques for interpretable machine learning, CACM, 2019, 63(1), pp. 68-77.

[2] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders and K. R. Müller, Explaining deep neural networks and beyond: a review of methods and applications." Proceedings of the IEEE, 2021, 109(3), pp. 247-278.

[3] M. T. Ribeiro, S. Singh and C. Guestrin, Why should I trust you? explaining the predictions of any classifier, SIGKDD, 2016, pp 1135-1144.

[4] T. Hastie and R. Tibshirani, Generalized additive models, Wiley Online Library, 1990.

[5] T. Monnier and M. Aubry, docExtractor: an off-the-shelf historical document element extraction, ICFHR, 2020, pp. 91-96.

[6] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, CVPR, 2016, pp. 770-778.

[7] O. Ronneberger, P. Fischer and T. Brox, U-Net: convolutional networks for biomedical image segmentation, MICCAI (3), 2015, pp. 234-241.

[8] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki and R. Ingold, DIVA-HisDB: a precisely annotated large dataset of challenging medieval manuscripts, ICFHR, 2016, pp. 471-476.

[9] A. Janik, Interpretability of a deep learning model for semantic segmentation : example of remote sensing application, Dissertation, 2019.

[10] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler and F. Viegas, Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), PMLR, 2018, pp. 2668-2677.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, Striving for simplicity: the all convolutional net, 2014.

[12] R. R. Selvaraju, M. Cogswell , A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, ICCV, 2017, pp. 618-626.

[13] K. Vinogradova, A. Dibrov and G. Myers, Towards interpretable semantic segmentation via Gradient-weighted class activation mapping, 2020, AAAI, 34(10), pp. 13943-13944.

[14] Y. J. Jung, S. H. Han and H. J. Choi, Explaining CNN and RNN using selective layer-wise relevance propagation, IEEE Access, 2021, 9, pp. 18670-18681.

[15] G. Chlebus, N. Abolmaali, A. Schenk and H. Meine, Relevance analysis of MRI sequences for automatic liver tumor segmentation, 2019.

[16] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, ECCV, 2014, pp. 818-833.

[17] G. Montavon, Gradient-based vs. propagation-based explanations: an axiomatic comparison, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, pp. 253-265.

[18] B. Zhou, D. Bau, A. Oliva and A. Torralba, Comparing the interpretability of deep networks via network dissection, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, pp. 243-252.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, 2017.

[20] S. Jain and B. C. Wallace, Attention is not explanation, 2019.

[21] S. Wiegreffe and Y. Pinter, Attention is not not explanation, 2019.

[22] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz and D. Terzopoulos, Image segmentation using deep learning: a survey, PAMI.

[23] A. Adadi and M. Berrada, Peeking inside the black-box: a Survey on explainable artificial intelligence (XAI), IEEE Access, 2018, 6, pp. 52138-52160.

[24] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs and H. Lipson, Understanding neural networks through deep visualization, 2015.

[25] P. E. Rauber, S. G. Fadel, A. X. Falcão and A. C. Telea, Visualizing the hidden activity of artificial neural networks, IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1), pp. 101-110.

[26] P. Linardatos, V. Papastefanopoulos and S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy, 2021, 23(1), pp. 18.

[27] M. Alber, Software and application patterns for explanation methods, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, pp. 399-433.

[28] C. Schorr, P. Goodarzi, F. Chen and T. Dahmen, Neuroscope: an explainable AI toolbox for semantic segmentation and image classification of convolutional neural nets, Applied Sciences, 2021, 11(5), pp. 2199.

[29] S. Poppi, M. Cornia,L. Baraldi and R. Cucchiara, Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis, IEEE/CVF, 2021,pp. 2299-2304.

# Segmentation, Recognition and Indexing of characters in CHAM documents

Tien-Nam Nguyen[1][0000−0002−2984−697X], Jean-Christophe Burie[1][0000−0001−7323−2855], Thi-Lan Le[2][0000−0001−9541−3905], and Anne-Valerie Schweyer[4][0000−0002−1058−8835]

[1] Laboratoire Informatique Image Interaction (L3i) La Rochelle University , Avenue Michel Crépeau, 17042, La Rochelle Cedex 1, France
{tnguye28,jcburie}@univ-lr.fr
[2] School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam.
lan.lethi1@hust.edu.vn
[3] Centre Asie du Sud-Est (CASE), CNRS, Paris, France
anne-valerie.schweyer@cnrs.fr

**Abstract.** The research topic of this PhD concerns document image analysis (DIA) on the historical Cham documents. This work includes: image denoising, text line segmentation, Cham characters recognition. We present the context of the thesis, the challenges as we encountered and the current approach we have planed to study for each task. A discussion of the advantage as well as the disadvantage of each approach is also provided. Finally, the future plan of each task is presented.

**Keywords:** Historical handwritten document · segmentation · recognition · image denoising · Cham inscription

## 1 Introduction

**Student's name**: Nguyen Tien Nam.
**University**: University of La Rochelle.
**Supervisor of the thesis**: Prof. Jean-Christophe Burie, Assoc. Prof. Le Thi Lan.
**Title of thesis**: Segmentation and Recognition for historical Cham inscription images.
**Starting and expected finalization date of the PhD**: The PhD started on 02/2020 and should end in 02/2023.

## 2 Context of the thesis

Exploring cultural heritage has attracted many researchers these last decades. Historical handwritten documents are important evidence in order to understand historical events and especially the ones of extinct civilizations. The Cham inscriptions are written from the Cham language system, which has been used

from the very early centuries AD in Champa (nowadays Vietnam coastal areas) and some nearby areas. The descendants of the Cham population represent one part of the community in Southeast Asia. Nowadays, the Cham inscriptions are mainly carved on steles of stone. Over time, the aging and climatic conditions have damaged the characters and created bumps. Many unwanted parts or gaps appeared on the stones making the visual quality of the image degraded significantly. The readability has become a real challenge for archaeologists, historians, as well as for people curious about Cham culture. The preservation of this cultural heritage is an important problem that needs to be considered. The final outcome of this work is to develop some approaches that can automatically analyze and recognize these inscriptions in order to index them and give access easily to the content of Cham documents.
In this thesis, we mainly focus on the analysis and recognition tasks. This work is supervised by researchers from different domains, linguistic : Anne-Valérie Schweyer (CASE CNRS, France), image processing and analysing: Jean-Christophe Burie (L3i, France) and image recognition Thi-Lan Le: (MICA, Vietnam).

## 3 Challenges

In document image analysis, image pre-processing is an important step to improve the quality of the images for further steps. This step can include: image denoising and image binarization. With Cham inscriptions, the main issue is the degradation and the noise in the images. The dataset used for this work is constituted of documents collected within a wide chronological range from the 6th to the 15th century AD. Over time, the aging and climatic conditions have damaged the characters and created bumps. Moreover, the writing system of the letters has evolved over time becoming more complex. Many inscriptions are highly degraded. By a simple observation, it is often impossible to tell whether certain strokes on the stone are part of a letter, a group of letters or even correspond to noise. An adaptive and robust approach is vital to reduce these kind of noise.
The next challenge concerns the text line segmentation task. Line fragmentation, missing part, inter-linear glosses are the common issues we find in our Cham document like other historical handwritten documents.
The recognition task of Cham language is also challenging. In the sentences, there is no space between words, more heuristic rules need to be applied to correctly make the meaning of the word. Moreover, diverse diacritics and uncommon notation such as flowers, sun, are the difficulties we have to consider.
Two additional big issues we have encountered for all these tasks are: (1) a limited amount of data for each period, (2) the time-consuming process to create the ground truth due to the difficulties for reading these inscriptions. Many approaches have been proposed in the literature for each task. If some of them provided interesting results, these approaches fail when processing unseen sets of data. So, new approaches are necessary to tackle these challenges.

## 4 Approaches

Because of the limitation of available data, we do not know exactly the variety of all data, so we choose the spiral model as the way we handle the problem. In the first phase of this work, we are only focusing on improving the quality of the inscriptions in order to have good data for the later stages of the whole pipeline. This task includes several steps but the mostly are image denoising and image binarization.

Firstly, we tackle image denoising step. Due to the specific properties of the noise present in our images, we can not consider that this noise is an additive Gaussian white noise. Consequently, the traditional image denoising approaches such as filtering, domain transform don't give the acceptable results. Hence, we studied some data driven based approaches. In the early, Independent Component Analysis (ICA) [1] achieved some interesting results on our small dataset but when we tested the method with new data, the results were not good. We found that, the trained model relies on defined parameters, it can not be robust with the variety of the Cham inscriptions. So we choose another strategy based on deep learning. This approach we used a model, namely, image translation (image to image) transforms image from the original domain to the new domain. We trained a model which directly mapping from high degradation (very noisy) to low degradation (clean image). The results with this deep learning approach are clearly better than the traditional image denoising on both quantitative and qualitative metric. However, the trained model considers equally the role of noise pixels and foreground pixels, it leads to the results more blurrier. In order to tackle this issue, we proposed global attention fusion module (GAF) which accumulates attention from different scales of image. This module integrates with reconstruction loss as an objective function in training, it helped the model to generate images with better visual quality besides removing noise in the images. The paper presenting the preliminary results on the image denoising has been accepted at ICDAR 2021 [4]. However, as aforementioned, making ground truth data to training model is very time-consuming, so we are considering some semi-supervised or weak supervised approaches to reduce the number of training data which have to do manually.

The second task concerns the process of text line segmentation. It consists in splitting the entire inscription image to the line by line. Till now, we have studied some approaches based on hand-craft features. We tried the work of [2]. Each line is separated with an energy function applied on the whole image. The drawback of this method with our data is that some parts of the characters (ascendant part, descendant part, diacritics) at the middle space of two lines are often misclassed. It leads to the wrong segmentation in these cases. The second approach is the work of [3] where the separation of two consecutive lines is guided by the minimization of an objective function. The main problem of this method is the initialization of the start and end points. From this observation, we proposed a method which leverages both advantages of these approaches by embedding an additional cost function when computing the accumulated energy function. With this approach, in some special cases where the lines are skewed or fragmented,

the proposed method fail. We are thinking of including some pre-processing or an improvement of the current method in order to be able to process such text lines. The methods developed for the denoising and text line segmentation tasks have been evaluated and give acceptable results. This is why, we have planned to move to the next step of the pipeline : the recognition of Cham characters.

## Acknowledgment

This work is supported by the French National Research Agency (ANR) in the framework of the ChAMDOC Project, n°ANR-19-CE27-0018-02.

## References

1. Hyvärinen, A., Hoyer, P. O, and Oja, E.: “Sparse Code Shrinkage: Denoising by Nonlinear Maximum Likelihood Estimation,” Proceedings of the 11th Int. Conf. on Neural Information Processing Systems, 1998.
2. Arvanitopoulos, N. and Süsstrunk, S.: ”Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts,” 2014 14th Int. Conf. on Frontiers in Handwriting Recognition, 2014.
3. Surinta, O., Holtkamp, M., Karabaa, F., Van Oosten, J-P., Schomaker, L., and Wiering, M.: ”A Path Planning for Line Segmentation of Handwritten Documents,” 2014 14th Int. Conf. on Frontiers in Handwriting Recognition, 2014.
4. Nguyen, T., Burie, J.C, Le, L., Schweyer, A-V.: ”On the use of attention in deep learning based denoising method for ancient Cham inscription images”, 2021 16th International Conference on Document Analysis and Recognition.

# Deep Learning Methods for Scientific Documents Understanding

Francesco Lombardi[1]

University of Florence, Florence, ITA

**Abstract.** Automatic scientific document reading, understanding and interpreting are hard tasks. This document aims to describe my research plan and my point of view on such topic. After an overall of my Ph.D. plan and organization, the approach i have been using and the objectives i have been fixing will be presented and shown. Such approach is twofold: the main branch is about the research that I am bringing on about deep learning methods for Natural Language Processing, and the other one concerns graphics and plots analysis within those documents. Semantic information extraction from plots and graphics can be an important advantage for scientific papers understanding, given that they are full of useful but implicit information (e.g. methods, results, implementation details). Therefore, my main research object is to put this information all together, in order to better interpret scientific literature documents and papers.ima Finally, my publications will be shortly listed.

**Keywords:** Deep Learning · NLP · Document Understanding.

## Personal Informations

My Ph.D advisor is Prof. Simone Marinai. I started my Ph.D on November 2019 and I will finish it on November 2022. The foreseeable thesis title is "Deep Learning Methods for Scientific Documents Understanding".

## 1 Introduction

As a first steps on my first research, I have deepened the analysis of scientific documents and the textual and the graphical information they contain, continuing the study already done and presented in [1]. The main objective of my first study has been the understanding and reinterpretation of a semantic information containing data such as lines within line plots in scientific papers, with the aim of being able to represent it in a different manner and shape, letting it be accessible for people with visual impairment. After this first step, I have been going on studying many types of writings, in order to enlarge the set of documents I could have chosen as a dataset. This study has helped me going towards the study of many deep learning methods and techniques applied in the major state-of-the-art works in the field of historical documents analysis and understanding, as can be seen in [2].

## 2 Research Plan

As of the Introduction, my research development up to now has been full of different "subjects" regarding the same field of study: information retrieval from digital or digitized documents. This has been done in order to better understand the topics I would have liked to improve and explore more, going deep on the latest technologies and deep learning methods to be used to solve some specific and hard tasks. The main macro-task i am actually working on, after the mentioned experiences, is structured documents semantic meaning interpreting and understanding. For humans, reading a scientific papers requires time and effort, while for machines this is nowadays an open problem. Moreover, in addition to plain text, scientific papers contain relevant information in graphics and tables. For these and many other reasons, structured document understanding is an hard task, and several Deep Learning state-of-the-art techniques have been proposed in recent years to address Natural Language Understanding (NLU) and Processing (NLP) tasks and problems.

### 2.1 Approach

Going in practice, I am planning to build a system able to understand semantic links and similarities between structured documents as scientific papers, based on the topics they belong to and the technologies they implement. This is surely not a simple task, and needs to be tackled from various points of view and using many different technologies. Another important problem, is related to data. I have been analyzing different open datasets, and I chose some of them for my research, because of the high amount of topics they cover and the accuracy of their labels. Some of those datasets:

- S2orc [tot. 81.1M papers] [3]
- Kp20k [tot. 20K papers] [4]
- Custom (Conferences/Journals) [tot. 2k, personally selected and chosen]

After data and context definition, I have been fixing a threefold strategy. This has been designed in order to test different deep learning methods when applied in order to solve the following different tasks:

- document clustering, dealing with structured papers from many different fields of study, authors and application fields (unsupervised learning)
- classification of papers, to be performed on their major field of study (supervised learning)
- definition or learning of a metric in order to compute a distance between documents, related to their semantic similarities and relations, based on the technologies they implement.

**Document Clustering** The definition of a class to be attributed to a scientific paper of the type of those under consideration is a problem difficult to solve.

Due to the length of the documents, their non-uniform conformation, and many other reasons related both to the structure of the document and to the absence of a groundtruth, it is very complex to define a method to associate a class to these same documents. For this reason, given the large availability of papers from the academic world, one of the ways I decided to undertake is to extract the most direct and important information for individual papers (e.g. title, abstract, authors, venue, etc.) and represent them in a vector space in which, then, perform clustering operations, possibly preceded by appropriate reductions in the dimensionality of the vector space obtained.

**Classification** Parallel to what has just been described, I have also pursued a different path, which is a supervised approach to the problem. In fact, one task that can be addressed if we consider data such as those from S2ORC or other datasets of that magnitude, is that of classifying the type of scientific paper based on its field of study. This attribute is part of the labels available within the groundtruth of the dataset, and is therefore of interest. Solving this type of problem is planned in the pipeline I would like to implement. This is to be considered as an interesting contribution to make an initial classification, in anticipation of solving more specific and complex problems from the same representation of input documents.

**Metric definition** In addition to what has been described in the previous paragraphs, my goal is, once I have obtained a defined and precise representation of the documents in question and all their most relevant contents, the definition or use of an accurate and effective metric to calculate a distance between the objects present in the feature space extracted from the documents. Such a metric could prove to be an excellent method of measuring similarity (even at the semantic level) between objects such as those in question: nearby elements could be defined as similar and vice versa.

In addition to what has been described in the previous paragraphs my goal is, once obtained a defined and precise representation of the presented documents and all their most relevant contents, the definition or use of an accurate and effective metric to calculate a distance between the objects present in the feature space extracted from the documents. Such a metric could prove to be an excellent method of measuring similarity (even at the semantic level) between objects such as those in question: nearby elements could be defined as similar and vice versa.

I also believe that an important contribution to such research could be represented by the analysis of different kinds of text and data within scientific papers. The greatest candidates for adding important information to the representation of documents are the images and graphics they contain, the tables, and the text contained in their captions. This assumption stems from the fact that, in most cases, a large amount of numerical data representing findings and conclusions

are encapsulated in these types of objects. An analysis of their entropy and the possible addition of that contribution to the pipeline I have in mind is definitely an experiment I plan to do soon.

## 3 Conclusion

As a conclusion, I look forward to deepen the study started on NLP methods, with the aim of creating a system which may be able to read, understand and extract (through deep learning methods) specific information contained in free text within structured literature documents, tables and images.

## References

1. F.Lombardi, C.Goncu, S.Marinai, *Line Recognition for Generating Accessible Line-Plots*, 13th IAPR International Workshop on Graphics Recognition (GREC), Sydney, Australia, 2019.
2. F.Lombardi, S.Marinai, *Deep Learning for Historical Document Analysis and Recognition—A Survey*, J. Imaging 6.10, 2020, p.110.
3. Lo, Kyle and Wang, Lucy Lu and Neumann, Mark and Kinney, Rodney and Weld, Daniel, *S2ORC: The Semantic Scholar Open Research Corpus*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
4. Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., and Chi, Y. (2017). Deep keyphrase generation. arXiv preprint arXiv:1704.06879.

# Font Design Analysis: Understanding Designers' Knowledge by Using Machine Learning

Student's name: Daichi Haraguchi
Supervisor name: Seiichi Uchida

Affiliation: Kyushu University, Japan
Starting date of my PhD: April 01, 2021
Expected end date of my PhD: March 31, 2024
{daichi.haraguchi,uchida}@human.ait.kyushu-u.ac.jp

**Abstract.** This paper proposes a research plan about a novel font design analysis considering designers' knowledge. There are a lot of subjective and small-scale font design analyses in previous works, such as questionnaire analysis, multivariate analysis, and so on. On the other hand, I propose objective font analysis by using a large-scale dataset. For this research, I propose the method in three steps. The first step is to decompose a font design into basic shape (e.g., skeleton) and design elements (e.g., color and thickness) to focus on only design elements in the next step of the analysis. The second step is to analyze the correlation between designers' knowledge and design elements. The last step is to generate a novel font design by using the result of the analysis step. This research might have an impact on the DAR community in terms of the first attempt to analyze considering designers' knowledge to the authors' best knowledge.

**Keywords:** font design analysis · font generation · machine learning.

## 1 Research Plan

### 1.1 Background and Purpose

I work on font design analysis that is interdisciplinary research between computer vision and design. Font designs that are used in many situations are selected by specialists (especially designers) who consider the many kinds of elements such as background, target reader, and so on. Font designs consist of two elements. First is basic shape including a message which specialists mainly want to send. The second is a design element that is possible to add impression. However, the method of selecting the design element has been to inherit among specialists or to find the trend by subjective experiment on a small scale. In previous works [1, 2], trends of font usage have been analyzed in subjective experiments.

I, therefore, try to objectively quantify the trend on designs (so-called tacit knowledge of designers) by using a large-scale dataset in order to support selecting fonts. In the end, I try to create novel font designs by using quantified tacit

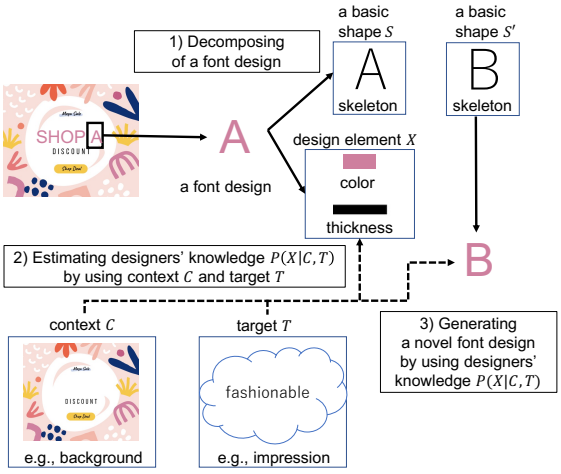



**Fig. 1.** Overview of the proposed research.

knowledge of designers. As I consider the above, my research purposes are three summarized below.

1. Decomposing elements of font design into basic shape and design element
2. Obtaining the value of tacit knowledge by analyzing large scale dataset of font designs
3. Creating novel font designs by using tacit knowledge of designers

Font designs are consisted based on "(i) basic shape of font designs $S$ (e.g., the skeleton of 'A'), (ii) design element $X$ (such as color and decoration) subject to (iii) target of font designs $T$ (such as the impression that designers want to send), (iv) context where fonts are used $C$ (such as background images, target readers)." From these (i) to (iv), probability distribution $P(X|C,T)$ is calculated. This is a designers' knowledge.

I conduct this research with three steps corresponding to the above three purposes. Step 1 is to decompose a font design into basic shape $S$ and design element $X$. Note that target $T$ and context $C$ are not in font design images but metadata (Figure 1-1). Step 2 is to analyze the correlation between target $T$, context $C$, and design element $X$, then to obtain designers' knowledge (Figure 1-2). Finally, step 3 is to generate the novel font designs considering designers' knowledge by using $P(X|C,T)$ applied into another basic shape $S'$ (Figure 1-3).

### 1.2 Difficulties of This Research

There are mainly third difficulties in this research. I consider the below difficulties in all steps.

First, this research introduces unwieldy metadata (target $T$, context $C$). The target $T$ and context $C$ might have outliers (e.g., intentional use of "unusual" designs) and noises (e.g., subjective variance).

Second, the ground truth of our task might be time-dependent, as human preferences change over time with shifts in fashion. In fact, the trend of font usage by era on movie posters was found in [3].

Finally, the most difficult is that not all design elements may be easily quantifiable. There is an infinite number of design elements, however, I address them as much as possible by using disentangled representation, and so on.

## 2 Steps and Action Plan So Far

I have already conducted three experiments for this research related to the above three purposes.

First, I conducted a font identification task (publication [1]) that identify whether the character pairs have the same font designs. I confirmed that it successfully obtained the font design features to identify font designs by analyzing the results. This is part of step 1.

Second, I conducted font designs generation (publication [2]) that match the background images of book covers. This task corresponds to part of steps 2 and 3. In this research, I didn't focus on the analysis; therefore, I couldn't understand what font designs match the background images (this is learned by neural network implicitly).

Third, I introduced shared latent space between font designs and their impressions (publication [3]). Font designs and impressions are in the same latent space; therefore, it is possible to generate font designs matching the impressions from latent variables of impressions. Although this research corresponds to step 3, the correlation between font designs and impressions is implicitly learned by a neural network.

So far, the above three steps are conducted separately. I am now preparing to experiment with the step 1 to 3 as one flow. It means that it is conducted by using font designs on the same material such as grayscale font designs or font designs on book covers from decomposing of font designs to generating font designs. Especially, I'm going to focus on grayscale font designs in the first year, and other font designs in the second and third year.

## 3 Originality and Novelty

There are two originality and novelty points in this research.

First, this research is combined with many fields such as information science, design psychology, and effective engineering. In the past, most of the impression analysis of font design in design psychology and effective engineering was done by subjective evaluation, such as showing the font design to the subjects and having them evaluate their impressions [1, 2]. In this research, I analyze the impressions of font designs using an information engineering and computer science approach. It means that we will analyze font designs based on a quantitative and objective method as much as possible by decomposing font designs into elements using a large-scale dataset and machine learning.

Second, this research introduces unwieldy metadata (target $T$, context $C$) into the analysis. A target is, for example, to give a specific impression to a font or make a web advertisement attractive and increase the click rate. The metadata is human sensory data and is considered to be a unique element of graphic design, including font designs. For this instability and situational dependency, we propose new analysis methods using various techniques (robust estimation and trend analysis). As for the context, it is necessary to deal with various viewpoints. For example, when I analyze the font designs printed on a book cover image, I might need to consider not only the image of the cover but also the title itself (word sequence), the genre of the book, and other multiple contexts.

## 4 Impact on Future from This Research

Successful analysis of impressions of design elements will make it possible to make appropriate choices without any special knowledge or experience. The success of automatic visual design generation will enable the acquisition of font designs that reflect the user's intentions. These results will impact on DAR community because font design A(nalysis) and font R(ecognition) are important in this research. Except for the academic side, it is expected that this will enable more effective and efficient marketing advertisements and materials. It will be a revolutionary technology for the marketing industry.

## Acknowledgements

I would like to express my deepest appreciation to Prof. Nicholas Howe for his support.

## Publications

1. **D. Haraguchi**, S. Harada, B. K. Iwana, Y.Sinahara and Seiichi Uchida, "Character-independent font identification," DAS2020.
2. T. Miyazono, B K. Iwana, **D. Haraguchi**, S. Uchida, "Font Style that Fits an Image–Font Generation Based on Image Context," ICDAR2021.
3. J. Kang, **D. Haraguchi**, A. Kimura, S. Uchida,"Shared Latent Space of Font Shapes and Impressions," arXiv preprint arXiv:2103.12347 (2021).

## References

1. K. Nonaka, J. Saito, and S. Nakamura. "Music Video Clip Impression Emphasis Method by Font Fusion Synchronized with Music," ICEC-JCSG2019.
2. T. Venkatesan, Q.J. Wang, and C. Spence. "Does the typeface on album cover influence expectations and perception of music?," Psychology of Aesthetics, Creativity, and the Arts (2020).
3. K. Tsuji, S. Uchida, B.K. Iwana, ""Using Robust Regression to Find Font Usage Trends," arXiv preprint:2106.15232 (2021).

# Structural Analysis and Understanding of Complex Layouts in Document Images

Student Name: Sanket Biswas[1]
Supervisor(s) of the Thesis: Josep Lladós[1], Pau Riba[1], and Umapada Pal[2]
Starting date of PhD: November 2020
Expected Finalization date of PhD: October 2023

[1] Computer Vision Center & Computer Science Department
Universitat Autònoma de Barcelona, Spain
sbiswas@cvc.uab.es
[2] CVPR Unit, Indian Statistical Institute, India

**Abstract.** Information extraction is a fundamental task of most business intelligence services which entail massive document processing. Understanding a document page structure in terms of its layout provides contextual support which is helpful in the semantic interpretation of the document terms. In this thesis, inspired by the progress of deep learning methodologies applied to the task of object recognition, we transfer these models to the specific case of document object detection, reformulating the traditional problem of document layout analysis. Moreover, we also contribute to prior arts by defining the task of instance-level segmentation on the document image domain. Our next problem was to handle the scarcity of labeled training data for the document object detection task. In this context, we have explored another research direction of synthetic document generation. A novel layout-guided document image synthesis framework using Generative Adversarial Networks(GANs) has been proposed in our study for creating synthetic document image datasets for augmenting real data during training for document layout analysis tasks. Finally, we have also explored a deep geometric approach with Graph Neural Networks (GNNs) to generate synthetic data with highly variable and plausible document layouts that can be used to train document interpretation systems with far more efficiency. This research plan has been organized into two subsections: Section 1 throws light on the research problem and the ongoing works of the thesis while Section 2 discusses the future directions of the thesis.

## 1 Thesis Overview

Layouts play a significant role in dictating the reader's attention and hence, the order by which it conveys the information. The layout structure of a document is fundamentally represented by layout objects (e.g. text or graphic blocks, images, tables, lines, words, characters and so on) while the logical structure conveys the semantic relationship between conceptual elements (e.g. company logo, signature, title, body or paragraph region and so on). The understanding of

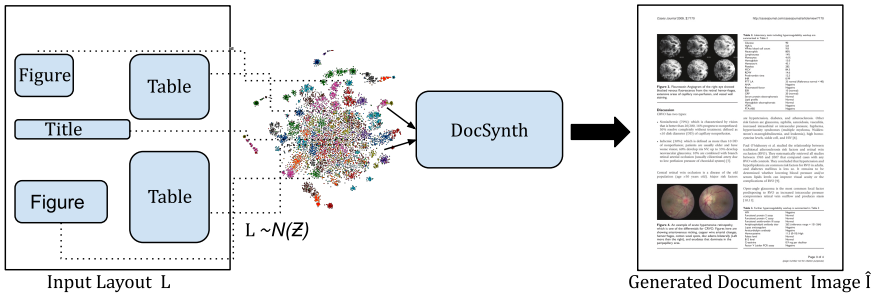



Fig. 1: **Layout-Guided Document Image Synthesis [2]:** Given an input document layout with object bounding boxes and categories, DocSynth samples the semantic and spatial attributes of every layout object from a normal distribution to generate realistic images.

complex document layouts is an important step towards the goal of information extraction. Business intelligence processes often require the extraction of useful information from document contents at a large scale, for subsequent decision-making actions. Consequently, there is a need for annotated data to supervise the learning tasks. In addition to the manual effort to annotate layout components, such types of images have privacy restrictions (personal data, corporate information) which prevent companies and organizations to disclose it. Data augmentation strategies provide a good solution. Among the different strategies for augmenting data, synthetic generation of realistic document images [12] is one of the solutions. Another strategy would be to generate plausible document layouts [7] which can define the structural information of a whole page.

Until now, the thesis has been driven towards progress in both the research domains of document object detection and synthetic document generation. In the first domain, we have focused on the idea of instance-level segmentation of complex document layouts inspired by original Mask-RCNN model [5] as shown in Figure 2. Having defined both detection and segmentation modules for document object detection, the proposed approach in our work [1] analyzes the problem of segmenting overlapping layout objects, specially in documents having a hierarchical content structure [11]. In the second domain, we have primarily focused on designing a GAN-based model called DocSynth [2] which can synthesize multiple realistic document images, guided by a spatial layout(bounding boxes with object categories) given as a reference by the user. This layout-guided document image synthesis has also been defined as a novel task in our work [2] as illustrated in Figure 1. The proposed method, is indeed able to learn and understand the complex interactions among the different layout components to generate synthetic document images that fulfills geometric and semantic consistency with the given layout. In one of our recent studies, we have also explored

a graph-based generative approach [3] using GNN's towards learning and generation of complex structured layouts using administrative invoice documents as a case study. Inspired by the READ framework for Document Layout Generation [9] and GNN-based table detection [10] and named-entity recognition in semi-structured documents [4], this work explores the power of graph representations to render synthetic document layouts in the form of diverse graphs that can perfectly match the structural characteristics of the target data.

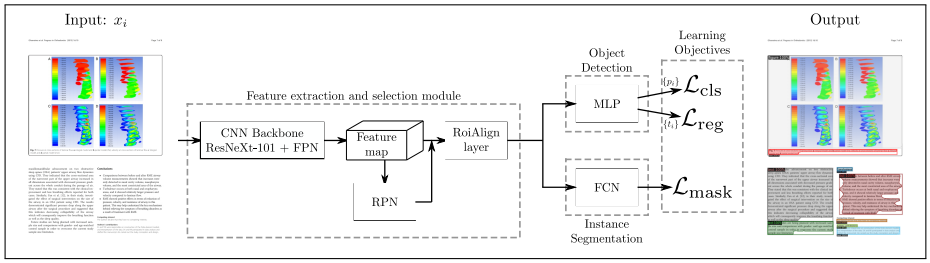



Fig. 2: **Proposed Instance-Level Layout Segmentation framework [1]:** Given an input image of a document, the model predicts the different layout elements, with object detection on one module head and instance-level segmentation on another module head.

## 2 Future Work

Our proposed end-to-end instance segmentation model [1] widened the scope of Document Object Detection, from coarse-grained detection with bounding boxes to a more fine-grained segmentation with instance masks. Experiments for document layout segmentation exhibit good results on both scientific articles and historical document datasets. Following this preliminary work, the architecture of our present model can be further improved and extended to more complex semi-structured administrative documents like invoices, forms and receipts where understanding geometrical relations between the layout elements is even more challenging. Also, there is great scope for improving this prototype using domain adaptation strategies since documents from the diverse domains may vary significantly in layout, language, and genre. The cross domain document object detection work by Li. et. al. [8] is an inspiration towards this direction.

The DocSynth [2] model for layout-guided document image synthesis opened a new direction in the field of synthetic document generation. Since this task can prove to be highly beneficial for augmenting training data to improve document interpretation systems, it provides more reason to broaden its scope. One of the next objectives in our thesis would be to incorporate more awareness to the content and template to this process, as done in case of layout generation of graphic

contents by Lee et. al. [6] in their recent work. The graph representation learning principle [3] adapted for generation of synthetic layouts for semi-structured documents also remains an area to be focused in future. In this direction, we would also want to explore evaluation metrics to introduce a proper quantitative comparison between the real and synthetically generated layouts in document graphs.

## References

1. Biswas, S., Riba, P., Lladós, J., Pal, U.: Beyond document object detection: Instance-level segmentation of complex layouts. International Journal on Document Analysis and Recognition (IJDAR) (2021)
2. Biswas, S., Riba, P., Lladós, J., Pal, U.: Docsynth: A layout guided approach for controllable document image synthesis. In: International Conference on Document Analysis and Recognition (ICDAR) (2021)
3. Biswas, S., Riba, P., Lladós, J., Pal, U.: Graph-based deep generative modelling for document layout generation. In: International Conference on Document Analysis and Recognition Workshops(ICDARW) (2021)
4. Carbonell, M., Riba, P., Villegas, M., Fornés, A., Lladós, J.: Named entity recognition and relation extraction with graph neural networks in semi structured documents. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9622–9627. IEEE (2021)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
6. Lee, H.Y., Yang, W., Jiang, L., Le, M., Essa, I., Gong, H., Yang, M.H.: Neural design network: Graphic layout generation with constraints. ECCV. Springer, Heidelberg (2020)
7. Li, J., Yang, J., Hertzmann, A., Zhang, J., Xu, T.: Layoutgan: Synthesizing graphic layouts with vector-wireframe adversarial networks. IEEE transactions on pattern analysis and machine intelligence (2020)
8. Li, K., Wigington, C., Tensmeyer, C., Zhao, H., Barmpalios, N., Morariu, V.I., Manjunatha, V., Sun, T., Fu, Y.: Cross-domain document object detection: Benchmark suite and method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12915–12924 (2020)
9. Patil, A.G., Ben-Eliezer, O., Perel, O., Averbuch-Elor, H.: Read: Recursive autoencoders for document layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 544–545 (2020)
10. Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., Lladós, J.: Table detection in invoice documents by graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 122–127. IEEE (2019)
11. Shen, Z., Zhang, K., Dell, M.: A large dataset of historical japanese documents with complex layouts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 548–549 (2020)
12. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8584–8593 (2019)

# Automatic recognition of historical handwritten parish records.

Solène Tarride
*Doptim, Univ Rennes,*
*Rennes, France*
*Email: solene.tarride@irisa.fr*

***Abstract*—French parish registers are handwritten books in which local religious ceremonies were recorded from the 16th century onward. These documents are especially useful to genealogists because they contain local information on births, marriages and deaths. Automatic information extraction would allow genealogists to index and search parish registers by content. The aim of the PhD research project is to develop a pipeline able to extract information from these registers. Processing these documents can be done in three steps: 1) layout analysis to localize text-lines and records, 2) handwriting recognition to get the text associated to each record, 3) information extraction to localize relevant words such as names, surnames, location and dates.**

***Keywords*-document layout analysis; handwriting recognition; information extraction; historical documents, deep neural networks.**

## I. About me

I am an industrial PhD student working for Doptim, a company specialised in data science. My research project is supervised by Bertrand Coüasnon who works in the Intuidoc team from Univ Rennes/IRISA. This research team is specialized in automatic document processing.
This PhD project began on February 2019 and will end in February 2022.

## II. Research Plan

### A. Introduction

French parish registers are handwritten books written by priests from the 16th to the 19th century. These documents are structured in records, which are paragraphs describing a religious ceremony (mainly baptism, marriage and burial). Parish registers are especially used by genealogists to find reliable information about their ancestors. They are especially relevant to find people born before French Revolution, as mandatory civil registration was established after 1792. If most parish registers are now accessible online, the search for ancestors remains time-consuming and laborious as it is necessary to search the archives to find relevant records. Moreover, reading these records is challenging, as the handwriting style is cursive. There are also many different writers with varying handwriting styles and phrasing, even within the same register. Finally, parish registers are often degraded, mainly with ink fading and bleed-through. As a result, there is a need for automatic methods able to extract relevant information from these documents. These records could then be indexed and searched by content (i.e. by name, date, location) which could substantially ease the search for ancestors.

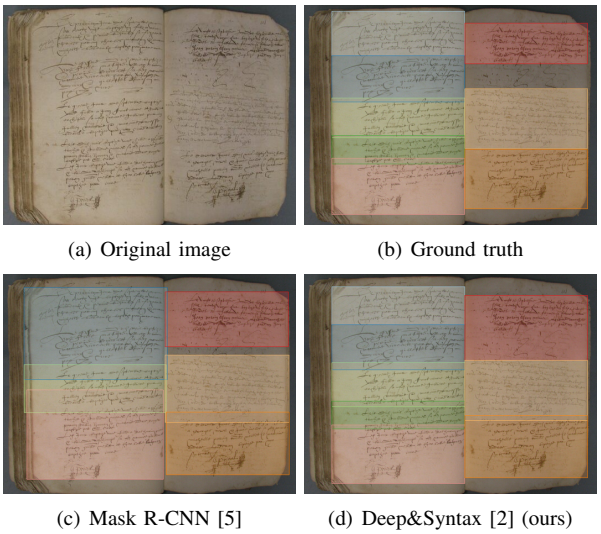
### B. Work achieved so far

First, we have focused on layout analysis, mainly to extract text-lines and records from these registers. Then, we focused on handwritten text recognition (HTR) and named entity recognition (NER) to extract relevant information for genealogists.

*1) Annotation of the database:* We labeled 410 images (3700 records) of parish registers by localizing page borders, text baselines, signatures, and record boundaries. A software has also been designed to allow user transcription. We are currently transcribing these records, but this process is difficult and time-consuming. So far, only 150 records (800 text-lines) have been transcribed.

*2) Structure analysis:* Record are complex objects that can vary in size and writing style. They can even overlap each other when a signature overlaps the text of the next record. As a result, training a model to localize record bounding boxes is challenging. To overcome this difficulty, we created a grammatical description of the records. First, we extract first text-lines, text-lines, signatures, and page borders from the image using U-shaped convolutional neural network [3] [4]. Then, these structural patterns are grouped using logical rules to form records. We typically define a record as a first text-lines followed by a group of text-lines and a group of signature. Other logical rules are also designed to overcome potential segmentation errors. This approach has the advantage of being easy to train with few labeled images, as it relies on the segmentation of more stable patterns. Training on 60 pages yield acceptable results. We compared this approach with object detection neural networks and found out that they performed worse, even when trained on more than 150 images. An example of prediction can be observed in Fig. 1. We have published two articles on our contribution for structure recognition [1], [2].

*3) Handwriting Recognition:* As we are still in the process of getting transcriptions from parish registers, all experiments were carried out on the Esposalles database. This database is composed of records that are similar to parish registers. We have explored the sequence-to-sequence



(a) Original image (b) Ground truth

(c) Mask R-CNN [5] (d) Deep&Syntax [2] (ours)

Figure 1. Illustration of a record segmentation obtained using object detection networks (c) and our original approach (d).

(seq2seq) architecture with attention mechanism for handwriting recognition, and have compared it with the usual CRNN-CTC approach. We found out that the seq2seq architecture achieves competitive performance. Few researchers have investigated this architecture for handwriting recognition, even though it is massively used for other research areas, such as automatic translation and image captioning.

Several experiments have been carried out on the seq2seq architecture. First, we investigated the loss function, and found out that using a hybrid loss, composed of a CTC loss after the encoder and a CE loss after the decoder, helps the network to learn a relevant representation of the image. We also compared several encoder's architecture. Typically, the encoder used for image captioning is a fully convolutional neural network (CNN). But for the task of handwriting recognition, it can be relevant to add recurrent layers to the encoder. We compared several convolutional-recurrent architectures (CRNN), mainly: CNN-GRU and CNN-LSTM. A study was also carried out to find the optimal number of recurrent layers and number of directions. The results indicate that a CNN coupled with a 2-layer bi-dimensional GRU improves handwriting recognition on the Esposalles database. Finally, we compared different ways to compute the attention vector and attempted to penalize the attention from left to right.

*4) Information Extraction:* This task is generally achieved by performing named entity recognition on the output of the handwriting recognition system. However, we believe that knowing the semantic context beforehand can be useful to handwriting recognition systems. An interesting approach that predicts characters and semantic tags simultaneously was proposed using the traditional CRNN-CTC approach [**?**]. In this scenario, the encoder learns a more global image representation to predict the context as well as the characters. For example, the network could predict the following sequence:

```
<child-name>Françoise <child-surname>Bouvet
daughter of <father-name>Julien <father-sur
name>Bouvet died in <location>Andouillé.
```

Two approaches for joint HTR and NER were explored during this thesis. First, we trained a seq2seq to model this tag sequence. We report a higher performance compared with the traditional CRNN-CTC architecture, proving the interest of the seq2seq architecture.

Second, we designed a multi-task seq2seq network composed of multiple decoders connected to a single encoder. This architecture allows to generate one sequence for each decoder. Multi-task seq2seq architectures have already been proved efficient for automatic translation [9] as compared with the single-task scenario. For handwriting recognition, multi-task learning can be used to predict simultaneously the sequence of characters and the sequence of semantic categories. As a result, the encoder learns an image representation that is adapted for both task. The most conclusive experiment consists to predict multiple tag sequences (1-category, 2-person, 3-mixed), such as:

```
1   <name>Joseph born in...
2   <husband>Joseph born in ...
3   <husband-name>Joseph born in ...
```

Multi-task seq2seq yields to a higher performance than for the single task scenario, even when evaluating only the third sequence. Results can be observed in Table I. This multi-task architecture could be improved by designing a post-processing step that takes advantage of the multiple outputs predicted by the network. Moreover, language modeling could also be used to correct linguistic errors.

Table I
COMPARISON OF DIFFERENT APPROACHES FOR INFORMATION EXTRACTION USING THE IEHHR EVALUATION PROTOCOL.

| Architecture | Basic score | Complete score |
|---|---|---|
| HMM-baseline [8] | 80.28 | 63.11 |
| CRNN-CTC + LM + NER [7][1] | 95.46 | 95.03 |
| CRNN-CTC-tag [6] | 90.59 | 89.40 |
| Seq2seq-tag (ours) | 91.42 | 89.44 |
| Multi-task seq2seq-tag (ours) | 93.18 | 92.13 |

### C. Future work

We are considering two objectives to achieve before the end of the thesis:

- We would like to publish our findings on the seq2seq architecture for information extraction in handwritten records. If this approach is promising, it cannot be realistically used by Doptim for now (more transcriptions of parish registers are needed).

[1]Not directly comparable as it is the only method using post-processing and language modeling

- We would like to design a pipeline that could be exploited by Doptim with few labeled images. To achieve this, we are considering training a generic HTR, for example a seq2seq trained on multiple available real and synthetic databases. Then, we are considering training a few-shot learning NER system, such as FLAIR [10], to tag the sequence predicted by the HTR system.

As a long-term outlook, Doptim is planning to deploy a collaborative annotation platform to obtain large-scale transcriptions of various records from different time periods and locations. Getting more labelled data would allow us to perform a transverse analysis on similar records to make the information extraction even more reliable. Indeed, records from a same family often contain the same names, surnames and location, and records from the same register often share the same handwriting style and phrasing.

## III. Conclusion

The aim of this thesis is the recognition of French parish registers. Extracting information from this documents would allow to search and index them by content, which could substantially ease the search for ancestors. Over the three years of this project, we have developed an original method for parish registers layout analysis with few labeled data. We have also focused on handwriting recognition and information extraction using a sequence-to-sequence architecture with an attention mechanism. Finally, we have labelled 3700 records and transcribed over 700 text-lines.

## References


[1] S. Tarride, A. Lemaitre, B. Coüasnon, S. Tardivel. *Signature detection as a way to recognise historical parish register structure*, In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP '19), Association for Computing Machinery, New York, NY, USA, 54–59, 2019. DOI:https://doi.org/10.1145/3352631.3352636

[2] S. Tarride, A. Lemaitre, B. Coüasnon, S. Tardivel. *Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples*. International Journal on Document Analysis and Recognition (IJDAR), 2021. https://doi.org/10.1007/s10032-021-00362-8

[3] Grüning, T., Leifert, G., Strauß, T. et al. *A two-stage method for text line detection in historical documents*. IJDAR 22, 285–302 (2019). https://doi.org/10.1007/s10032-019-00332-1

[4] S. Ares Oliveira, B. Seguin and F. Kaplan, *dhSegment: A Generic Deep-Learning Approach for Document Segmentation*, 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 7-12, doi: 10.1109/ICFHR-2018.2018.00011.

[5] K. He, G. Gkioxari, P. Dollár and R. Girshick, *Mask R-CNN*, 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[6] Carbonell, Manuel Villegas, Mauricio Fornés, Alicia Lladós, Josep. (2018). *Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model*.

[7] Animesh Prasad, Hervé Déjean, Jean-Luc Meunier, Max Weidemann, Johannes Michael, Gundram Leifert (2018). *Benchmarking Information Extraction in Semi-Structured Historical Handwritten Records*. ArXiv, abs/1807.06270.

[8] A. Fornés et al., *ICDAR2017 Competition on Information Extraction in Historical Handwritten Records*, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1389-1394, doi: 10.1109/ICDAR.2017.227.

[9] Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, Lukasz Kaiser (2016). Multi-task *Sequence to Sequence Learning*. In International Conference on Learning Representations.

[10] Akbik, R. (2019). *FLAIR: An Easy-to-Use Framework for State-of-the-Art* NLP. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 54–59). Association for Computational Linguistics.

# Automated Summarization of Legal Judgements

The Case of Supreme Court of Pakistan

Sahar Arshad *(PhD Scholar)*
Dr. Ahmed Slaman *(Supervisor),* Prof. Dr. Faisal Shafait *(Co-Supervisor)*
Department of Computing, School of Electrical Engineering and Computer Science, NUST
Islamabad, Pakistan
Starting Date PhD: October 2020
Expected Final Date: December 2024
sarshad.phdc20seecs@seccs.edu.pk

***Abstract*— Legal text summarization is an emerging area in the field of natural language processing. Legal text has its own characteristics such as size, structure, vocabulary, ambiguity, and citations of precedence of similar nature that makes this domain unique in its nature. Although some studies have addressed the legal text summarization for a particular country, there is no systematic evidence that a methodology developed for a particular country can be generalized for the legal context of another country. This research work is focused on the text summarization of legal documents in the context of judicial system of Pakistan. We explore the potential of the pre-trained transformer model as well as the state-of-the-art deep learning libraries for text summarization of Supreme Court of Pakistan judgements. Experimental results suggest that our proposed approach effectively model the corresponding relations between judicial orders and its generated summaries, with a much better evaluation score as compared to other baseline approaches.**

***Keywords: Text Summarization, Legal Documents, Supreme Court of Pakistan, BART, FastAI***

## I. INTRODUCTION

In countries that follow the Common Law jurisdictions (for example, the United Kingdom, Canada, India, and Pakistan), there are two main origins of laws: legislative statutes (constitution) and precedents (judge-made law). The precedents aid a legal expert to synthesizes how the court has handled those similar cases in the past as applicable to the current facts which are mostly archived as legal reports / case decisions that run independently with long, dense legal text. This makes it difficult for even a legal expert to comprehend the full text of a case [1].

To address a legal problem, lawyers have to look through several past decisions to support their cases so they may extract richer and relevant legal information from large archives. On the other hand, novice users are often interested to know if there is prior evidence of such court judgements (summaries). Thus, automated text summaries can be a vital tool to help these stakeholders to cut down cognitive effort as well as to explore key points that must be included in case summaries [2]. One of the key challenges in this area is that most of the previous researches of automated summaries are limited to newspaper articles and scientific papers [3]. However, legal text has its own characteristics such as size, structure, vocabulary, ambiguity, and citations of precedence of similar nature that makes this domain unique in its nature. Therefore, the summarization of legal text requires special treatment and demands a study in itself, which is different from the summarization of the general text and news articles [2].

This research work is focused on the text summarization of legal documents in the context of the judicial system of Pakistan. It is worth mentioning that some studies have addressed the legal text summarization, e.g., UK case judgements [4], Singapore case judgements [5], and Indonesian case judgements [6] but there is no systematic evidence that a methodology developed for a particular country can be generalized for the legal context of another country. Every country has its own legal requirements, standards, structure, and terminologies which vary according to the social traditions of another country. Therefore, a summarization approach developed in the context of a particular country cannot be generalized, keeping in view the legal requirements of another country [1]. In this work, our contributions are:

- To the best of our knowledge, this is the first research effort in the domain of legal documents of the Supreme Court of Pakistan to address the text summarization approach.
- We investigated the significance of transformer architecture to effectively employ for the summarization task by using BART pre-trained model along with FastAI and Pytorch libraries. This is also a first step in the legal domain direction.
- We compare the proposed model as a stepping stone in this domain by achieving better results in contrast to the baseline approaches being used in this domain.

## II. BRIEF RESEARCH PLAN

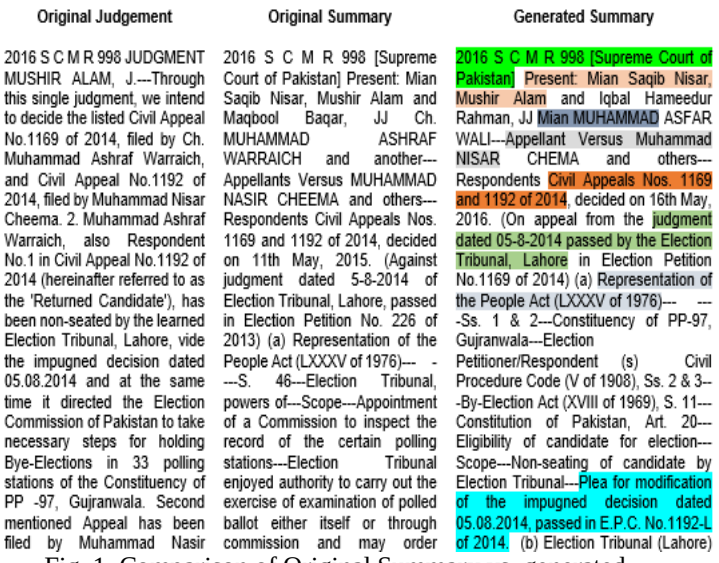
Language model pre-training has been emerged as an advanced approach in NLP tasks, ranging from sentiment classification, question answering, named entity recognition, and semantic similarity. Many of the state-of-the-art NLP models are developed using this transformer architecture,

based on the orientation of attention encoder-decoder framework to convert the input sequences into output sequences. There are two types of transformer architectures that are commonly used: the transformer encoder and the decoder of the transformer. State-of-the-art pre-trained models include ELMo, GPT, BERT, and most recently BART transformer model. BART is a pre-trained de-noising auto-encoder seq2seq model for the generation, translation, and comprehension of natural languages. This model can be treated as a generalized BERT with a typical seq2seq/machine translation architecture with a bidirectional encoder as well as a GPT by using a decoder from left to right. Training is performed by masking text with an arbitrary noise function and then learn a model to rebuild the original text. The solution to noise entails random shuffling of the order of the initial sentences and the use of a new in-filling scheme. In this study, we explore the potential of BART for text summarization of Supreme Court of Pakistan judgements.

In the first step, the dataset containing 200 Supreme Court of Pakistan judgements and their summaries were loaded from a json file into a data frame and were tokenized using FastAI tokenizer. The data was preprocessed by removing extra spacing and word repetition and adjusting the maximum sequence length. The data was then split into a training testing set with a ratio of 80, 20. In the subsequent step, the text summarization model was initiated and trained using BART transformer based on the PyTorch system and CUDA, an Nvidia- developed parallel computing platform. New summaries are then generated using the trained models and summarization results are evaluated using ROUGE-1, ROUGE-2, and ROUGE-L on a full-length F1 Score.

The proposed model was fine-tuned after a series of experiments. In the first epoch, the model produced an average evaluation score as compared to other models especially LSA. However, after a number of iterations, the proposed model was able to provide a better comprehension and understanding of the judgments and the evaluation score as well as the generated summaries which were much relevant to the original summaries [7].

The proposed model was able to guess the judgement number, name of the judges, and judgement key decision lines. However, there were few inconsistencies in the names of respondents that were results of the stemming and masking anomalies. In general, the generated summaries were meaningful and much relevant to the original decision. A part of the original judgement, actual summary, and the generated summary is presented in Fig. 1.



| Original Judgement | Original Summary | Generated Summary |
|---|---|---|
| 2016 S C M R 998 JUDGMENT MUSHIR ALAM, J.---Through this single judgment, we intend to decide the listed Civil Appeal No.1169 of 2014, filed by Ch. Muhammad Ashraf Warraich, and Civil Appeal No.1192 of 2014, filed by Muhammad Nisar Cheema. 2. Muhammad Ashraf Warraich, also Respondent No.1 in Civil Appeal No.1192 of 2014 (hereinafter referred to as the 'Returned Candidate'), has been non-seated by the learned Election Tribunal, Lahore, vide the impugned decision dated 05.08.2014 and at the same time it directed the Election Commission of Pakistan to take necessary steps for holding Bye-Elections in 33 polling stations of the Constituency of PP -97, Gujranwala. Second mentioned Appeal has been filed by Muhammad Nasir | 2016 S C M R 998 [Supreme Court of Pakistan] Present: Mian Saqib Nisar, Mushir Alam and Maqbool Baqar, JJ Ch. MUHAMMAD ASHRAF WARRAICH and another---Appellants Versus MUHAMMAD NASIR CHEEMA and others---Respondents Civil Appeals Nos. 1169 and 1192 of 2014, decided on 11th May, 2015. (Against judgment dated 5-8-2014 of Election Tribunal, Lahore, passed in Election Petition No. 226 of 2013) (a) Representation of the People Act (LXXXV of 1976)--- ---S. 46---Election Tribunal, powers of---Scope---Appointment of a Commission to inspect the record of the certain polling stations---Election Tribunal enjoyed authority to carry out the exercise of examination of polled ballot either itself or through commission and may order | 2016 S C M R 998 [Supreme Court of Pakistan] Present: Mian Saqib Nisar, Mushir Alam and Iqbal Hameedur Rahman, JJ Mian MUHAMMAD ASFAR WALI---Appellant Versus Muhammad NISAR CHEMA and others---Respondents Civil Appeals Nos. 1169 and 1192 of 2014, decided on 16th May, 2016. (On appeal from the judgment dated 05-8-2014 passed by the Election Tribunal, Lahore in Election Petition No.1169 of 2014) (a) Representation of the People Act (LXXXV of 1976)--- ---Ss. 1 & 2---Constituency of PP-97, Gujranwala---Election Petitioner/Respondent (s) Civil Procedure Code (V of 1908), Ss. 2 & 3---By-Election Act (XVIII of 1969), S. 11---Constitution of Pakistan, Art. 20---Eligibility of candidate for election---Scope---Non-seating of candidate by Election Tribunal---Plea for modification of the impugned decision dated 05.08.2014, passed in E.P.C. No.1192-L of 2014: (b) Election Tribunal (Lahore) |

Fig. 1. Comparison of Original Summary vs. generated summary.

## III. FUTURE COURSE OF ACTION

We explored the potential of BART as well as the state-of-the-art deep learning libraries for text summarization of Supreme Court of Pakistan judgements. The proposed model was able to provide a better comprehension and understanding of the judgments and the evaluation score as well as the generated summaries. However, several research directions will be further explored in the future:

- Since this was the first study of its kind, the dataset size was limited and we aim to explore the effect of a larger dataset on the evaluation score as well as using other performance evaluation metrics such as BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation for Translation with Explicit Ordering).
- Although hyper-parameters were fine-tuned to the maximum extent, there were some memory restrictions on a P100 GPU which was used during the course of experimentation. In the future, we want to explore the effect of varying hyper-parameters on a server with high memory specifications.
- ROUGE scores are not necessarily the appropriate criterion for determining the quality of domain-specific summaries. ROUGE counts just n-gram overlaps and does not consider if the sentences adequately reflect the actual text (e.g., rhetorical categories for legal documents) [1]. Human evaluation is always preferred especially for the domain where the research effort is in inception. In collaboration with the Supreme Court of Pakistan, we will involve different stakeholders from the legal domain to better synthesize and evaluate the generated summaries so that it may fulfil the requirements of law practitioners.
- We aim to prepare a template to be used for legal judgement summarization so that the text summaries may be more precise and related to the requirements of legal community.
- We also plan to leverage domain knowledge to enrich the semantic understanding of the proposed model by incorporating much relevant and down-streamed model trained on the legal corpora such as Legal-Bert for a better summary generation.

## REFERENCES


[1] Bhattacharya, P., et al. A comparative study of summarization algorithms applied to legal case judgments. in European Conference on Information Retrieval. 2019.

[2] Ambedkar, K., S. Pal, and R. Pamula, Text Summarization from Legal Documents: A Survey. Artificial Intelligence Review, 2019. 51(3): pp. 371-402.

[3] Polsley, S., P. Jhunjhunwala, and R. Huang. Casesummarizer: a system for automated summarization of legal texts. in COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations. 2016.

[4] Grover, C., et al. Automatic summarisation of legal documents. in 9th International Conference on Artificial Intelligence and Law. 2003.

[5] Howe, J.S.T., L.H. Khang, and I.E. Chai., Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments. arXiv preprint, 2019.

[6] Adelia, R., S. Suyanto, and U.N. Wisesty. Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit. in Procedia Computer Science. 2019.

[7] Arshad S, Latif S, Hasan A, Latif R, Shafait F. Automatic Summarization of Legal Judgements using Deep Learning. Submitted for publication in Computers, Materials & Continua.

# Information Extraction from Legal Documents Based on Natural Language Processing (NLP)

Student's name: Iqra Basharat
Supervisor/s: Dr. Rabia Irfan, Dr. Faisal Shafait
School of Electrical Engineering and Computer Science, NUST
Islamabad, Pakistan
Starting date of the PhD: October 2020
Expected finalization date of the PhD: July 2024
Email: ibasharat.phdcs20seecs@seecs.edu.pk

*Abstract*— **The proposed research study is focused to automate the key elements of judicial system of Pakistan using deep learning and natural language processing techniques. The overarching objective is to cut down the cognitive efforts in legal document analysis and information extraction as well as to explore the key points that must be included in case reviews and hence accelerate the legal processes. Thus, we aim to propose a sole solution as an ontological-based legal knowledge base that involves, construction of domain ontologies from legal corpus as an essential step building the knowledge base. Furthermore, the idea is to extend this into semantic search system that will act as an assistance tool for legal practitioners to have semantically enhanced information and rule-based reasoning.**

**Keywords-- information extraction; legal judgements; legal ontologies**

## I. SHORT RESEARCH SUMMARY

Extracting richer and relevant legal information from a large archive is significantly important for its various stakeholders such as lawyers, scholars, and the public [1]. To address a legal problem, lawyers must look through several past decisions to support their cases. On the other hand, novice users are often interested to know if there is previous evidence of such court judgements. Legal text has its own characteristics such as size, structure, vocabulary, ambiguity, and citations of precedence of similar nature that makes this domain unique in its nature. Due to this reason, already implemented and proposed approaches cannot be applied on native legal data.

This research work is focused on the information extraction of legal documents using ontological-based knowledge base system and semantically enhanced search system for Pakistan's Judicial System. It is worth mentioning that some studies have addressed the information extraction from legal text in the form of text summarization, named entity recognition and catchphrase identification, from UK case judgements [2], Singapore case judgements [3], Indonesian case judgements [4] and Indian case judgements [5] but there is no systematic evidence that a methodology developed for a particular country can be generalized for the legal context of another country [1].

Semantic search for legal domain in Pakistan has not been the area of focus for Pakistani researchers. This proposed project will contribute significantly to the research study of judicial system of Pakistan. Similarly, no prior work has been found on Ontology development for Pakistan's legal landscape. Overall, this research will not only make groundbreaking contributions in the field of applying semantic search based on legal ontologies but will also paved the foundation for high level applications, e.g., knowledge-graph of Pakistani judicial system, verdict assistance tool for honorable judges, etc.

## II. INFORMATION EXTRACTION FROM LEGAL DOCUMENTS OF PAKISTAN

In literature limited study on legal documents of Pakistan was found primarily focusing the development of annotated dataset on civil judgments considering that admissible data is not available and moreover developing a machine learning system for entity extraction from legal judgment cases [6]. To model the named entity recognition (NER) three classifiers, conditional random field (CRF), maximum entropy (MaxEnt), and trigram N tag (TNT), were used to train the model on the dataset of 244 out of 304 tagged with nine named entities criminal miscellaneous judgments of the Lahore High Court, Pakistan. Overall, the sequence modeling NER using three classifiers illustrate encouraging results and can be further extended to extract more NEs. The constraint to this proposed study is that it is applied on only one type of criminal judgments since there are several other types of judgments as well such as criminal appeal, criminal revision, civil petition, etc., on which this research should be applied to assure the stated results. Another study from Pakistan [7] carried out by researchers from Government University of Faisalabad have presented the findings of their studies that extract information and identify all cases related to their relevant case under investigation from the corpora of 99 legal judgments. In their proposed study they have used GATE (General Architecture for Text Engineering) for information extraction and achieved desired results. Despite this, too small corpora could not generalize the findings.

### *A. Information Extraction using Ontologies*

In research studies [8] and [9], authors have presented approaches for information extraction using ontologies from legal documents with least effort. Likewise, authors in [10] have used semantic annotation and query expansion techniques of ontology-based information retrieval

techniques. Ontologies have been demonstrated as an effective method for providing semantically rich information representation. In recent years, this approach has been used for data representation and information extraction. In various research studies [11,12,13] ontological information extraction is being used to improve the user search experience. Generally, ontologies represent a set of concepts of information of a certain challenge [14]. The benefit of semantic modeling is its ability to represent the data association and dependencies between different content. This association and semantic linkage provide ways to innovative searching approaches.

### *B. Key phrase/Catchphrase Extraction Techniques*

Various approaches have been used for the extraction of key phrases and important sentences to present a summary of the legal case documents. In [15] scoring techniques using deep neural networks was proposed, and a key phrase extraction tool (KEA) [16] and RAKE [17] were implemented for key phrase extraction and Xin Jiang and his colleagues [18] have proposed SVM technique. In [19] authors proposed a solution for extraction of catchphrases by providing a concise representation of the core legal issues through candidate phrase generation and creating word vector representation further applying Long- short term memory (LSTM) model on 400 legal documents of Indian Supreme Court case documents. This model yields poor results and the dataset used was too small.

### *C. Catchphrase extraction from legal judgements of Pakistan*

As a steppingstone towards the proposed research, we have carried out catchphrase extraction from 300 legal judgements of Pakistan Supreme court. We have adapted the approach used by authors of [20] to extract the catchphrases. This approach uses the sentence embedding model SIFRank+ combined with ELMO, a pre-trained language model, compared with KeyBERT [21] and Rake [17]. We compared the results of KeyBERT unified with ELMO and Rake and further performed human evaluation on results obtained from these three approaches. Preprocessing methods such as lexical analysis, tokenization and stemming were performed to filter irrelevant words and phrases. The performance of models was then evaluated by calculating accuracy through comparison of generated output with correct output along with the consultation of expert opinion from the legal system. The results presented in table I below shows better accuracy for KeyBERT unified with ELMO deep learning model and table II displays the extracted catchphrases from different models. Nevertheless, the research results can be improved by using large corpus. We have submitted this work in a conference that is under review.

TABLE I. PERFORMANCE EVALUATION OF CATCHPHRASE EXTRACTION MODELS

| CatchPhrase Extraction model | Dataset | Accuracy |
|---|---|---|
| ELMO + KeyBERT | Legal Judgements of Supreme Court | 0.43 |
| RAKE | Legal Judgements of Supreme Court | 0.30 |

TABLE II. CATCHPHRASES EXTRACTED USING DIFFERENT MODELS

| KeyBERT + ELMO (+ SIFRankPlus) | RAKE |
|---|---|
| ('filed review petition terms section 14 ordinance dismissed' 0.71), ('leave granted consider petitioners remedy representation president', 0.60), ('competently filed light whereof merit appeal dismissed mwa', 0.60), ('32 ordinance accepted president 2005 aggrieved respondent challenged',041), ('review federal tax ombudsman sets aside earlier', 0.5556), ('sets aside earlier decision irrespective recommendation', 0.5477), ('federal tax ombudsman passed order recommendation favor respondent',0.539), ('14 ordinance empowers federal tax ombudsman review finding', 0.51), ('ordinance instead appellant filed review', 0.5296), ('division period days date receipt complaint reference motion', 0.51) | [(26.08, 'federal tax ombudsman within 30 days'), (25.0, 'dismissed vide order dated 26'), (25.0, '1481 order mian saqib nisar'), (22.75, 'federal tax ombudsman sets aside'), (17.33, 'pakistan within 30 days'), (16.75, 'federal tax ombudsman passed'), (16.75, 'federal tax ombudsman ordinance'), (16.0, 'particularly section 32 thereof', (16.0, 'note:- perhaps even'), (16. 'following terms: - leave'), |

## III. FUTURE WORK

With this comprehensive literature study, we have concluded that there are several future directions and important aspects that need serious attention to improve and automate the information extraction process of judicial system in Pakistan. Following are some key objectives for future research.

### *A. Development of a Legal Knowledge-Base*

Our aim is to develop a legal knowledge base based on legal domain ontological framework. Building ontologies is an essential step in the development of a knowledge-based system. Therefore, we are aiming to put our efforts in the direction of developing well-founded legal domain ontologies for building rule-based reasoning model. With this, developing a knowledge base would accelerate and ease the information extraction process to assist the legal practitioners. In the best of our knowledge this would be a first ever ontological knowledge-based system for Pakistan's judicial system.

### *B. Development of Semantic Search System*

Another possible research direction can be to develop an intelligent legal semantic search system for Pakistan legal setup. Based on findings of the previously stated research objective we can use state-of-the-art natural language processing and deep learning techniques to build different

components of semantic search system, such as query expansion, searching and indexing etc.

### C. *Prposed Research Work Flow*

The proposed research implementation workflow is illustrated in the figure 1 below.

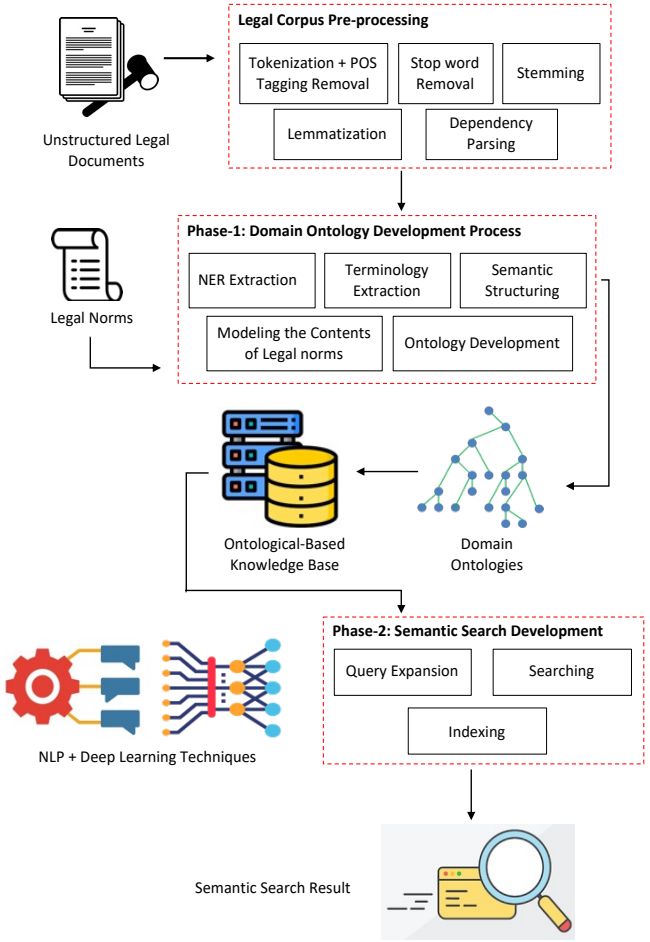



Figure 1. Implementation Workflow of Proposed Research

### D. *Expected Outcomes:*

The major technological outcomes of the proposed research will be as follows:

- Legal ontological knowledge base system from existing legal judgement corpus.
- A semantically enhanced and efficient information search and retrieval system based on state-of-the-art deep learning algorithms.

## REFERENCES


[1] S. Polsley, P. Jhunjhunwala, R. Huang, "Casesummarizer: A system for automated summarization of legal texts," *In Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, 2016, pp. 258-262.

[2] C. Grover, B. Hachey, I. Hughson, C. Korycinski, "Automatic summarisation of legal documents," *In Proceedings of the 9th international conference on Artificial intelligence and law*, 2003, pp. 243-251.

[3] J.S.T. Howe, L.H. Khang, I.E. Chai, "Legal area classification: A comparative study of text classifiers on singapore supreme court judgments,", arXiv preprint arXiv:1904.06470, 2019.

[4] R. Adelia, S. Suyanto, U.N. Wisesty, "Indonesian abstractive text summarization using bidirectional gated recurrent unit," Procedia Computer Science, 2019, 157, 581-588.

[5] M. Saravanan, B. Ravindran, S.Raman, "Improving legal document summarization using graphical model," Frontiers in Artificial Intelligence and Applications, 2016, 152, 51.

[6] A. Iftikhar, S. W. Ul Qounain Jaffry and M. K. Malik, "Information Mining From Criminal Judgments of Lahore High Court," in *IEEE Access*, vol. 7, pp. 59539-59547, 2019

[7] M. R. Talib, M. K. Hanif, Z. Nabi, M.U. Sarwar, N. Ayub, N, "Text mining of judicial system's corpora via clause elements," *International Journal on Information Technologies & Security*, 9(3). 2017.

[8] M.G. Buey, A.L. Garrido, C. Bobed, S. Ilarri, "The AIS Project: Boosting Information Extraction from Legal Documents by using Ontologies," I*n Proceedings of the 8th International Conference on Agents and Artificial Intelligence* ICAART 2016. Volume 2, pages 438-445.

[9] M.G. Buey, C. Roman, A.L. Garrido, C. Bobed, E. Mena, "Automatic legal document analysis: Improving the results of information extraction processes using an ontology." In Intelligent Methods and Big Data in Industrial Applications, Springer, Cham. 2019, pp. 333-351.

[10] I.O. Oyefolahan, E.F. Aminu, M.B. Abdullahi, M.T. Salaudeen, "A review of ontology-based information retrieval techniques on generic domains," Int J Appl Inf Syst, 2018, 12(13), 8-21.

[11] N. Paton, R. Stevens, P. Baker, C. Goble, S. Bechhofer, "A. Brass, Query processing in the tambis bioinformatics source integration system," *In Proceedings of the IEEE International Conference on Scientific and Statistical Databases (SSDBM),* 1999, pp. 138–147;;

[12] A. Meštrovic´ , A. Calì`, "An ontology-based approach to information retrieval," Semantic Keyword-based Search on Structured Data Sources, Springer, 2016, pp. 150–156

[13] F. Ramli, S. Noah, T. Kurniawan, "Ontology-based information retrieval for historical documents," *Third International Conference on Information Retrieval and Knowledge Management (CAMP*), IEEE, 2016, pp. 55–59.

[14] T.R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis. 5 1993 199–220.

[15] Tran, M. Nguyen, K. Satoh, "Automatic catchphrase extraction from legal case documents via scoring using deep neural networks," *In Neural Networks for Natural Language Processing*, Hershey, PA: IGI Global, 2020. pp 143-158.

[16] T.D. Nguyen, M.Y. Kan, "Keyphrase extraction in scientific publications," *In International conference on Asian digital libraries*, Springer, Berlin, Heidelberg. 2017 pp. 317-326.

[17] A. Bhat, C. Satish, N. D'Souza, N. Kashyap, "Effect of Dynamic Stop List on Keyword Prediction Using Rake. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2018. 4(6).

[18] A. Mandal, K. Ghosh, A. Bhattacharya, A. Pal, S. Ghosh, "Overview of the FIRE 2017 track: Information Retrieval from Legal Documents (IRLeD)," *In Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation*, 2017.

[19] R. Bhargava, S. Nigwekar, and Y. Sharma, "Catchphrase Extraction from Legal Documents Using LSTM Networks," *In FIRE* 2017 *Bangalore, India*, pp. 72-73, 2017

[20] Y. Sun, H. Qiu, Y. Zheng, Z. Wang and C. Zhang, "SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model," in IEEE Access, vol. 8, pp. 10896-10906, 2020.

[21] M. Grootendorst, "KeyBERT: Minimal keyword extraction with BERT," Published by Zenodo, v0.1.3 2020.

## I. Informations

- **Title of the thesis**: Weakly-Supervised Scene Text Detection
- **Student's name**: Mengbiao Zhao
- **Affiliation**: National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences
- **Supervisor of the thesis**: Cheng-Lin Liu
- **Starting date of the PhD**: September 2019
- **Expected end date of the PhD**: June 2023

## II. Research goals

Scene text detection has received increasing attention in recent year due to its wide applications in document analysis and scene understanding. Benefited from deep neural networks, many previous methods have made great progress. However, deep learning based methods usually require large scale strongly annotated data during training. And acquiring such a strongly annotated dataset is expensive and time consuming, which has impeded its application in large scale real problems. In addition, existing works on scene text detection often rely on a single form of annotation, however, real-world annotations are often diverse in form, which challenges these existing works.

Therefore, our research goal is to alleviate the excessive dependence of scene text detection algorithm on strongly-annotated data and reduce the cost of data annotation. Specifically, it includes the following contents:

- To propose a detection framework that can utilize data with different annotation formats for training.
- To propose a series of simple and easily labeled weak supervision forms, which can be regarded as the alternative of the strong supervision.
- To propose a weakly-supervised learning algorithm, which can utilize the weak labels to boost the performance of text detectors.
- To propose a arbitrary-shaped text detector, which can facilitate the incorporation of weak supervisions.

## III. Research plans

In this section, we propose a series of research plan to the problems mentioned above.

### A. Joint Training of Data with Multiple Annotation Formats

Nowadays, there are two mainstream annotation formats for scene text datasets: word-level and line-level annotations. And accordingly, the existing methods were designed to detect either words or text lines. The difference between the two annotation formats lies in the definition of the text box. Word-level annotation gives the bounding box of each word in the image, while line-level annotation provides the bounding box of each text line (See Fig. 1). The two types of annotated data were used to train either word detector or text line detector, which ignores the reciprocity between them.


(a) (b)

Figure 1: Visualization of two different annotation formats. (a) Word-level annotation gives the bounding box of each word. (b) Line-level annotation gives the bounding box of a text line.

Word-level and line-level text detection are closely related, and thus can benefit each other: (1) The text line detector can be used to filter out non-text regions in images. After filtering out the background, the word detector can pay more attention to distinguish words. (2) One of the difficulties of text line detection is viewing the background between words as the positive sample. By using word detector to provide the prior position of each word, the line-level detection degenerates into a simpler task: integrating the adjacent words into text line. Therefore, both the two tasks could be improved by the mutual guidance.

Therefore, we could propose a novel framework to perform the word-level and line-level text detection simultaneously (See Fig. 2). Firstly, for the joint training of two types of annotated data, we design a dual-task network with two detection heads for word-level and line-level detection, respectively. Considering the similarity between the two tasks, the two detectors share one backbone network for feature extraction. During training, the backbone network extracts features and then sends them to the corresponding detection heads. Secondly, to take advantage of the complementary information between the two kinds of data, we propose a mutual guidance scheme, which is implemented by two modules. Specifically, based on the feature map extracted by the text line detector, a line filtering module is proposed to filter out the non-text regions for the word detector. Meanwhile, according to the output feature map of the word detector, a word enhancing module is proposed to provide word prior to the text line detector. Experimental results on CTW1500 [1] are shown in Table I, respectively, which demonstrate the effectiveness of the proposed dual-task network and mutual guidance scheme.

This work had been published in *IEEE ICPR 2020*, and won the "**Best Scientific Paper Award**".

*M. Zhao, W. Feng, F. Yin, X.-Y. Zhang and C.-L. Liu, Mutually Guided Dual-Task Network for Scene Text Detection, in IEEE International Conference on Pattern Recognition (ICPR), 2020.*

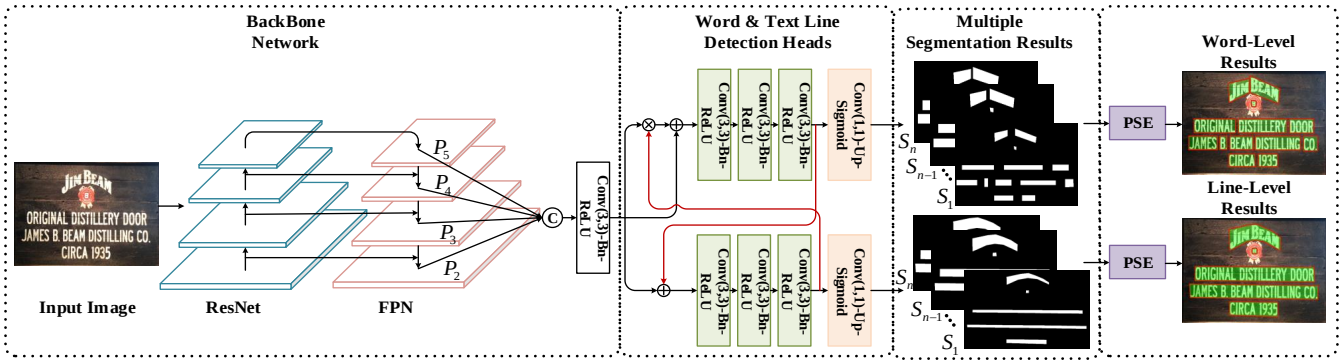



Figure 2: The pipeline of the proposed framework. The backbone network is ResNet with FPN. The two detectors have the same structure, and they can guide each other which is shown by the red lines in the figure. The output results of the two detectors are multiple segmentation masks. In the inference stage, the progressive scale expansion (PSE) algorithm is executed to get the final predictions of text instances.

Table I: Effects of joint learning and mutual guidance on CTW1500. 'P', 'R' and 'F' represent the precision, recall and F-measure, respectively.

| Method | P | R | F |
|---|---|---|---|
| Baseline [3] | 80.6 | 75.6 | 78.0 |
| Baseline+joint [3] | 74.87 | 71.96 | 73.39 |
| Dual-task | 79.68 | 76.17 | 77.89 |
| Dual-task+guidance | **81.48** | **78.39** | **79.92** |

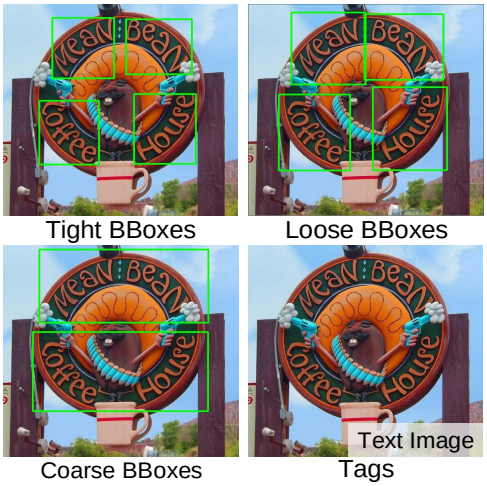



Figure 3: Examples of four weak supervision forms.

### B. Weak Supervision Forms of Arbitrary-Shaped Scene Text

As analyzed above, labeling scene text with polygons is laborious and tedious. Therefore, we propose four forms of weak supervision (See Fig. 3) for scene texts as alternatives from the perspective of annotators. The tight bounding box is defined as the rectangle enclosing the text instance tightly, so the annotator needs to find the four extreme points of text contour, which is still time-consuming. In order to speed up annotation, we then propose the loose bounding box, which could be broader than the tight bounding box. Later, based on the observation of the text distribution in the scene, we further simplify the annotation process, using coarse bounding box to roughly locate the position of a cluster of texts (multiple instances). However, when the number of images is very large, the cost of any box-level annotation is also very large, and the image-level tag becomes the easiest choice. With the decrease of annotation complexity, the time costs of labeling with different annotation policies are gradually decreasing.

Public datasets of arbitrary-shaped text detection are mostly annotated with polygon-level labels, from which the proposed four forms of weak labels can be generated. The tight bounding box can be obtained from the external bounding box of the polygon. The loose bounding box can be obtained by expanding the height and width of the tight bounding box by 0:1 to 0:2 times respectively. The coarse bounding box can be generated from the external bounding box of the text cluster. We adopt Mean Shift algorithm to cluster the centers of text regions, where the clustering radius is set to 0.3 times of the short side length of the image. The image-level tag indicates whether an image contains text or not, which can be easily obtained.

### C. Arbitrary-Shaped Text Detector

Inspired by [4], we propose a contour based arbitrary-shaped text detector. An overview of the detector is illustrated in Fig. 4. After extracting original features by the backbone network, a text localization network is used to generate bounding-box-level text proposals. Then, we adopt a contour initialization network to produce the initial text contour for each text proposal. Finally, initial text contours

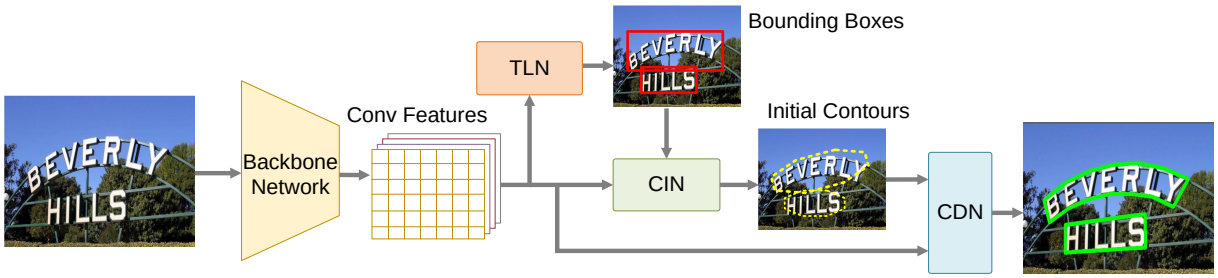



Figure 4: Illustration of the structure of the proposed text detector."TLN", "CIN", and "CDN" represent text localization network, contour initialization network and contour deformation network, respectively.

together with original features are sent to the contour deformation network, which performs iterative contour regression to obtain the text instance boundary. Since the proposal mechanism in the pipeline is very similar to the proposed weak supervision forms (*coarse*, *loose*, and *tight bounding boxes*), the proposed method can fully utilize the weak labels to boost the detection performance.

**Text Localization Network.** We adopt CenterNet [5] to generate text proposals, which reformulates the detection task as a keypoint detection problem. The detection head has two branches: (1) The classification branch calculates a heatmap, where the peaks are supposed to be the text instance centers; (2) The regression branch predicts the height and width of the proposal bounding box for each peak.

**Contour Initialization Network.** Since the initial input has a non-ignorable influence on the contour deformation and the detected text proposals usually have some offsets or errors, we propose this network to produce more accurate and suitable initial contours for text instances. In [6] and [4], octagon enclose the arbitrary-shaped object much tighter than the rectangle. Therefore, we also choose it as the initial contour. In fact, the octagon could be formed by four extreme points, which are top, leftmost, bottom, rightmost pixels in an object, denoted by $\{z_i^{ex}|i=1,2,3,4\}$. Therefore, the problem is how to get the extreme points from the bounding box.

Given a bounding box, we could extract the four center points at the top, left, bottom, right box edges, denoted by $\{z_i^{bb}|i=1,2,3,4\}$, and then connect them to get a diamond contour. After that, we adopt the Deep Snake, a contour regression model proposed by [4]. It takes the diamond contour as input and outputs four offsets that point from each $z_i^{bb}$ to the extreme point $z_i^{ex}$, namely $z_i^{ex}-z_i^{bb}$. Finally, we extend a line in both directions at each extreme point, whose length is 1/4 to the corresponding edge, and connect their endpoints to get the octagon.

**Contour Deformation Network.** Contour deformation network is used to regress the offsets from points on the initial contour to the corresponding points on the ground-truth. We adopt the same regression method as the contour

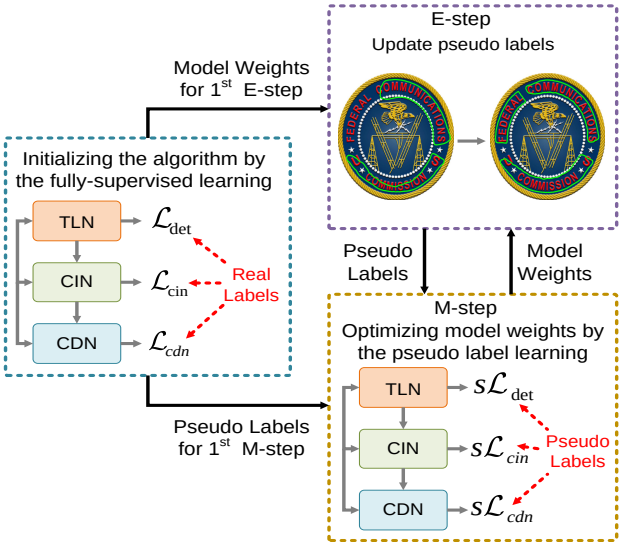



Figure 5: The pipeline of the EM-based learning algorithm. The algorithm is firstly initialized by the model trained with a small amount of strongly annotated data (The blue box part). Then the algorithm alternates between updating the pseudo labels (The purple box part) and optimizing the model weights with the pseudo label learning, where the confidence weighted loss is adopted (The yellow box part).

initialization. To make the output contour smoother, we sample 128 points along the octagon contour. Similarly, the ground-truth is generated by uniformly sampling 128 vertices along the polygon. In addition, in order to simplify the difficulty of regression, an iterative optimization strategy is adopted. Specifically, the output contour of the previous iteration is used as the initial contour of the next iteration.

### *D. Weakly-Supervised Learning Algorithm*

Specifically, let $x$ denotes the image values, $y$ denotes the polygon-level labels of the image, and $t$ denotes the weak labels of the image. As for the weakly annotated image, we can observe the image values $x$ and the weak labels $t$, but the text instances polygons are latent variable. We have the

following probabilistic graphical model:

$$P(x, y, t; \theta) = P(x)P(y|x; \theta)P(t|y). \quad (1)$$

Then, in order to learn the model parameters $\theta$ from the weakly annotated data, we can adopt an EM-based learning strategy as follows:

**E-step.** The purpose of E-step is to estimate the complete-data log likelihood. Given the previously estimated parameter $\theta^{'}$, the expected complete-data log likelihood for weakly annotated image $x$ and it's label $t$ is given by

$$\begin{aligned} Q(\theta; \theta^{'}) &= \sum_{y} P(y|x, t; \theta^{'}) \log P(y|x; \theta) \\ &\approx \log P(\hat{y}|x; \theta), \end{aligned} \quad (2)$$

where we adopt a hard-EM approximation, estimating the latent variable by

$$\hat{y} = \arg\max_{y} P(y|x, t; \theta^{'}). \quad (3)$$

**M-step.** The M-step is to maximize the $Q(\theta; \theta^{'})$ with respect to $\theta$. According to Eq. 2, the key to maximize $Q(\theta; \theta^{'})$ is maximizing $\log P(\hat{y}|x; \theta)$. Here, we treat $\hat{y}$ as ground truth polygons, and optimize $\log P(\hat{y}|x; \theta)$ by the mini-batch SGD algorithm.

We integrate the detector in Sec. III-B into the learning algorithm, and obtain a pipeline for weakly-supervised text detection, which is shown in Fig. 5. The parameter $\theta$ is equivalent to the weights of the detection model. We use a small amount of strongly annotated data to train a model, which provide the initial state for the 1st M-step. And the estimated latent polygon-level label $\hat{y}$ is given by the output of contour deformation network in the detection model. As shown in Eq. 3, $\hat{y}$ is related to weak labels $z$. Different weak supervisions will provide different information, so there are different approaches to estimate the latent variables.

Experimental results on Total-Text [7] and CTW1500 [1] are shown in Table II and III, which demonstrate that all the weakly-supervised models have outperform the baseline model, and using only 10% strongly annotated data our method yields comparable performance to state-of-the-art methods.

This work had been submitted to *IEEE Transactions on Image Processing*.

*M. Zhao, W. Feng, F. Yin, X.-Y. Zhang and C.-L. Liu, Mixed-Supervised Scene Text Detection with Expectation-Maximization Algorithm, IEEE Transactions on Image Processing, Under review.*

Table II: Detection results on Total-Text. "P", "R", and "F" represent the precision, recall and F-measure, respectively.

| Method | P | R | F | FPS |
|---|---|---|---|---|
| TextField [8] | 81.2 | 79.9 | 80.6 | - |
| PSENet-1s [3] | 84.0 | 78.0 | 80.9 | 3.9 |
| SPCNet [9] | 83.0 | 82.8 | 82.9 | - |
| CRAFT [10] | 87.6 | 79.9 | 83.6 | - |
| DB-ResNet-50 [11] | 87.1 | 82.5 | 84.7 | 32.0 |
| PAN-640 [12] | **89.3** | 81.1 | 85.0 | **39.6** |
| ContourNet [13] | 86.9 | **83.9** | 85.4 | 3.8 |
| 100%Poly | 88.2 | 83.3 | **85.6** | 24.2 |
| 10%Poly & 90%Tight | 85.4 | 83.8 | 84.6 | |
| 10%Poly & 90%Loose | 86.6 | 82.1 | 84.3 | |
| 10%Poly & 90%Coarse | 84.7 | 79.6 | 82.0 | |
| 10%Poly & 90%Tag | 82.9 | 78.8 | 80.8 | |
| 10%Poly | 80.2 | 78.5 | 79.4 | |

Table III: Detection results on CTW1500. "P", "R", and "F" represent the precision, recall and F-measure, respectively.

| Method | P | R | F | FPS |
|---|---|---|---|---|
| ABCNet [14] | 83.8 | 79.1 | 81.4 | 9.5 |
| PSENet-1s [3] | 84.8 | 79.7 | 82.2 | 3.9 |
| DB-ResNet-50 [11] | 86.9 | 80.2 | 83.4 | 22.0 |
| PAN-640 [12] | 86.4 | 81.2 | 83.7 | **39.8** |
| ContourNet [13] | 83.7 | **84.1** | 83.9 | 4.5 |
| 100%Poly | **87.0** | 81.8 | **84.3** | 32.3 |
| 10%Poly & 90%Tight | 86.4 | 81.1 | 83.7 | |
| 10%Poly & 90%Loose | 86.3 | 80.1 | 83.1 | |
| 10%Poly & 90%Coarse | 84.0 | 80.6 | 82.3 | |
| 10%Poly & 90%Tag | 83.4 | 79.1 | 81.2 | |
| 10%Poly | 81.3 | 79.1 | 80.2 | |

## IV. Remaining Problems

### A. About The Unified Scene Text Detection Framework

In the real world, the annotations are often diverse in forms. Incorporating strong supervisions (*e.g.*, polygons) and various forms of partial supervision (*e.g.*, boxes, scribbles, points, and class tags) to perform omni-supervised learning is particularly beneficial in practice. However, it still remains a challenge problem to develop a unified omni-supervised scene text detection framework to simultaneously handle direct supervision and various forms of partial supervision.

### B. About The Weakly-Supervised Text Spotting

As text recognition is commonly the uppermost goal for the detection, it would be meaningful to study the weakly-supervised text spotting. Sun *et al.* proposed a large dataset, where each image is only annotated with one dominant text. They also design an algorithm to utilize these partially labeled data. However, because the supervision information is too weak, it will not greatly improve the performance. For the text spotting, how to define a more reasonable weak supervision form and a more efficient algorithm need to be further studied.

## REFERENCES


[1] Y. Liu, L. Jin, S. Zhang, and S. Zhang, "Detecting curve text in the wild: New dataset and new solution," in *arXiv preprint arXiv:1712.02170*, 2017. 1, 4

[2] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *International Conference on Document Analysis and Recognition*, 2015.

[3] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4

[4] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[5] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019. 3

[6] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[7] C. S. C. Chee Kheng Chng, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2018. 4

[8] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "Textfield: Learning a deep direction field for irregular scene text detection," *IEEE Transactions on Image Processing*, vol. 28, no. 11, p. 55665579, Nov 2019. 4

[9] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2019. 4

[10] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[11] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proceedings of AAAI Conference on Artificial Intelligence*, 2020. 4

[12] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 4

[13] Y. Wang, H. Xie, Z.-J. Zha, M. Xing, Z. Fu, and Y. Zhang, "Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[14] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

# Optical Handwritten Named Entity Recognition

Thomas CONSTUM
*LITIS (University of Rouen Normandy)*
*Saint-Étienne-du-Rouvray, France*
*Advisor : Pr. Thierry PAQUET*

## I. Introduction

This document presents my research plan for the doctoral consortium of the ICDAR 2021. In 2020, I graduated from the INSA Rouen Normandy and obtained a master's degree of IT with a major in data science. I am currently a research engineer for the LARHRA (CNRS) and the LITIS within the project POPP (Project for the Oceration of the Parisian Population) and I will do a thesis at the LITIS from October 2021 to September 2024.

## II. Presentation of the thesis topic

During my thesis, it is planned to study in depth the question of Optical Handwritten Named Entity Recognition (OH-NER). While named entities represent a very large diversity of semantic information that can occur in a specific language and in a specific domain (such as health, law, cooking, etc.), we plan to focus our attention on a specific corpus to make our experimentation. It is sufficiently specific to constrain the diversity of the possible named entities, while representative of a real task of Named Entities Extraction in handwritten documents. This corpus is made of the handwritten marriage certificates of Paris and its suburb from 1880 to 1922. It is already digitized, and studied by historians and demographers, notably by Dr. Sandra Brée (CR-CNRS-LARHRA) who is an historian and demographer, specialist in historical demography, especially with large quantitative databases. We are currently collaborating with her within the POPP project. The marriage certificates contain about 30 different types of named entities to be extracted and recognized, and we will count on the collaboration with Dr. Sandra Brée and her team of the LARHRA to get a sufficient amount of annotated data for our experimentations. Figure 1 shows one example of a marriage certificate and highlight the complexity of the task at end. All along the project, we expect to process about 150 000 marriage certificates and advance the state of the art both from the methodological point of view and the experimental setup. As a side effect of this study, a unique and very large dataset of handwritten material for Named Entities Recognition will be built.

Depending on the progress of the thesis, it is also planned to address writer adaptation and handwritten Named Entity Disambiguation. In the next section, I present these 3 topics and the associated research plan.

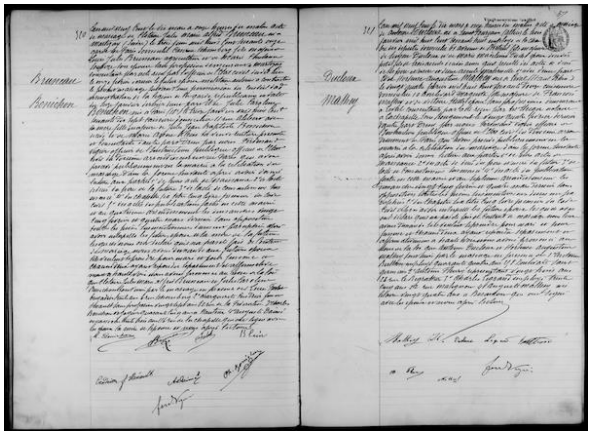


Figure 1. A marriage certificate of Paris (1880-1922)

## III. Research plan

### A. Optical Handwritten Named Entity Extraction (OH-NER)

The main topic of my thesis is to explore in depth new strategies and Deep Neural Network architectures for end to end OH-NER. Handwritten Named Entities are very difficult to recognize because in general they pertain to very large lexicons, sometimes infinite lexicons, which make them very hard to disambiguate during the recognition. The use of character n-gram language models is generally the default solution introduced on top of the optical model in the aim to disambiguate the recognition hypothesis using a language driven decoding stage (a.k.a. Viterbi beam search decoding). Named Entity Recognition is generally performed sequentially after the recognition phase, on the recognized sequences of characters. This is the strategy adopted on printed text [1], [2]. This may be the default strategy for Handwritten Named Entity Recognition as well, provided handwriting recognition performs sufficiently well, for example on a single writer corpus such as the Esposalles dataset [3]. End to end approaches have been proposed recently for handwritten Named Entities Recognition [4]. But these approaches do not exploit external resources to disambiguate the recognition process. In the context of the POPP project, we have been able to use external resources such as lists of family names, lists of surnames, locations, list of jobs to disambiguate the recognition of each Handwritten

entity in a table. These lists come from the French mortality database (INSEE). For some entities such as date of birth, regular expressions are used to constrain the recognition. Lists of words, or regular expressions are compiled in a weighted Finite State Transducer (WFST) and used during a Viterbi beam search decoding process. This is possible because a first segmentation process locates each entity by exploiting the known structure of the table, prior to launching the appropriate decoder (recognizer) on each cell of the table.

We plan to explore a similar approach here, but with a higher level of complexity. The key issue lies in the possibility to locate each Named Entity in the text, before launching a specific decoder for each of them, that would embed the appropriate knowledge (encoded into a specific WFST). In this respect, we will explore a shallow Named Entity Recognizer, that will be devoted to labelling the text line images with the different Named Entity categories. Inspired by [4], and a deep CNN based recognizer architecture developed at LITIS [5] we will develop an End-to-End architecture for Named Entity localization. The architecture will also serve for the recognition stage as well, by providing the optical representation of the image to the decoder. The shallow Named Entity Recognizer approach is expected to perform well, as the marriage certificates are written with very regular linguistic patterns that should ease the localization process.

### *B. Writing Adaptation*

In order to improve the recognition performance, another direction of research will be the writer adaptation of the recognition system. It is a well-known property that recognizing a single known writer is easier than recognizing any, unknown, writers. The corpus of marriage certificates is typical of this mono-writer context, as the marriage certificates were written by the same person in each administrative service, during several years. Key issues are to detect the different handwritings in the corpus then adapt the recognizer to each new detected handwriting. We assume that each page is written by one single writer and plan to explore unsupervised learning for handwriting detection. Each page contains enough information to learn an embedding representation of the writer. Then, a decision function can be trained based on the triplet loss for example. The writer adaptation task is to specialize a general recognizer (omni-writer recognizer) to one writer, providing sufficient annotated data of that writer are available. It is planned to explore more data of each writer for the specialization through semi-supervised learning.

### *C. Learning Handwritten Named Entity Disambiguation*

Whereas handwritten word disambiguation is at the core of the recognition system that will be developed, the last direction of research will be to study Named Entity Disambiguation (NED) [6]. The NED is the task of assigning a unique identity to entities mentioned in a collection of texts. It is intended to provide augmented information to a reader by providing links to external resources that provide additional validated information about a famous person, a place, a company, etc. This subject has not been studied yet regarding handwritten named entities, above all, the question on how to validate the recognition results by the exploitation of contextual information has not been studied. Since each marriage certificate will encode tuples of Named Entities about the spouses, the NED could be applied to perform reference checking.

Indeed, even if we expect good recognition performance of the recognition module, these tuples will be prone to recognition errors. Errors on marriage certificates occur each time one single Named Entity is erroneous. Thus, the recognition performance may decrease dramatically if we focus on error-free recognized marriage certificates. This is why reference checking is needed, in order to validate as much as possible, the Named Entities extracted. Unfortunately, we cannot expect that general external resources such as wikidata can bring relevant knowledge to help the disambiguation process (which is generally the proposed approach in the literature), because this can only serve in the case of known people which are referenced in historical records, books, etc... Fortunately, the INSEE French mortality database[1] which contains the list of deaths in France since 1970, is the appropriate external knowledge resource for our purpose. It gives, for the persons who died since 1970, the name, surname, sex, date of birth, place of birth, date of death and place of death. This resource will allow to detect the inconsistencies in the extraction process by checking the extracted tuples of information with those of the INSEE French mortality database. A second resource is the POPP dataset that is currently under construction by LITIS within the POPP project.

From these resources we expect to build a marriage dataset where each entry will contain the key information that uniquely identifies a marriage. Notice that a marriage certificate can be uniquely identified by the names of spouses, their date of birth, place of birth, and date of marriage. A first vanilla approach for reference checking is to look for exact matches between the resource and the tuple extracted. However, in our context, we are prone to a significant number of errors in the extracted tuples, and we need a more elaborated strategy for checking. The key information will serve as discriminative features, possibly augmented by the other extracted information. By exploiting these inputs, we plan to learn a latent representation on which a decision can be made about the similarity of the two inputs to be paired.

[1] https://www.insee.fr/fr/information/4769950

## IV. Novel research ideas

As my thesis has not yet started, I have not been able to conduct actual experiments. However, thanks to the discussions with Prof. Oriol Ramos Terrades, my mentor for this doctoral consortium, I have a more precise idea of the first experiments that should be made regarding the Named Entity Recognition (NER) architecture.

First, it has been shown that using contextual information improve the performance of NER [7]. Therefore, working separately on each word or line could remove information that would have been useful for NER. Thus, one first step to keep this context could be to use the attention module from [5] to implicitly segment the marriage certificates into lines using an attention module. Then, the obtained line features maps would be concatenated into a unique feature map for the whole text. A similar approach was used in [8], where bounding boxes of located words were concatenated in reading order to predict the semantic tags as a sequence. An alternative could be to use an architecture such as [9] that directly learns to unfold an input multi-line image into a single line image.

In certain cases, the interdependence between the task of handwriting recognition and NER makes it beneficial to learn both tasks jointly [8]. Thereby, we could in addition to a branch dedicated to NER, include another branch for handwriting recognition. Moreover, another solution could be to develop a multi-task decoder that would perform both tasks at the same time by including named entity tags in the charset as special characters [10].

These decoders would initially be based on BLSTMs, but it would be interesting to replace them with transformers [11]. The transformer is an attention-based architecture used for learning a representation of sequences that is faster to train and less resource consuming than BLSTMs. This architecture was first developed for machine translation and was then adapted to performs NER on strings [12] since the vanilla version does not reach the state of the art for this task. However, the adaptation of the transformer architecture for NER on handwritten documents has not been studied yet and could thus be an interesting research axis. A first idea would be to simply replace the character embedding of string from [12] with the image encoder used in [5]. However, a transformer-based encoder will also be studied since architectures using transformers have shown promising results for visual recognition [13].

## References


[1] W. Swaileh, T. Paquet, S. Adam, and A. Rojas Camacho, "A named entity extraction system for historical financial data," in *Document Analysis Systems*, X. Bai, D. Karatzas, and D. Lopresti, Eds. Cham: Springer International Publishing, 2020, pp. 324–340.

[2] C. Neudecker, "An open corpus for named entity recognition in historic newspapers," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4348–4352. [Online]. Available: https://www.aclweb.org/anthology/L16-1689

[3] V. Romero, A. Fornés, N. Serrano, J.-A. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós, "The esposalles database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognition*, vol. 46, 06 2013.

[4] C. Wigington, B. Price, and S. Cohen, "Multi-label connectionist temporal classification," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 979–986.

[5] D. Coquenet, C. Chatelain, and T. Paquet, "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network," *arXiv:2012.03868 [cs]*, Dec. 2020, arXiv: 2012.03868. [Online]. Available: http://arxiv.org/abs/2012.03868

[6] C. Brando, F. Frontini, and J.-G. Ganascia, "REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets," *Complex Systems Informatics and Modeling Quarterly*, no. 7, pp. 60 – 80, Jul. 2016. [Online]. Available: https://hal.sorbonne-universite.fr/hal-01396037

[7] J. I. Toledo, M. Carbonell, A. Fornés, and J. Lladós, "Information extraction from historical handwritten document images with a context-aware neural model," *Pattern Recognition*, vol. 86, pp. 27–36, 2019.

[8] M. Carbonell, A. Fornés, M. Villegas, and J. Lladós, "A neural model for text localization, transcription and named entity recognition in full pages," 2020.

[9] M. Yousef and T. E. Bishop, "Origaminet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold," *CoRR*, vol. abs/2006.07491, 2020. [Online]. Available: https://arxiv.org/abs/2006.07491

[10] E. Boros, V. Romero, M. Maarand, D. Stutzmann, K. Zenklová, J. Křečková, E. Vidal, and C. Kermorvant, "A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters," in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, ser. 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). Dortmund, Germany: IEEE, Sep. 2020, pp. 79–84. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02935087

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[12] H. Yan, B. Deng, X. Li, and X. Qiu, "Tener: Adapting transformer encoder for named entity recognition," 2019.

[13] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," 2021.

# Multivalent Graph Matching and Ant Colony Optimization for Pattern Recognition

Student's name: HO Kieu Diem
Supervisors: Prof. Jean-Yves RAMEL and Assoc. Prof. Nicolas MONMARCHE
University: University of Tours, France
Starting date: 01/11/2018
Expected finalization date: 15/01/2022
Email: kieu.ho@univ-tours.fr

***Abstract*—The use of a graph to represent the semantic content of images is becoming more and more frequent due to its strong power of representation and interpretability. Locating, recognizing, or interpreting content then turns into comparing or matching graphs. The exact Graph Matching (GM) methods are not tolerant to noise and they follow strict constraints. Meanwhile, the inexact ones relax these constraints so they permit good management in noisy contexts such as handling with over-segmentation. Due to that, my Ph.D. thesis concentrates on inexact GM, the multivalent GM in particular. This kind of GM is very general and flexible in various contexts because one node in one graph can match more than one node in another. As a consequence, this problem is very combinatorial and complex. We formalize this problem as an Extended Graph Edit Distance (ExGED) by adding possibilities of split and merge operations in addition to the classical ones (substitution, insertion, and deletion). Then, the Ant Colony Optimization (ACO) is applied to solve the ExGED problem due to its effectiveness for combinatorial problems such as the traveling salesman problem. Currently, in my Ph.D. thesis, we illustrate the feasibility of the proposed approach for a set of pre-segmented symbol images in graphic document images where these images are noised in various ways. For the first step, the numerical results show the interest in using multivalent GM to recognize noisy symbols. Also, the results provide meaningful interpretability about different elements constituting the images besides the similarity measures between symbols. For the second step, we plan to clarify some theoretical aspects related to ExGED. Then, we would like to apply this approach to compare human brain connectomes where the graphs represent the functional or structural brain networks. Specifically, we intend to detect the differences or modified patterns between the brains of patients, Alzheimer's malady, for instance, and brains of healthy control people. In addition, we would like to apply the proposed algorithm on more dataset to have more reliable evaluations.**

***Keywords*-multivalent graph matching; graph edit distance; ant colony optimization; symbol recognition; brain connectome comparison;**

## I. Short research plan

### A. Introduction

Graph matching is more and more applied in the computer vision field because of its advantages: 1) providing the sub-part correspondence and 2) providing a general similarity measure [1]. Particularly, it witnesses success in some domains like 2D and 3D image analysis [2]. Therefore, it is pretty promising to apply to other fields.

Graph matching can be exact or inexact depending on the application. However, the inexact GM is more flexible because it relaxes the strict constraints in exact GM to permit more matching possibilities. In this category, we focus on multivalent GM. The reason is that multivalent GM is more general than the other inexact GM since it allows one node in one graph can be associated with more than one node in another. So, it is more tolerant to noise and more adaptable in real context [3]. For example, in over-segmented image segmentation, multiple regions can correspond to one region in the original image. So, in my thesis, I attempt to contribute the solution to solve this problem.

Dealing with such combinatorial problems as multivalent GM, ones usually formalize it as an ExGED problem. It is done by adding splitting and merging operations besides substitution, insertion, and deletion [4], [5]. However, in existing works, the authors have not been clarified the costs for splitting and merging operations and the edge operations when having splitting and merging. Also, ExGED's properties have not been discovered or clarify. On the other hand, some ACO-based methods have been applied for multivalent GM [2]. However, the number of works is limited, and it cannot deal with various kinds of attributes.

The basic idea of our research project is to bring the multivalent GM and the ACO together for the task of pattern recognition. More precisely, we focused on the feasibility of this combination. Then, we attempt to make clear the properties of the employed method. Moreover, we expect that this combination can help us to explore the large-size graphs. So that we can apply them to various domains. Specifically, we hope that it will be useful for studying the human brain connectome because the graph theory-based recent methods only provide a general similarity but not the difference inside brain regions.

From what has been discussed, we identify these main lines of research that we hope to pursue in the recent project:

1) Demonstrating the feasibility of applying ACO for multivalent GM problem
2) Enhancing the approach to work on big graphs with node and edge labels
3) Clarifying the theoretical aspects of multivalent GM

formalized as an ExGED
4) Attempting to apply the proposed method to human brain connectome comparison.
5) Applying the method to other dataset to have more reliable evaluation.

### B. Completed Work

At the beginning of the thesis, we try to demonstrate the feasibility when applies ACO to solve the multivalent GM problem. However, due to the lack of Benchmark datasets for this kind of GM and a limited number of related works in this domain, we have generated synthetic data from a symbol one called SESYD. That is done by adding noise to the models to make the noisy symbols such as deformation, rotation, or scale the model [1]. Then, a graph representation is done for each symbol as an undirected attributed graph. Nodes are lines, and edges are spatial relations between nodes. Node's labels are its length, edge's labels are angles between lines and relation types (parallel, successive). Then, costs for node and edge edit operations are defined to establish the cost matrix for the ExGED. This cost matrix plays as a heuristic factor during the search process for ACO to find good matchings. Specifically, the Max-Min Ant System (MMAS), a variant of ACO, is implemented. At the end of the process, the ants give the best mapping. Note that this mapping can be optimal or not because ACO is a meta-heuristic approach. During the process of building matching, a local search is applied to enhance the solution quality. In the context of symbol recognition, we analyze the numerical results in three main aspects: 1) tuning parameters of MMAS to obtain a suitable parameter set for the dataset; 2) analyzing mapping quality by considering the sub-part correspondences between symbols; 2) evaluating mapping quantity by considering the matching result as a distance for classifying symbols. The preliminary results are promising, and they are presented in a workshop paper in SSPR 2021 [6].

Nevertheless, after that, we found that computational time is still a challenge. So, we attempt to accelerate it by reducing the search space of ants based on the neighborhood search strategy. The numerical results show the improvement of computational time while it maintains a good solution quality. This work will be presented in the GLESDO workshop, a part of the ICDAR 2021 conference [2].

### C. Future Work

After demonstrating the feasibility of the proposed method, we intend to do the two following main tasks: 1) clarifying the theory aspect of ExGED; 2) applying the proposed approach to human brain connectome comparison; 3) apply the proposed method on other datasets.

[1]link: http://www.rfai.lifat.univ-tours.fr/PublicData/ExGED/home.html
[2]Glesdo: https://www.glesdo-icdar2021.ml/home

For the first task, we would like to investigate the metric properties of ExGED. That is because ExGED relies on GED which is a distance. But, with extended operations of splitting and merging, is ExGED a distance? Does that mean ExGED will satisfy the non-negativity, identity, and triangle inequality properties or not? Additionally, when having splitting and merging operations, how to define the implied edge relations is a question. Thus, we hope that we can answer these inquiries.

For the second task, we intend to apply the proposed method to the brain connectomic field. Through our literature review, the application of GM for a comparison of brain connectome is increasing recently. Because the GM techniques provide both the similarity measure and the explanation between brain regions. As a result, the researchers take these benefits to explore more deeply the brain connectomes, especially for brain networks of patients for establishing a subject-specific bio-marker [7]. Nevertheless, these studies often compare the human connectomes at the same scales and atlases, meaning all the graphs having the same size. That is not pretty natural because of our diversity. Moreover, the GM technique is usually GED, and the GM solver is frequently the Hungarian method. These are only available for one-to-one mapping. In contrast, our approach allows more matching possibilities leading to more adaptability to the real context. Also, we expect that the ACO can be effective for exploring large graphs like the human connectomes.

For the third task, we would like to apply the proposed method to another dataset like the letter dataset (suggestions from the mentor) to have more comparisons. Besides, this dataset has a higher level of distortion. Therefore, testing the algorithm on this dataset can help us evaluate its performance and resistance to noise.

## REFERENCES


[1] D. Conte, P. Foggia, C. Sansone, and M. Vento, "Thirty years of graph matching in pattern recognition," *International journal of pattern recognition and artificial intelligence*, vol. 18, no. 03, pp. 265–298, 2004.

[2] O. Sammoud, C. Solnon, and K. Ghédira, "Ant algorithm for the graph matching problem," in *European Conference on Evolutionary Computation in Combinatorial Optimization*. Springer, 2005, pp. 213–223.

[3] S. Sorlin, C. Solnon, and J.-M. Jolion, "A generic graph distance measure based on multivalent matchings," in *Applied Graph Theory in Computer Vision and Pattern Recognition*. Springer, 2007, pp. 151–181.

[4] R. Ambauen, S. Fischer, and H. Bunke, "Graph edit distance with node splitting and merging, and its application to diatom identification," in *International Workshop on Graph-Based Representations in Pattern Recognition*. Springer, 2003, pp. 95–106.

[5] M. C. Boeres, C. C. Ribeiro, and I. Bloch, "A randomized heuristic for scene recognition by graph matching," in *International Workshop on Experimental and Efficient Algorithms*. Springer, 2004, pp. 100–113.

[6] K. D. Ho, J.-Y. Ramel, and N. Monmarché, "Multivalent graph matching problem solved by max-min ant system." in *S+SSPR*, 2020, pp. 227–237.

[7] R. S. Shen, J. A. Alappatt, D. Parker, J. Kim, R. Verma, and Y. Osmanlıoğlu, "Graph matching based connectomic biomarker with learning for brain disorders," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020, pp. 131–141.

# Transformers for Historical Handwritten Text Recognition

Killian Barrere
*PhD Student*
*killian.barrere@irisa.fr*

Yann Soullard
*PhD Supervisor*
*yann.soullard@irisa.fr*

Aurélie Lemaitre
*PhD Director*
*aurelie.lemaitre@irisa.fr*

Bertrand Coüasnon
*PhD Director*
*bertrand.couasnon@irisa.fr*

Univ Rennes, CNRS, IRISA, France

This work is part of a thesis entitled
"Deep Neural Networks and Attention Mechanisms for Handwritten Text Recognition".
The thesis started on October 1st 2020 and is expected to finish by September 30th 2023.

***Abstract*—Handwritten documents are recently getting more and more publicly available, but searching efficiently information through them is difficult. Handwritten Text Recognition systems automatically transcribe documents and offer excellent solutions to make the content of handwritten documents available. Neural networks are currently the state-of-the-art approaches for this task. Recently, Transformer architectures have gained in popularity in many fields. We present the works we have done so far toward an efficient architecture using transformer layers for the field of Handwritten Text Recognition. Architectures using Transformer for Handwritten Text Recognition are presented. Our architectures aim to replace recurrent layers with transformers, while combining optical recognition and language modeling in end-to-end model. We manage to obtain state-of-the-art results on the IAM dataset with one of our architecture.**

## I. Introduction

Nowadays, a considerable number of handwritten documents have been digitalized and made available to the public to ease their access. However, searching information through them efficiently remains a complex task. While human transcribers could be considered to make the textual content available, this is typically a long and expensive process. This is especially true for difficult historical documents, where even an experienced transcriber could experience difficulties to decipher the textual content.

Offline Handwritten Text Recognition aims to automatically read scanned handwritten documents and output a computer-understandable text. However, they require text-line images to perform their task. Usually, a first step of document layout analysis segments the text from a page into text-line images. Handwritten text recognition systems are then used to obtain their transcripts. Finally, a post-processing step consisting of applying a Language Model is applied to correct eventual errors.

While initial models in the field of handwritten text recognition were mostly based on Hidden Markov Models, deep learning and neural networks have been showing groundbreaking improvements in the field [1]. Model based on recurrent layers have been widely used in the following years [1], [2], [3] thanks to their abilities to model sequential dependencies, paired with the well-known Connectionist Temporal Classification [1]. Convolutional layers have then been considered and added into existing architectures [4], [5].

Recent models of Handwritten Text Recognition perform very well and obtain low error rates on common datasets. However, for harder documents like historical documents, existing models often fall short to obtain low error rates. This is principally due to the inherent difficulties of historical documents caused by complex writings, various styles, deteriorated documents and old languages. These difficulties also impact skilled transcribers which usually result in a longer and costly annotation process. Therefore, historical documents typically contain a very few amount of annotated data. Such documents are challenging for existing models and remain interesting data with much room for improvement before attaining low error rates.

In the field of Natural Language Processing, Transformer models using the so-called Multi-Head Attention have been proposed and showed ground-braking results [6], [7]. Vaswani et al. introduced Multi-Head Attention, which is able to handle long-range context, while providing efficient parallelism, which is crucial in current deep learning approaches. Multi-Head Attention is, therefore, an efficient replacement to long used recurrent layers. In the years that follow, Transformer models have been employed more and more in the field of Natural Language Processing, while also showing promising results in other fields like Automatic Speech Recognition [8], Natural Image Classification [9] or even more recently in the field of Handwritten Text Recognition [10], [11].

In this work, we aim to use Transformer to recognize complex handwritten text in historical documents. As Transformer models represent an efficient alternative to recurrent neural networks, we aim to exploit their superior training capabilities to overcome the gap between the results obtained

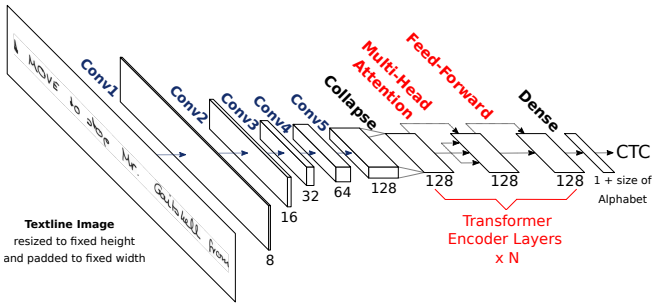



Figure 1. Our Convolutional Transformer Model.

with existing models and low error rates. Furthermore, we believe such models represent an efficient alternative to Language Models, that might be difficult to apply for historical documents, and we aim to combine both the Optical Recognition step and the Language Modeling in an end-to-end model, as achieved by Kang et al. [10].

Transformer models for Handwritten Text Recognition are introduced in this work. In Section II, we introduce Convolutional Transformer, directly derived from Convolutional Recurrent Neural Network where we replaced recurrent layers. Section III then follow and proposes a Sequence-to-Sequence Transformer where we included a Transformer Decoder to combine both Optical Recognition and Language Modeling in an end-to-end fashion. Such architectures are then evaluated on common datasets with preliminary results in Section IV.

## II. Convolutional Transformers

Convolutional Recurrent Neural Networks (CRNN) are widely used for Handwritten Text Recognition [12], [13]. They are good enough to extract local and global features, while providing good information from the past and future context for text lines images by combining both convolutions and recurrent layers.

As Transformers models introduced Multi-Head Attention as an efficient alternative to recurrent layers, we propose replacing the recurrent layers in CRNN with Multi-Head Attention. However, applying transformer to 2D images is far more difficult than applying it to 1D data [9]. To alleviate this issue, we use the inherent sequentiality related to the reading order (from left to right for Latin scripts) of text-line image. This is made possible by a vertical dimension collapsing enabling our model to work on a sequence of features. Connectionist Temporal Classification (CTC) [1] is used to handle the differences in sizes between the predicted character probability and the ground truth.

Our architecture (illustrated in Figure 1) follows the general trend of CRNN architectures. It is composed of a sequence of five convolutional layers aiming to extract local features from the images. The image is quickly reduced in size with 2x2 max pooling following each of the first 3 convolution layers. Following the convolutional backbone, we collapse the vertical dimension with a convolution layer resulting in a sequence of features containing meaningful information for each column of the text-line image. We then use stacked Transformer encoder layers as an alternative to recurrent layers, that are in charge of handling long range dependencies. Each Transformer encoder layer is composed of a layer of Multi-Head Attention and a Feed-Forward Layer, with each sub-layer using residual connections. Following these stacked transformer layers, we apply a dense layer with a number of output neurons equals to the size of the character set, while including the CTC Blank character, before training it with an usual CTC loss function.

## III. Seq2Seq Transformers

Following works on Convolutional Transformer, we tried to propose another model closer to the initial Transformer model by proposing a sequence-to-sequence encoder-decoder (seq2seq) architecture.

Seq2seq models for Handwritten Text Recognition takes as input text-line images as well as the sequence of what the model has already predicted. They output characters one by one. This process, therefore, enables the architecture to model the language and adapt its output character based on the characters before.

With such architecture, we aim to combine both the Optical Recognition with the Language modeling inside mutual Multi-Head Attention layers, therefore resulting in an end-to-end model. Mutual Multi-Head Attention layers combine output of the layer before it with the encoding matrix, obtained after feeding the text-line image to the encoder layers.

Our Sequence-to-Sequence Transformer model (illustrated in Figure 2) is composed of a CRNN (or a Convolutional Transformer) as the encoder, and of a stack of Transformer decoder layers as the decoder. The encoder is used to encode the text-line image in a matrix shape, which is directly obtained from the previous hidden layer. The decoder takes as input the sequence of what the model has already produced. A character-level embedding and positional encoding are applied before feeding that sequence to the decoder. However, at training time, we employ teacher forcing. We feed the target transcription to the decoder, after being shifted right to assure that the model only sees the previous characters and not the characters it should output.

Following the work done by Michael et al. [14], we find it beneficial to use a hybrid loss to train the model. Therefore, in addition to Cross Entropy Loss used to train seq2seq models, we additionally use a CTC loss function which is applied to the output of the encoder (following a dense layer).

## IV. Preliminary Results

We evaluated our architecture on the IAM Dataset consisting of modern English text-line images and the READ

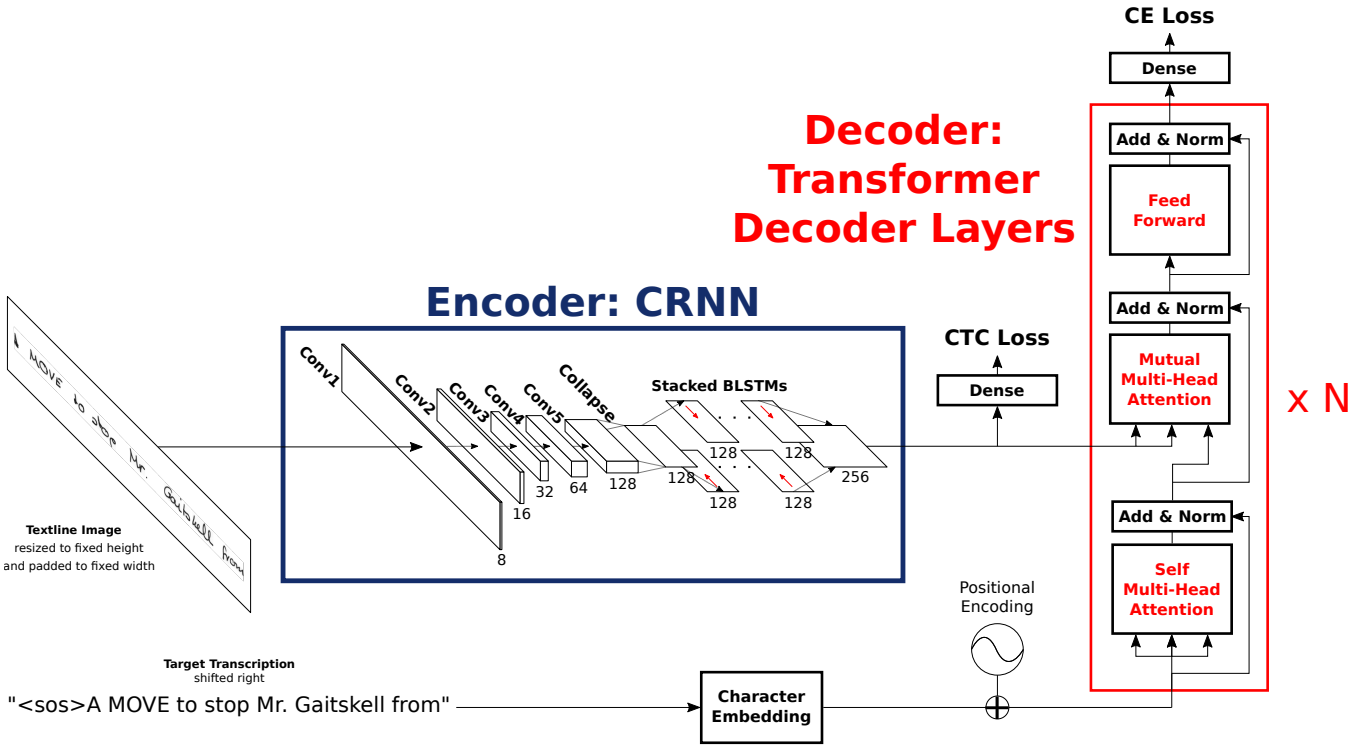



Figure 2. Our Sequence-to-Sequence Transformer Model

Table I
RESULTS ON IAM DATASET (MODERN ENGLISH).

| Model Architecture | Validation | | Test | |
|---|---|---|---|---|
| | CER | WER | CER | WER |
| GCRNN + additional data + LM [4] | | | **3.2** | **10.5** |
| FCN [15] | **3.3** | | 4.9 | |
| Seq2Seq CRNN + LSTM [14] | | | 4.87 | |
| VAN [16] | **3.32** | **13.60** | 4.95 | 16.24 |
| Transformer [10] | | | 7.62 | 24.54 |
| Transformer + synthetic data [10] | | | 4.67 | 15.45 |
| Our Convolutional Transformer | 5.99 | 24.17 | 7.42 | 29.09 |
| Our Seq2Seq Transformer | 4.02 | 16.55 | 5.07 | 21.47 |

Table II
RESULTS ON READ 2018 (HISTORICAL DOCUMENTS). THE DATASET CONSISTS OF 17 HISTORICAL DOCUMENTS FROM VARIOUS LANGUAGES WITH OUR CUSTOM TRAIN, VALIDATION, AND TEST SPLITS.

| Model Architecture | CER | WER |
|---|---|---|
| Our CRNN | 14.27 | **39.35** |
| Our Convolutional Transformer | 16.76 | 45.70 |
| Our Seq2seq Transformer | **12.81** | 41.00 |

2018 dataset [5], which is composed of 17 documents from various languages. For the READ dataset, we used our own custom train, validation and test split. Table I and Table II report the results for each dataset respectively. We also compared our results with the state of the art on the IAM dataset.

Regarding our Convolutional Transformer, we only managed to train small models. It obtains results under the state of the art. Despite the fact that it might be capable to possess more capabilities than typical CRNN models, we find it difficult to exploit such model. Despite that, with such architecture, we managed to obtain promising results. However, we believe there is still room for improvement.

Our seq2seq Transformer meanwhile managed to obtain state-of-the-art results on the IAM dataset without additional data. On the READ dataset, we also obtained a character error rate below what we obtain with a regular CRNN. This architecture shows very promising results, however, we find it difficult to train as it has many hyper-parameters and still have difficulties to converge efficiently. As future works, we plan to invest these problems and pursue our work toward improving further our results with this architecture.

Transformer models, nonetheless, remain a challenging architecture for historical datasets. As they require a fair amount of annotated data to obtain the best out of the architecture.

## V. CONCLUSION AND FUTURE WORKS

In this work, we presented two architectures using Transformer for Handwritten Text Recognition. We use Transformer to replace recurrent layers of a CRNN, resulting in

a Convolutional Transformer model. We proposed a second architecture: seq2seq Transformer in which we included a Transformer Decoder. With this architecture, we aim at combining both Optical Recognition and Language Modeling in an end-to-end fashion. We obtain state-of-the-art results on the IAM dataset with our seq2seq transformer with a character error rate of $5.07$ on the test set, without additional data. However, we find it difficult to train such models.

Ongoing works are dedicated to efficiently training such architectures by improving the training optimization, while also using more data augmentation techniques and synthetic data. This is significant as these models seem to require a substantial amount of annotated data, whereas historical documents contain few annotated data. Furthermore, we would like to work on augmentation techniques fitted especially for historical documents.

Upcoming works will be dedicated to pushing the results further, while focusing our works on historical documents.

Following that, we hope to publish our works, therefore providing a more detailed view of our models and the training procedure while making our code available to ease reproduction.

## ACKNOWLEDGMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012550 made by GENCI.

## REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[2] T. Bluche, J. Louradour, M. Knibbe, B. Moysset, M. F. Benzeghiba, and C. Kermorvant, "The a2ia arabic handwritten text recognition system at the open hart2013 evaluation," in *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 161–165.

[3] V. Frinken and S. Uchida, "Deep blstm neural networks for unconstrained continuous handwritten text recognition," in *13th ICDAR*. IEEE, 2015, pp. 911–915.

[4] T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," in *14th IAPR ICDAR*, vol. 1. IEEE, 2017, pp. 646–651.

[5] T. Strauß, G. Leifert, R. Labahn, T. Hodel, and G. Mühlberger, "Icfhr2018 competition on automated text recognition on a read dataset," in *216th ICFHR*. IEEE, 2018, pp. 477–482.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[9] 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.

[10] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *arXiv preprint arXiv:2005.13044*, 2020.

[11] S. S. Singh and S. Karayev, "Full page handwriting recognition via image to sequence extraction," *arXiv preprint arXiv:2103.06450*, 2021.

[12] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" in *14th IAPR ICDAR*, vol. 1. IEEE, 2017, pp. 67–72.

[13] Y. Soullard, W. Swaileh, P. Tranouez, T. Paquet, and C. Chatelain, "Improving text recognition using optical and language model writer adaptation," 2019.

[14] J. Michael, R. Labahn, T. Grüning, and J. Zöllner, "Evaluating sequence-to-sequence models for handwritten text recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1286–1293.

[15] M. Yousef, K. F. Hussain, and U. S. Mohammed, "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks," *arXiv preprint arXiv:1812.11894*, 2018.

[16] D. Coquenet, C. Chatelain, and T. Paquet, "End-to-end handwritten paragraph text recognition using a vertical attention network," *arXiv preprint arXiv:2012.03868*, 2020.