ICDAR2019 Doctoral Consortium

(in conjunction with ICDAR 2019)

Sydney, Australia

Date: 2:00 pm – 6:15 pm, September 22, 2019
Location: CB11.04.103 and 105

Chairs: Véronique Eglin 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) and ICDAR 2017 (in Kyoto, Japan). The Doctoral Consortium at ICDAR 2019 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 2019 Doctoral Consortium has accepted 19 students, which have been mentoring by 19 senior active researchers of all nationalities working in the field of Document Image Analysis and Recognition. During the DC, each research proposal is presented 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 to attend the poster session in which the students and their mentors discuss project details. The prize for the best poster will be delivered at the conference gala evening.

PROGRAM

2:00 – 2:15 pm	Opening - Introduction to ICDAR Doctoral Consortium 2019 *J-C Burie & V Eglin*
2:15 – 3:15 pm	Teaser presentation of each PhD project *Tutees*
3:15 – 4:00 pm	*Setting-up of Posters and Coffee Break*
4:00 – 4:30 pm	Talk "How to succeed in your Ph.D. degree" ***Jean-Marc Ogier,** professor La Rochelle University, France*
4:30 – 6:00 pm	Poster session and discussions
6:00 – 6:30 pm	Concluding remarks and Best Poster Award

OVERVIEW OF CONTRIBUTIONS

Most of the Ph.D propositions attempt to bring innovative solutions with the aim of facing new major societal challenges. The DC 2019 is the opportunity to raise new issues about how to make data more secure (fight against counterfeits), how to access larger scale datasets, how to produce more efficient representation models (from characters levels to structures) seeking to take advantage of very recent machine learning solutions, mostly based on deep artificial neural networks approaches.

The 19 projects may be clustered in those main topics:

- Document Layout analysis
- Multilingual texts: OCR, handwriting recognition and transcription
- Multimodal machine learning for scene and document interpretation
- Large scale document image processing and accessibility
- Semantic understanding
- Information (textual) extraction:
  - From document template free
  - From images and videos

Allan Marvin Ssemambo	A model to automate uganda sign recognition and translation to English	Uganda
Nishatul Majid	Developing an offline Bangla handwriting recognition system	USA
Raul Gomez	Exploiting the Interplay between Visual and Textual Image Content for Scene Interpretation	Spain
Showmik Bhowmik	An Integrated Document Layout Analysis System	India
Florian Westphal	Training Data and Time Efficient Algorithms for Historical Document Analysis	Sweden
Chandra Sekhar	Interpretable Online Signature Verification through Generative Adversarial Networks (GANs).	India
Julien Maitre	Detection and analysis of weak signals. Development of a digital investigation framework for a hidden alert launcher service	France
Lady Viviana Beltrán Beltrán	Multimodal Machine Learning: Representation, Fusion and Applications	France
Shreya Goyal	Semantic Understanding of Floor Plan Images through Machine Learning Techniques	India
Xenofon Karagiannis	Layout Analysis and Recognition in historical documents using machine learning methods	Greece

Camille Guerry	Historical big-data: modelization of strategies to analyze collections of documents	France
Bhargava Urala Kota	Automated Lecture Video Summarization via Extraction and Feature Representation of Text Content	USA
Thi Tuyet Hai Nguyen	Multilingual OCR correction for ancient books: Looking at multiple documents to fix multiple words	France
Brian Davis	Template-Free Information Extraction From Arbitrary Form Images	USA
Olfa Mechi	Transcription and indexing of text in archival documents using deep architectures	Tunisia
Clément Sage	Table Information Extraction from Business Documents	France
Antoine Pirrone	Multimodal analysis and reconstruction of ancient papyrus fragments using image processing and deep learning	France
Qingqing Wang	Detect and recognize text from images	Australia
Rohit Saluja	Interactive Systems for Reading Texts in Indian Streets and Documents	Australia

List of mentors

Barney Smith	**Elisa H.**	Boise State University, USA
Coüasnon	**Bertrand**	Irisa, Rennes, France
Coustaty	**Mickaël**	Laboratoire L3i, Université de La Rochelle, France
Essoukri Ben Amara	**Najoua**	Ecole Nationale d'Ingénieurs de Sousse, Sousse, Tunisia
Fischer	**Andreas**	University of Fribourg, Switzerland
Fornés	**Alicia**	Computer Vision Center, Universitat Autònoma de Barcelona, Spain
Jawahar	**C.V.**	International Institute of Information Technology, Hyderbad, India
Lins	**Rafael**	Federal University of Pernambuco, Brazil
Llados	**Josep**	Computer Vision Center, Universitat Autònoma de Barcelona, Spain
Marcelli	**Angelo**	DIEM - Universita di Salerno, Italy
Palaiahnakote	**Shivakumara**	University of Malaya, Malaysia
Paquet	**Thierry**	Laboratoire LITIS, Université de Rouen, France
Pratikakis	**Ioannis**	Democritus University of Thrace, Greece
Ramos Terrades	**Oriol**	Computer Vision Center, Universitat Autònoma de Barcelona, Spain
Shafait	**Faisal**	National University of Sciences and Technology, Pakistan
Sidère	**Nicolas**	Laboratoire L3i, Université de La Rochelle, France
Uchida	**Seiichi**	Human Interface Laboratory, Kyushu University, Japan
Zanibbi	**Richard**	Dept. Computer Science, Rochester Institute of Technology, USA

Short bio of the DC Chairs

Veronique Eglin is full professor in computer science at INSA Lyon since 2015 and member of IMAGINE team in LIRIS – UMR CNRS 5205 since 2005. She obtained her PhD degree in Computer science in 1998 and her *Habilitation à Diriger les Recherches* in Computer Science in 2014 at INSA de Lyon. She is today head of IMAGINE Team in the LIRIS laboratory and deputy director of the teaching First Cycle of INSA de Lyon. Her scientific publications deal with the topic of document analysis and content recognition, mainly focused on document segmentation and recognition, handwriting identification, word spotting and automatic transcription. In that context, her current topics of interest deal with multiscale analysis, incremental learning, graph-embedding representation and recently pattern mining for symbolic information spotting. Her industrial, academic and multidisciplinary collaborations contributed those last years to the supervision of 14 Ph.D theses in computer vision and document image analysis and recognition, twelve papers in international journals, five books chapters, more than seventy publications in selective international IEEE conferences and workshops. Since 2000, she has also contributed to the development of several research associations in the field of document analysis and recognition (GDR-I3 of the CNRS (GDR 722), Cluster ARC5 in the Rhône-Alpes Region, GRCE, Valconum).

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 Vice-Chair of the TC10 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 old 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.

A MODEL TO AUTOMATE LUGANDA SIGN RECOGNITION AND TRANSLATION

Ssemambo Marvin, 210025242, 2016/HD05/344U

sallanmarvin@gmail.com; sallanmarvin@yahoomail.com, mssemambo@cis.mak.ac.ug
OPTION
COMPUTER VISION & NATURAL LANGUAGE PROCESSING
Master of Science in Computer Science

College of Computing and Informatics Science,

Makerere University

ABSTRACT

A sign always suggests the presence of an information, circumstance, or quality. Signs are universally in our lives. They ensure our lives are easier when we are conversant with them. But most times they posture problems. For example, a traveller/tourist might not be able to comprehend signs in a foreign country. This paper talks about problems of automated sign recognition and conversion/translation. A model capable of apprehending images, noticing and identifying signs, and translating them into a target language is proposed. This research work will look into the numerous methodologies for automatic sign recognition, extraction and translation will be described. User-centered approach will be used in model development and natural language translation (NLT) which is a key application in the field of natural language processing and its requirements to deliver more robust and dependable system that will be resistant to failure irrespective of users' inputs. The methodology takes advantage of human intelligence and leverage human capabilities. Currently the focus is working on Luganda sign translation. In this proposal, NLP for translating Luganda language to English language is proposed. In the design, the input is images with Luganda text only and the direction of translation is Luganda to English language, although in the future extension of this work, speech would also be accepted as input and the direction of translation would be expanded. NLP of Luganda Language would enhance knowledge transfer and communication using the Luganda language.

1. Introduction

21st century is period of information. Images, as an optical basis for observing the world, is the key media for information attainment, communication and broadcast. Digital image processing procedures (DIPT) started in 1920s' and advanced after 80s'.Currently, with the speedy development of computer techniques and related theories, applications of DIPT in many fields have received wide attention and have made pioneering achievement in such as biomedical engineering, industrial examination, machine vision system, public safety and justice, and military guidance [34]. Persons communicate with others using a collection of information systems and media in progressively varied atmospheres. One method of common communication media is a sign. A sign always suggests the presence of an information, circumstance, or quality. Signs are universally in our lives. They ensure our lives are easier when we are conversant with them. But most times they posture problems. For example, a traveler/tourist might not be able to comprehend signs in a foreign country that stipulates military caution or dangers [5].

Text that is enclosed in images taken from natural scene using digital cameras or either in a form of document like scanned CD/book covers or video images. Video text can be widely classified into two categories: overlay text and scene text. Overlay text refers to those characters created by graphic titling machines and overlaid on video frames images, such as video captions, while scene texts happens naturally as part of the scene, like text in information boards/signs, nameplates, food flasks, etc. Since the text data can be implanted in an image or video in dissimilar front styles, sizes, orientations, color, and against a complex background, the difficult of extracting the candidate text region becomes a challenging one [22].

2. Problem Statement

Text that is implanted in images comprises of significant and valuable semantic information, which can be used to completely comprehend images. The recognition and extraction of text region in an image is commonly known difficulty in the computer vision research and the efficiency of different approaches strongly relies on character size. Since in natural scene the perceived characters may be extensively of different sizes, it therefore problematic to extract all text areas from the image by means of only a single method. These solutions have not been possible with most African languages like Luganda since its phonetics and transcriptions are not added into these methods to enable the recognition and translation for Luganda text.

United Nation Education Scientific and Cultural Organization (UNESCO) plays a very important part in ensuring that the values of the World Heritage Site are maintained and requires Buganda kingdom to translate most of the signs in its cultural site to English [39]. However the kingdom wants to reserve the Luganda language in existence and most especially promote Luganda literature.

2.1 General Objective

To develop a model for automation of Luganda sign recognition and translation of texts in signs to English.

2.2 Objectives

This study will be directed by the following specific objectives;

i. To identify current methods of character recognition and text translation in order to establish methods of modeling character recognition and text translation for Luganda.

ii. To generate adequate experimental dataset for character recognition and text translation for Luganda.

iii. To design character recognition and text translation model for Luganda.

iv. To evaluate the model for accuracy.

2.3 Significances of the Research

The research provides the following values and benefits to users, researchers and competitors.

A Luganda sign recognition and translation model will be developed, with images of natural scene signs as input and then a translation of the text in the images is output which is further converted into other international languages using the existing translation engines like for Microsoft, Babylon, Google, festival, marytts among others.

·The automated model will establish a mathematical theory of exactly how units of syntax making source language are translated to units of syntax making Target Language; gives arranged and clear-cut meanings of English words and simple terms in Luganda Language.

The research will provide means of training and teaching Luganda Language to non-indigenous leaners and work as an educational instrument with high reliability on electronic media.

· Innovative products in text translation concerned with Luganda linguists will be realized through software developers who will develop solutions on mobile, web and desktop that capture Luganda text and generate corresponding English text.

· Using the image dataset of the Luganda signs, software developers in the Ugandan industry will be motivated to build open source platforms in order to have purely innovative products that will solve local problems in language linguistics.

· The technology can be improved in the future to help visually handicapped persons to increase environmental awareness.

3.0 Methodology

Previous methods to scene text detection fall short when dealing with challenging situations, even when equipped with deep neural network models because the general performance is determined by a connection of multiple stages and components in pipeline. We have used a simple pipeline that yields fast and accurate Luganda text detection in natural scenes. The pipeline predicts words or lines of text of arbitrary orientations and shapes in full images, eliminating unnecessary immediate shapes.

- Text data is a very important element of modern day communication. However, there is a caveat, text can appear in multiple languages most of which everyone cannot understand. In this project we detect common text in the wild and use modern computer vision techniques to extract and interpret the text. We also attempt to provision our pipeline to be able to translate the output text from luganda to English.

Dataset:

Our dataset consists of colored images obtained from the natural environment. Each sample of the images contain articulate text in luganda which we use to train our model. The images were obtained from natural scene settings and comes in various shapes, fonts and sizes. A sample of the dataset is shown below.

References

1. H.K. Kim, Efficient Automatic Text Location Method and Content-Based Indexing and Structuring Of Video Database, Journal of Visual Communication and Image Representation vol. 7, no. 4 ,1996, pp. 336–344.
2. C. Y. Suen, L. Lam, D. Guillevic, N. W. Strathy, M. Cheriet, J. N. Said, and R. Fan, Bank Check Processing System, International Journal of Imaging Systems and Technology, vol. 7, No. 4 1996, pp. 392–403.
3. D.S. Kim, S.I. Chien, Automatic Car License Plate Extraction using Modified Generalized Symmetry Transform and Image Warping, Proceedings of International Symposium on Industrial Electronics, Vol. 3, 2001, pp. 2022–2027.
4. A.K. Jain, Y. Zhong, Page Segmentation using Texture Analysis, Pattern Recognition, Vol. 29, No. 5, Elsevier, 1996, pp. 743–770.
5. T.N. Dinh, J. Park and G.S. Lee, Low-Complexity Text Extraction in Korean Signboards for Mobile Applications, IEEE

Developing an Offline Bangla Handwriting Recognition System - Research Highlights

Nishatul Majid
Supervisor: Elisa H. Barney Smith
Department of Electrical and Computer Engineering
Boise State University
Boise, Idaho, USA
nishatulmajid@u.boisestate.edu

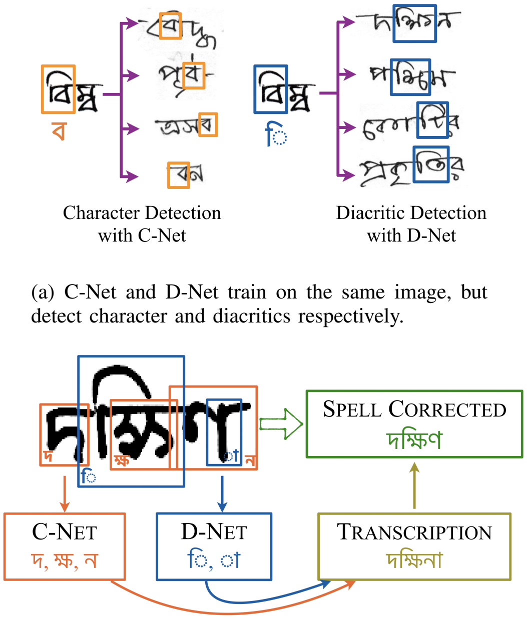
***Abstract*—The aim of this research is to develop an offline Bangla handwriting recognition system using sequential detection of characters and diacritics. This is an entirely segmentation-free approach where the characters and associated diacritics are detected separately with different networks and later the results are merged to form a transcription. This method capitalizes on segmented training data, and the Boise State Bangla Handwriting dataset was developed to complement this approach. This contains offline handwriting both in essay and isolated character format. This is prepared carefully to cover a major portion of the Bangla script. Also an experiment of comparing recognition performance between scanned vs camera-acquired data has been done using this dataset. Future plans are expanding the dataset as well as the handwriting transcription approach by means of using machine printed data, developing advanced data augmentation technique using stroke features, etc. Also, testing this system on a third party Bangla dataset to verify the strength of this approach and use Natural Language Processing to further improve transcription accuracy.**

I. Brief About the Bangla Script

Bangla/Bengali is one of the most used languages in the world. With over 205 million people, it is the 7th most spoken native language. Bangla script, along with the almost identical Assamese alphabet, is the fifth most widely used writing system in the world. It is the national and official language of the People's Republic of Bangladesh, and official language of several states in India such as West Bengal, Tripura, Assam, Andaman etc. Bangla belongs to the Abugida class of writing system. Despite being a major script, not much progress has been made in terms of handwriting recognition. Also, handwriting recognition is particularly useful in developing countries like Bangladesh where handwritten documents are prevalently used.

II. Segmentation-free Transcription by Sequential Detection of Character/Diacritics

The idea is to look for each possible character element (like basic characters, diacritics, conjuncts, punctuation etc.) of the script inside the words and combine the results into text, rather than segmenting the words followed by recognizing the characters [1]. This is a lexicon independent approach. The technique relies on character level ground truth coordinates for training and performs character or diacritic spotting on word images. Individual networks, named C-Net (Character Network) and D-Net (Diacritic Network) are separately trained for detecting characters and diacritics. The training images are the same for both of these networks while the detection output classes are different, as shown in Fig 1a. The idea is to look through each possible character and diacritic and sequentially spot where (if) each character/diacritic is found in the word or page to be recognized using the networks. Afterwards, the detection results are combined to form a transcription as shown in Fig. 1b.



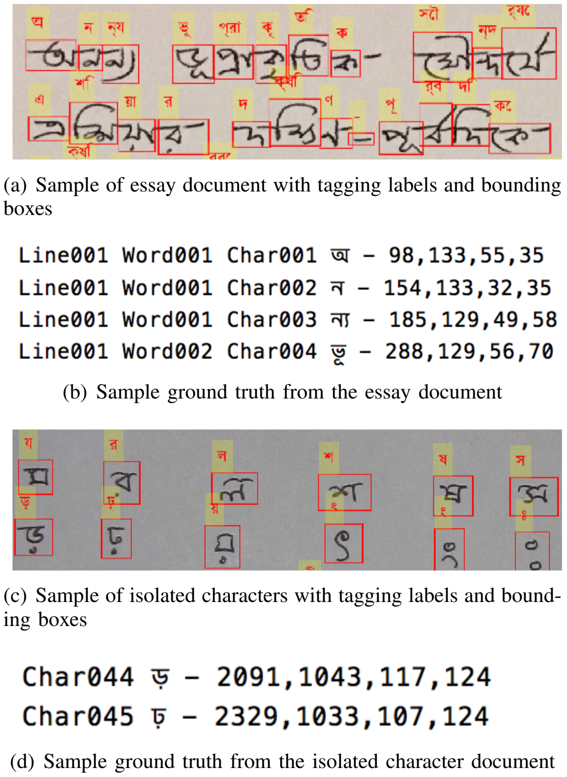
(a) C-Net and D-Net train on the same image, but detect character and diacritics respectively.

(b) Combining the detection results to form a transcription

Fig. 1. Schematic overview of the segmentation-free transcription framework

III. Boise State Bangla Handwriting Dataset

The Boise State Bangla Handwriting Dataset is a publicly available dataset developed during this program [2] [3]. It contains both essay scripts and isolated characters. One of the highlight features of this dataset is all of its content are tagged at the character, word and line level with associated ground



(a) Sample of essay document with tagging labels and bounding boxes

```
Line001 Word001 Char001 অ - 98,133,55,35
Line001 Word001 Char002 ন - 154,133,32,35
Line001 Word001 Char003 ন্য - 185,129,49,58
Line001 Word002 Char004 ভু - 288,129,56,70
```

(b) Sample ground truth from the essay document

(c) Sample of isolated characters with tagging labels and bounding boxes

```
Char044 ড় - 2091,1043,117,124
Char045 ঢ় - 2329,1033,107,124
```

(d) Sample ground truth from the isolated character document

Fig. 2. Boise State Bangla Handwriting dataset. (a), (c) ground truth overlay and (b), (d) ground truth tag files of the essay and isolated character document.

truth. Images are digitized using a flat-bed scanner and cell-phone cameras. This is an up-growing dataset and the plan is to sufficiently cover a major portion of the entire Bangla script.

IV. Experiments Done

Several experiments have been done with this method using the Boise State Bangla Handwriting Dataset. First, an isolated Bangla basic character recognizer was designed with features being extracted based on zonal pixel counts, structural strokes and grid points with U-SURF descriptors modeled with bag of features [2]. This was used to benchmark this dataset with the other publicly available Bangla datasets. The highest recognition accuracy of 96.8% was found with an SVM classifier based on a cubic kernel. Afterwards, we developed the segmentation-free handwriting recognition method as described in Sec II that gave us a transcription with CER of 11.2% and WER of 24.4% [1]. A spell checker was developed which further minimized the errors to 8.9% and 21.5% respectively. Lastly, we also compared recognition performance between scanned and cell-phone camera acquired data using these two methods [4]. As expected, we found that the performance is better with higher quality images but not by a significant margin. Also surprisingly, we found training with higher quality images produced better results than lower quality images when both were tested on lower quality images.

V. Future Plans

A. Finding Effect of Tagging Accuracy

Since ground truth tagging is the most time consuming and laborious process of any dataset preparation and machine learning, we want to test how much the recognition performance varies if the tagging is slightly inaccurate or done differently, Fig. 3. Outcome of this research has the potential to drastically reduce the labor and time required for preparing such dataset.

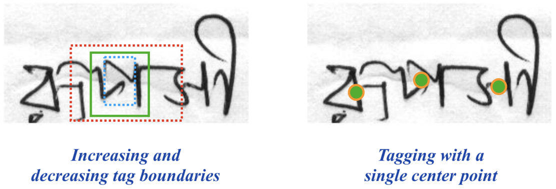
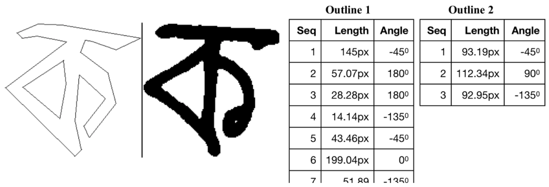

Increasing and decreasing tag boundaries
Tagging with a single center point

Fig. 3. Monitoring recognition performance varying tagging accuracy

B. Advanced Data Augmentation with Stroke Features

Data augmentation is proved to be very useful particularly when working with a small dataset. A stroke feature extractor is developed which extracts the important transition points as well as tracks the stroke outlines as shown in Fig. 4. The plan is to alter the direction, thickness or uniformity of the strokes to obtain augmented data.



Outline 1

Seq	Length	Angle
1	145px	-45^0
2	57.07px	180^0
3	28.28px	180^0
4	14.14px	-135^0
5	43.46px	-45^0
6	199.04px	0^0
7	51.89	-135^0

Outline 2

Seq	Length	Angle
1	93.19px	-45^0
2	112.34px	90^0
3	92.95px	-135^0

Fig. 4. Working dynamics of stroke feature [2].

C. Testing on Third Party Dataset with NLP

There are other publicly available datasets (such as CMATERdb) which although can't be used for training in proposed method, but can be used for testing. The plan is to test the segmentation-free method with such a dataset and use Natural Language Processing to further improve the result.

References

[1] N. Majid and E. H. Barney Smith, "Segmentation-free bangla offline handwriting recognition using sequential detection of characters and diacritics with a Faster R-CNN," in *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019.

[2] N. Majid and E. H. Barney Smith, "Introducing the Boise State Bangla Handwriting dataset and an efficient offline recognizer of isolated Bangla characters," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 380–385.

[3] N. Majid and E. H. Barney Smith, "Boise State Bangla Handwriting Dataset," https://doi.org/10.18122/saipl/1/boisestate, 2018.

[4] N. Majid and E. H. Barney Smith, "Performance comparison of scanner and camera-acquired data for Bangla offline handwriting recognition," in *8th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, 2019.

Exploiting the Interplay between Visual and Textual Image Content for Scene Interpretation

Raul Gomez
Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
Eurecat, Centre Tecnològic de Catalunya, Unitat de Tecnologies Audiovisuals, Barcelona, Spain
raulgomez@cvc.uab.com

I. Thesis Information

Title: Exploiting the Interplay between Visual and Textual Image Content for Scene Interpretation
University: Universitat Autònoma de Barcelona
Supervisors: Dimosthenis Karatzas, Lluis Gomez and Jaume Gibert
Starting date: 01/03/2017
Expected ending date: 2020

II. Research Plan

A. Self-Supervised learning from images and associated text

In this PhD thesis I research how to design systems that learn together from images and associated information, with an special focus on web and social media data. In the first year the main focus has been to explore the state of the art of multi-modal learning from images and associated text, to do a performance comparison of existing techniques, and to apply those in specific problems exploiting social media data. In [4], [5] we compare the performance of different text embeddings to learn a common space for images and words for multimodal image retrieval when learning from Web and Social Media data. We analyze the semantic structure of the learnt joint multimodal space and explore its possibilities in the image retrieval by text task. In [3] we apply that approach to learn relations between words and images from a dataset made of Instagram images associated to Barcelona, and show how that can be useful to analyze the differences between visual relations established by tourists and locals. In [6] we explore the effectiveness of visual features learnt with the different text embeddings for image classification.

B. Multi-Modal models for Hate Speech Detection

I worked on multi-modal understanding pipelines where visual and textual data have to be analyzed jointly to make a decision, and applied them to the problem of hate speech detection on multimodal publications. Multi-modal hate speech detection is a problem where information of two different modalities (image and text) has to be analyzed together to make a decision. It's a problem that has not been addressed before, so I created a new database using Amazon Mechanical Turk. Then I implemented state of the art multi-modal nets and applied them to this problem. Finally I wrote a report which has been submitted for publication to GCPR [2].

C. Text Style Transfer

I explored neural style transfer to transfer the style between text images. Models were trained to transfer styles in three different text modalities: Scene text, machine-printed text and handwritten text. I evaluated style transfer as a data augmentation technique to train scene text detectors, proving that it was effective [1].

D. Ongoing Research

Currently I'm researching on extending the work on self-supervised learning from images and associated text to other modalities, such as image geolocation, date, or author. I'm interested on exploring the possibilities of learning joint multi-modal semantic embeddings with multiple data modalities.

III. Publications

Exploring Hate Speech Detection in Multimodal Publications
Raul Gomez, Jaume Gibert, Lluis Gomez, Dismosthenis Karatzas
Submitted to GCPR, 2019

Selective Text Style Transfer
Raul Gomez, Ali Biten, Dismosthenis Karatzas, Lluis Gomez, Maral Rossinyol, Jaume Gibert
ICDAR, 2019

Self-Supervised Learning from Web Data for Multimodal Retrieval
Raul Gomez, Dismosthenis Karatzas, Lluis Gomez, Jaume Gibert
Book Chapter, Multi-Modal Scene Understanding, 2019

Learning from Barcelona Instagram data what Locals and Tourists post about its Neighbourhoods
Raul Gomez, Dismosthenis Karatzas, Lluis Gomez, Jaume Gibert
ECCV MULA Workshop, 2018

Learning to Learn from Web Data through Deep Semantic Embeddings
Raul Gomez, Dismosthenis Karatzas, Lluis Gomez, Jaume Gibert

ECCV MULA Workshop (Oral), 2018

TextTopicNet - Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces
Yash Patel, Lluis Gomez, **Raul Gomez**, Maral Rusiol, Dismosthenis Karatzas and CV Jawahar.
arXiv preprint arXiv:1807.02110, 2018

ICDAR2017 Robust Reading Challenge on COCO-Text
Raul Gomez, Baoguang Shi, Lluis Gomez, Lukas Numann, Andreas Veit, Jiri Matas, Serge Belongie and Dismosthenis
ICDAR, 2017

FAST: Facilitated and Accurate Scene Text Proposals through FCN Guided Pruning
Dena Bazazian, **Raul Gomez**, Anguelos Nicolaou, Lluis Gomez, Dimosthenis Karatzas and Andrew Bagdanov
Pattern Recognition Letters, 2017

Improving Text Proposals for Scene Images with Fully Convolutional Networks
Dena Bazazian, **Raul Gomez**, Anguelos Nicolaou, Lluis Gomez, Dimosthenis Karatzas and Andrew D. Bagdanov
ICPR Workshop on Deep Learning for Pattern Recognition, 2016

IV. CV

A. Education

Industrial PhD student: Eurecat and Computer Vision Center, Universitat Autónoma de Barcelona
Master in Computer Vision: Universitat Autónoma de Barcelona
Bachelor's Degree in Telecomunications Engineering: Universistat Politécnica de Catalunya
Master Thesis: Efficient discovery of text in the wild using Fully Convolutional Networks Research the possibilities of fully convolutional networks in text prediction tasks. Implementation of a method to efficiently retrieve images with textual content from a huge dataset with high precission. Implementation of a pipeline to produce object proposals with high recall. Superior state of the art results achieved in both tasks. Gave rise to a publication on the Deep Learning and Pattern Recognition Workshop 2016
Bachelor Thesis: Video and object syncronization for video content enhancement Development of an online web-based video object segmentation and edition tool. Developed in HTML5 and JavaScript for the client side, and in C and C++ for the server-side, using a MySQL database, AJAX and CGI for the client-server comunication. The tool is based on the video object segmentation software Sensarea, developed by Pascal Bertolino.

V. Projects

Personal Blog: I have a blog where I explain my PhD work, toy experiments and general machine learning concepts. *https://gombru.github.io/*
SetaMind: I developed this Android App that, given a photo of a mushroom, recognizes its species. It uses a CNN that runs locally in the phone. *play.google.com/store/apps/details?id=gombru.setamind*
FaceFCN: A fully convolutional network that performs pixel level face and hair segmentation. *gombru.github.io/2018/01/08/face_hair_segmentation*
InstaBarcelona: I developed tools that learn from images and associated text, and applied that to Instagram images related to Barcelona. I presended this work in the TurisTIC Forum of Barcelona. *gombru.github.io/2018/01/12/insta_barcelona*
Multi-Modal Retrieval Demo: I developed an online demo to show the performance of the developed multi-modal semantic image retrieval algorithms. It was shown in the Barcelona' Mobile World Congress. *https://gombru.github.io/MMSemanticRetrievalDemo/*

A. Talks

MULA ECCV Workshop: I gave a talk about my paper Learning to Learn from Web Data through Deep Semantic Embeddings in this Multimodal Learning and Applicattions Workshop.
ForumTurisTIC: I presented in this tourism forum in Barcelona the possibilities of applying my research on learning from social media images and associated text to tourism analysis in the inspirational session. *gombru.github.io/2018/02/11/forumTurisTIC_presentation/*

References

[1] Raul Gomez, Ali Furkan Biten, Lluis Gomez, Jaume Gibert, Marçal Rusiñol, and Dimosthenis Karatzas. Selective Style Transfer for Text. *ICDAR*, 2019.

[2] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring Hate Speech Detection in Multimodal Publications. *Submitt. to GCPR*, 2019.

[3] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Learning from #Barcelona Instagram data what Locals and Tourists post about its Neighbourhoods. *ECCV MULA Work.*, 2018.

[4] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Learning to Learn from Web Data through Deep Semantic Embeddings. *ECCV 2018, MULA Work.*, 2018.

[5] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Self-Supervised Learning from Web Data for Multimodal Retrieval. *Multi-Modal Scene Underst.*, 2019.

[6] Yash Patel, Lluis Gomez, Raul Gomez, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. TextTopicNet - Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces. *arXiv*, 2018.

An Integrated Document Layout Analysis System

Student's Name: Showmik Bhowmik,
Visvesvaraya PhD Scholar, Department of
Computer Science & Engineering, Jadavpur
University, Kolkata, India.
E-mail: showmik.cse@gmail.com

Supervisor's Name: Ram Sarkar, PhD
Associate Professor, Department of Computer
Science & Engineering, Jadavpur University,
Kolkata, India.
E-mail: raamsarkar@gmail.com

***Starting date**: 23rd Feb 2016 & **Expected finalization date:** Feb 2021*

Jadavpur University, Kolkata, India.

I. Overview

Documents are in use as a popular media for storing and transferring knowledge since long. The archives stored in almost all corners of this world have high societal and cultural values. Even in the current days, massive amount of documents are generated in the form of Newspapers, Magazines, Forms, Books, etc., thereby possessing enormous information. Knowledge contained in these documents are needed to be secured and electronically available. Thus a large belt of researchers across the world have indulged themselves in developing Document Image Processing (DIP) system, and Document Layout Analysis (DLA) is considered as one of the major components for any comprehensive DIP system.

DLA can be defined as a method of detecting and categorizing the regions of interest in the scanned image of the text documents. Dissection of text regions from non-text ones is one of primary necessities of the reading system along with organization of their correct reading orders, detection and labeling of the different blocks as text body, illustrations, symbols, tables implanted in a document. The process of doing this is, in general, called *physical layout analysis*. Nevertheless, text regions play diverse logical roles inside the document such as title, subtitle, caption, footnote, etc., and this is kind of semantic labeling which comes under the purview of the *logical layout analysis*. Therefore, DLA is the union of physical and logical labeling of the document.

Due to the importance of this problem, the research on DLA has been initiated long time ago. Many methods have been introduced [1][2][3]. However, these methods have considered mostly the document with simple layout as well as textual content. In a survey [4] published in 2016, *Eskenazi et al.* have categorized and summarized these methods introduced in the literature since 2008. But as the technology advances, documents with more complex layout come into the existence. Mostly the presence of different types of non-textual contents makes these documents structurally complex. Thus recently researchers have started focusing on the separation and processing of non-textual contents present in a document image, in order to get better result [5][6]. Besides, many commercial and non-commercial tools like Tesseract, ABBYY (FineReader), Aletheia, LAREX have also been introduced for DLA [7].

Although a good number of methods have been found in the literature, none of these produces desired result for all kinds of documents. As the structure of the documents is getting complex over the time, the earlier methods introduced for DLA have become unusable. Even though the outcomes produced by recent methods are good but not at par. That means still there is a scope for improvement. For that reason PRImA Research Lab, University of Salford, UK regularly arranges competitions on complex layout analysis [8], under the banner of ICDAR competitions. In every new competition, they introduce some additional challenges for the participants. This also indicates that the DLA is still an open research problem.

Generally, a DLA system by and large consists of a *preprocessing stage*, *text/non-text separation stage* and a *region generation and classification stage*. Different methods available in the literature follow different techniques for these stages.

II. Current progress

The current progresses so far is listed below,

A. Preprocessing

Binarized documents are usually easy for further processing. Almost all the methods introduced for DLA have considered binarized images as input. But a poor binarization method can have ill-effect on the performance of next stages. If the input document is noisy then an efficient binarization technique a pre-requisite.

a) Considering this fact, we have introduced a novel background estimation and elimination technique to remove background noise or pixel level variation [9].

b) We have also designed a game theory based binarization technique for degraded document images [10].

B. Text/non-text separation

Identification as well as separation of non-textual content present in a document image becomes a pressing need in order to get a desired result in DLA. So, in the present scope of the work, we have made the following contributions:

a) We have prepared a thorough survey on text/non-text separation to summarize the current status of the research made in this domain [11].

b) We have developed a text/non-text separation method for printed documents [12].

c) We have also developed two text/non-text separation methods for handwritten document images - one for handwritten class notes [13] and another for handwritten laboratory reports [14].

d) We have made an empirical study on text/non-text separation on handwritten pages using different LBP based features [15].

C. *BINYAS*: a complex document layout system

The above mentioned contributions can be considered as the preprocessing steps for DLA. Besides, we have also developed a DLA system named as “BINYAS” (a Bengali word means arrangement) for the printed documents.

a) **BINYAS: a brief overview**

Our input is a color image I which we first convert into its grayscale counterpart I_g and then we subsequently use a contrast stretching operation and a region filling operation which gives us I_{fill}. We binarized I_{fill} with the help of a global threshold based method and get I_b. We also consider another binarized image I_b' which we get by applying our binarization technique GiB [10]. Next we check all the connected components (CCs) in I_b in terms of their height, width and aspect ratio and based on that we take a decision to exclude the separators and margins from both I_b and I_b'. After that we again examine the CCs of I_b to classify the input document as geometrically complex or simple.

For complex document, we partition the components present in it into two images - I_{large} and I_{small}. We generate I_{large} from I_b and I_{small} from I_b'. After that we perform text and non-text classification on both the I_{large} and I_{small} separately. From I_{large} we get text only image I_t^l and non-text only image I_{nt}^l. Similarly from I_{small} we get I_t^s and I_{nt}^s. Then we combine I_{nt}^l and I_{nt}^s to generate the final non-text only image I_{nt}^{final}.

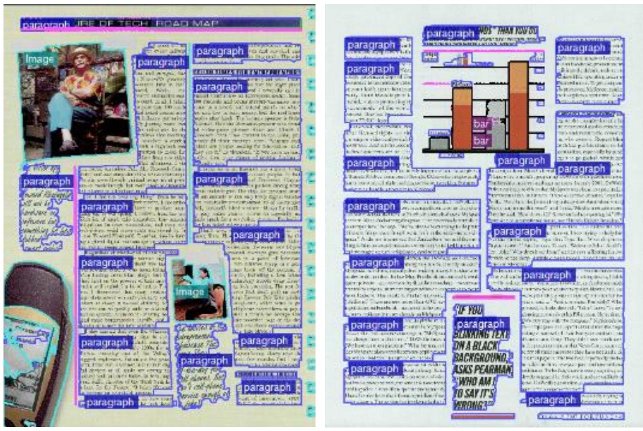
We further classify the CCs in I_{nt}^{final} into table, bar chart, image, and inverted text. From I_t^l we identify and separate the drop capital, if any. Apart from that, we perform stroke-width analysis on the components of I_t^s to generate I_{thick} and I_{thin}. After that we form the regions from I_t^l , I_{thick} and I_{thin}separately. For this we employ an iterative and adaptive morphology based approach using a rotating structuring element. We then refine all the regions to get the closest polygon and the paragraph separator. At the final step we combine all the segmented text images to get I_t^{final}.

However, if we find the input image as a geometrically simple one, then we avoid the component partition step, rather we directly perform text non-text separation on the I_b' to get I_t^s and I_{nt}^{final}. We perform rest of the operations on these images as mentioned earlier.

The recently evaluated performance of BINYAS on RDCL2017 dataset [8] is given in Table I. The outputs of BINYAS on some samples taken from RDCL2017 dataset are given in Figure 1.

TABLE I. EVALUATION OF BINYAS ON RDCL2017 DATASET

Method	Accuracy (in %)		
	Segmentation	*Segmentation+ Classification*	*Text region only*
Tesseract	75.83	75.83	75.83
ABBYY	83.87	83.87	83.87
LIPADE	81.15	81.15	81.15
MHS2017	92.32	92.32	92.32
CVML	83.96	83.96	83.96
AOSM	82.75	82.75	82.75
JU_AEGEAN	76.31	76.31	76.31
BINYAS	**93.22**	**92.33**	**93.66**



(a) (b)

Figure 1. Output of BINYAS on some samples of RDCL2017 dataset. Here the regions with blue outline are texts, with magenta outline are separators, with brown outline are tables, with black outline are bar charts, with green outline are graphics and with mint outline are images.

D. *List of Publications*

- **S. Bhowmik**, R. Sarkar, B. Das, and D. Doermann, “GiB: a Game theory Inspired Binarization technique for degraded document images,” *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1443–1455, 2019.
- B. Das, **S. Bhowmik** , A. Saha, R. sarkar, “An Adaptive Foreground-Background Separation Method for Effective Binarization of Document Images,” *8th Int. Conf. Soft Comput. Pattern Recognit.*, 2016.
- **S. Bhowmik**, R. Sarkar, M. Nasipuri, and D. Doermann, “Text and non-text separation in offline document images: a survey,” *Int. J. Doc. Anal. Recognit.*, vol. 21, no. 1–2, pp. 1–20, 2018.
- **S. Bhowmik**, R. Sarkar, and M. Nasipuri, “Text and Non-text Separation in Handwritten Document Images Using Local Binary Pattern Operator,” in *Proceedings of the First International Conference on Intelligent Computing and Communication*, 2017, pp. 507–515.
- **S. Bhowmik**, S. Kundu, B. K. De, R. Sarkar, and M. Nasipuri, “A Two-Stage Approach for Text and Non-text Separation from Handwritten Scientific Document Images,” *in Information Technology and Applied Mathematics*, *Springer*, 2019, pp. 41–51.
- A. K. Sah, **S. Bhowmik**, S. Malakar, R. Sarkar, E. Kavallieratou, and N. Vasilopoulos, “Text and non-text recognition using modified HOG descriptor,” in *Calcutta Conference (CALCON), 2017 IEEE*, 2017, pp. 64–68.

- S. Ghosh, D. Lahiri, S. Bhowmik, E. Kavallieratou, and R. Sarkar, “Text/Non-Text Separation from Handwritten Document Images Using LBP Based Features: An Empirical Study,” *J. Imaging*, vol. 4, no. 4, p. 57, 2018.

III. FUTURE PLAN

It has been mentioned earlier that the method is still under process. Our ultimate plan is to develop an integrated DLA system, which not only will produce suitable outcome for printed documents but also it will perform well for unconstrained handwritten documents. Our current objectives are listed as follows:

- Designing a comprehensive table detection technique.
- Designing a technique which can handle complex color background.

REFERENCES

[1] L. O’Gorman, “The document spectrum for page layout analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993

[2] J.-L. Meunier, “Optimized xy-cut for determining a page reading order,” in Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, 2005, pp. 347–351.

[3] K. Kise, “Page segmentation techniques in document analysis,” in Handbook of Document Image Processing and Recognition, Springer, 2014, pp. 135–175.

[4] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, “A comprehensive survey of mostly textual document segmentation algorithms since 2008,” Pattern Recognit., vol. 64, pp. 1–14, 2017.

[5] N. Vasilopoulos and E. Kavallieratou, “Complex layout analysis based on contour classification and morphological operations,” Eng. Appl. Artif. Intell., vol. 65, pp. 220–229, 2017.

[6] T. A. Tran, I. S. Na, and S. H. Kim, “Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology,” Int. J. Doc. Anal. Recognit., vol. 19, no. 3, pp. 191–209, 2016.

[7] C. Reul, U. Springmann, and F. Puppe, “LAREX: A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books,” in Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, 2017, pp. 137–142.

[8] C. Clausner, A. Antonacopoulos, and S. Pletschacher, “ICDAR2017 Competition on Recognition of Documents with Complex Layouts-RDCL2017,” in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, 2017, vol. 1, pp. 1404–1410.

[9] B. Das, S. Bhowmik , A. Saha, R. sarkar, “An Adaptive Foreground-Background Separation Method for Effective Binarization of Document Images,” 8th Int. Conf. Soft Comput. Pattern Recognit., 2016.

[10] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, “GiB: a Game theory Inspired Binarization technique for degraded document images,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 1443–1455, 2019.

[11] S. Bhowmik, R. Sarkar, M. Nasipuri, and D. Doermann, “Text and non-text separation in offline document images: a survey,” Int. J. Doc. Anal. Recognit., vol. 21, no. 1–2, pp. 1–20, 2018.

[12] A. K. Sah, S. Bhowmik, S. Malakar, R. Sarkar, E. Kavallieratou, and N. Vasilopoulos, “Text and non-text recognition using modified HOG descriptor,” in Calcutta Conference (CALCON), 2017 IEEE, 2017, pp. 64–68.

[13] S. Bhowmik, R. Sarkar, and M. Nasipuri, “Text and Non-text Separation in Handwritten Document Images Using Local Binary Pattern Operator,” in Proceedings of the First International Conference on Intelligent Computing and Communication, 2017, pp. 507–515.

[14] S. Bhowmik, S. Kundu, B. K. De, R. Sarkar, and M. Nasipuri, “A Two-Stage Approach for Text and Non-text Separation from Handwritten Scientific Document Images,” in Information Technology and Applied Mathematics, Springer, 2019, pp. 41–51.

[15] S. Ghosh, D. Lahiri, S. Bhowmik, E. Kavallieratou, and R. Sarkar, “Text/Non-Text Separation from Handwritten Document Images Using LBP Based Features: An Empirical Study,” J. Imaging, vol. 4, no. 4, p. 57, 2018

SHOWMIK BHOWMIK

Current Status	:	Fulltime PhD Scholar in CSE Dept., Jadavpur University (**Visvesvaraya Fellow**)
Date of Birth	:	22-08-1986
Gender	:	Male
Email and Contact no.	:	showmik.cse@gmail.com, Mob. +919475528462

1. Education

- Pursuing **PhD** in Department of Computer Science & Engineering, Jadavpur University **since February 2016.**
- **M.E.** in Computer Science & Engineering from Jadavpur University in **2014** with 1st Class.
- **B.Tech** in Computer Science & Engineering from West Bengal University of Technology in **2008** with 1st Class.
- **Higher Secondery** from WBCHSE in **2004** with 1st Division.
- **Secondary** from WBSE in **2002** with 1st Dividion.

2. Award

- Selected as a research fellow under **Visvesvaraya PhD Scheme** by **Electronics & IT under Ministry of Electronics and Information Technology, Government of India.**

3. Academic Experience

- ***Fulltime PhD Scholar*** in the Department fo Computer Sc. and Engg., Jadavpur University, West Bengal, India, ***Feb 23, 2016 - till date.***
- ***Assistant Professor*** in Computer Sc. and Engg., Dumkal Institute of Engineering and Technology, West Bengal, India, ***Feb 01, 2010- Feb 22, 2016.***
- ***Technical Assistant*** in Computer Sc. and Engg., Dumkal Institute of Engineering and Technology, West Bengal, India**,** ***Nov 04, 2008 – Jan 31, 2010***

4. International Competition

- "*RDCL2017: ICDAR Competition on Recognition of Documents with Complex Layouts (Continuous)", "***BINYAS: A Complex Document Layout Analysis System" (secured 2nd position out of 9 methods**).
- *"ICDAR2017 Competition on Document Image Binarization (DIBCO 2017)",* (**secured 16th position out of 26 methods**)

5. Publication

Journal

1. **S. Bhowmik,** R. Sarkar, B. Das, D. Doermann, " GiB: A Game Theory Inspired Binarization Technique for Degraded Document Images". *IEEE Transactions on Image Processing*, 28(3), 1443-1455, **2019 [Impact Factor 6.79]**
2. S. Malakar, M. Ghosh, **S. Bhowmik,** R. Sarkar, M. Nasipuri, "A GA based Hierarchical Feature Selection Approach for Handwritten Word Recognition", *Neural Computing and Applications, Springer*, **2019. (DOI:** https://doi.org/10.1007/s00521-018-3937-8 **) [Impact Factor 4.664]**
3. **S. Bhowmik,** R. Sarkar, M. Nasipuri, D. Doermann, "Text and Non-text Separation in Offline Document Images: a Survey", *International Journal on Document Analysis and Recognition (IJDAR*), Springer, *21*(1-2), 1-20, **2018** [**Impact Factor 0.846]**
4. **S. Bhowmik**, S. Malakar, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, "Off-line Bangla Handwritten Word Recognition: a Holistic Approach", *Neural Computing and Applications, Springer*, **2018. (DOI:** https://doi.org/10.1007/s00521-018-3389-1**) [Impact Factor 4.664]**
5. S. Sahoo, S. K. Nandi, S. Barua, P. Priyam, **S. Bhowmik,** S. Malakar, R. Sarkar, "Handwritten Bangla Word Recognition using Negative Refraction based Shape Transformation", Special Issue on Ambient Advancements in Intelligent Computational Sciences for *Journal of Intelligent & Fuzzy Systems- Applications in Engineering and Technology, IOS Press*, 35(2), pp. 1765-1777, **2018. [Impact Factor 1.637]**
6. S. Ghosh,., D. Lahiri, **S. Bhowmik,** E. Kavallieratou R. Sarkar, "Text/Non-Text Separation from Handwritten Document Images Using LBP Based Features: An Empirical Study", *Journal of Imaging*, 4(4), 57, (**2018**).

7. **S. Bhowmik, S.** Polley, M. G. Roushan, S. Malakar, R. Sarkar, M. Nasipuri, "A Holistic Word Recognition Technique for Handwritten Bangla Words", *Int. J. Applied Pattern Recognition*, 2(2), pp.142–159, (**2015**).

Book Chapter

1. Ghosh M., Malakar S., **Bhowmik S**., Sarkar R., Nasipuri M. (2019) Feature Selection for Handwritten Word Recognition Using Memetic Algorithm. In: Mandal J., Dutta P., Mukhopadhyay S. (eds) Advances in Intelligent Computing. Studies in Computational Intelligence, vol 687. pp. 103-124, Springer, Singapore

Conference

1. Ghosh, S., Ghosh, K.K., Chakraborty, S., **Bhowmik, S.**, Sarkar, R. (2019) A Filter Ensemble Feature Selection Method for Handwritten Numeral Recognition, International Conference on Emerging Technologies for Sustainable Development (ICETSD '19), pp. 394-398.
2. Ghosh, S., Bhattacharya, R., Majhi, **S., Bhowmik**, S., Malakar, S., Sarkar, R., (2018) Textual content retrieval from Filled-in Form Images, 4th Workshop on Document Analysis and Recognition (DAR), 18 December 2018, IIIT Hyderabad, India, doi: https://doi.org/10.1007/978-981-13-9361-7_3.
3. Sah, A. K., **Bhowmik, S.,** Malakar, S., Sarkar, R., Kavallieratou, E., & Vasilopoulos, N. (2017). Text and non-text recognition using modified HOG descriptor. In Calcutta Conference (CALCON), (pp. 64-68). IEEE.
4. Ghosh, M., Malakar, **S., Bhowmik**, S., Sarkar, R., Nasipuri, M., (2017) Memetic Algorithm based Feature Selection for handwritten City Name Recognition, In: Mandal J., Dutta P., Mukhopadhyay S. (eds) Computational Intelligence, Communications, and Business Analytics. CICBA 2017. Communications in Computer and Information Science, vol. 776, pp.599-613, Springer, Singapore.
5. **Bhowmik, S**., Kundu, S., De, B.K., Sarkar, R., Nasipuri, M., (2019) A two-stage approach for Text and Non-text Separation from Handwritten Scientific Document Images", In: Chandra P., Giri D., Li F., Kar S., Jana D. (eds) Information Technology and Applied Mathematics. Advances in Intelligent Systems and Computing, vol 699. pp. 41-51, Springer, Singapore,.
6. **Bhowmik, S**., Sen, S., Hori, N., Sarkar, R., Nasipuri, M. (2017). Handwritten Devanagari Numerals Recognition using Grid based Hausdroff Distance. In Computer, Communication and Electrical Technology: Proceedings of the International Conference on Advancement of Computer Communication and Electrical Technology (ACCET 2016), (p. 15). CRC Press.
7. Barua, S., Malakar, S., **Bhowmik, S**., Sarkar, R., Nasipuri, M. (2017). Bangla Handwritten City Name Recognition Using Gradient-Based Feature. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 343-352). Springer, Singapore.
8. Das, B**., Bhowmik, S**., Saha, A., Sarkar, R. (2018). An Adaptive Foreground-Background Separation Method for Effective Binarization of Document Images. In: Abraham A., Cherukuri A., Madureira A., Muda A. (eds) Eighth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016). Advances in Intelligent Systems and Computing, vol 614. Springer, Cham (pp. 515-524).
9. **Bhowmik, S.,** Sarkar, R. Nasipuri, M. (2016) Text and Non-text separation in Handwritten Document Images using Local Binary Pattern Operator, In: Mandal J., Satapathy S., Sanyal M., Bhateja V. (eds) Proceedings of the First International Conference on Intelligent Computing and Communication. Advances in Intelligent Systems and Computing, vol. 458. pp. 507-515, Springer, Singapore.
10. Singh, P. K., Mondal, A., **Bhowmik, S.,** Sarkar, R., Nasipuri, M. (2015) Word-Level Script Identification from Handwritten Multi-script Documents. In: Satapathy S., Biswal B., Udgata S., Mandal J. (eds) 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Advances in Intelligent Systems and Computing, vol. 327. Springer, Cham (pp. 551-558)..
11. **Bhowmik, S**., Polley, S.,Roushan, M.G., Malakar, S., Sarkar, R. Nasipuri, M, (2014) Handwritten Bangla Word Recognition using HOG Descriptor, In Emerging Applications of Information Technology (EAIT-2014), 4th International Conference on (pp. 193-197) IEEE.
12. **Bhowmik, S**., Malakar, S., Sarkar, R. Nasipuri, M, (2014) Handwritten Bangla Word Recognition using Elliptical Features, In Computational Intelligence and Communication Networks (ICCICN) 2014, The 6th International Conference on (pp. 257-261) IEEE.

Training Data and Time Efficient Algorithms for Historical Document Analysis

Florian Westphal

Supervisors of the thesis: Håkan Grahn, Niklas Lavesson
University: Blekinge Institute of Technology, Sweden
Starting date of the PhD: 14.01.2015
Expected finalization data of the PhD: 31.01.2020
Email: florian.westphal@bth.se

Abstract. Over the last decades companies and government institutions have gathered vast collections of images of historical handwritten documents. In order to make these collections truly useful to the broader public, images suffering from degradations, such as faded ink, bleed through or stains, need to be made readable and the collections as a whole need to be made searchable. Developing algorithms which support this through image binarization or word spotting, and achieve reasonable performance, is a difficult task. Additionally, these algorithms need to execute fast enough to be able to process vast collections of images in a reasonable amount of time, and to be able to deal with the limited access to labelled training data specific to the target image collection.

Based on this motivation, the main aim of my thesis is to identify different techniques, which reduce the execution time and the amount of training data required for performing document analysis tasks. While, I have mostly focused on document image binarization, my focus has shifted further to character recognition and word spotting. For image binarization, I have explored heterogeneous computing, parameter tuning and architectural changes to achieve shorter execution times. Furthermore, I have explored two potential guided machine learning approaches for reducing the required amount of training data, namely human based training sample selection, for binarization, and learning using privileged information (LUPI), for character recognition. In future work, I am planning to use LUPI for word spotting.

1 Introduction

It has never been easier to access historical documents than now, since different companies and government institutions provide access to high resolution color images of historical documents to a broad public via the Internet. Different document image analysis techniques aim to further improve this access by means of image binarization to make images more readable, or word spotting to make them searchable. However, developing these techniques is a challenging task, due to common degradations in historical documents, such as faded ink, bleed through or stains, as well as general irregularities in those documents.

Furthermore, it is not enough for algorithms to produce reasonable results in image binarization or word spotting. In order to be truly useful when dealing with vast collections of document images, these algorithms need to perform fast to be able to process these collections in a reasonable amount of time. Another challenge for the application of document image analysis techniques in practice is that most used algorithms, especially those based on machine learning, require ground truth samples from the target image collection to tune the algorithm's parameters or train the algorithm to perform well on this target collection. However, those ground truth samples are generally not available and costly to acquire which limits or prevents the application of those algorithms in practice. In my research, I am exploring different techniques to address these two challenges, i.e., improved execution performance and dealing with the limited availability of training data for specific image collections.

1.1 Methodology

Execution Time Efficiency For studying possible ways to speed up document analysis algorithms, I have chosen the task of image binarization. Image binarization is interesting in this context, since it is a common pre-processing step for many document analysis algorithms. As such, it should take only a short amount of time to not prolong the processing time of those algorithms needlessly. However, the binarization quality must be reasonably high to not influence the following algorithms negatively.

In a first study [7], I have focused on improving the execution performance of Howe's binarization algorithm (HBA) [3]. This has been done by splitting the algorithm into three parts and finding the mapping of these parts to CPU and GPU, which yields the best execution performance. Another approach taken in this study was to replace Howe's parameter tuning algorithm for HBA [4] with random forest based multivariate regression to predict suitable parameters. This study has shown that mapping all parts of HBA to the GPU results in a 3.5 times faster execution time compared to mapping all parts to the CPU. Furthermore, the parameter prediction results in a 2.5 times faster processing for large images compared to Howe's original algorithm.

In a second study [9], I have proposed an approach for image binarization, which is similar to the approach by Afzal et al. [1] in that it uses recurrent neural networks (RNNs). With respect to time efficiency, this study investigated the impact of the network's footprint size on the execution performance and found that a footprint size of 4×4 results in the best trade-off between execution time and binarization quality. This study proposed also a dynamically weighted loss function to penalize binarization errors more, which affect the readability.

Training Data Efficiency In order to reduce the required amount of training data for document analysis tasks, I focus on guided machine learning (gML), also known as interactive machine learning (iML) or human-in-the-loop approaches [2]. In particular, I have explored user based sample selection and learning using privileged information (LUPI) [6] as possible gML mechanisms.

User based sample selection has been applied to the task of document image binarization [8]. In this study, users were shown the current binarization result of an RNN based binarization algorithm and could select the parts of the document image, which should be used to re-train the model. This selection could be aided by an additional visualization of the model's uncertainty, as defined by the difference between the predicted value and the chosen label value. For example, a predicted value of 0.6 for one pixel would result in the assignment of the label 1 to the pixel, which would lead to an uncertainty value of 0.4. This study has shown that samples chosen by users based on perceived readability issues and the visualized uncertainty result in better training results than random sample selection or user selection based on readability alone.

The LUPI framework has been used to develop one approach to reduce the amount of required training data for character recognition. Based on the idea of privileged information, i.e., instance representations, which are only available during training time, but not during test time, I have proposed the use of graph representations of characters to train a convolutional neural network (CNN) based character recognizer [11]. The proposed approach trains a Siamese network to predict the graph edit distance between two character graphs. Since this training requires a large amount of character graphs, it makes this type of privileged information unsuitable for gML. However, since the character graphs and the graph edit distance can also be obtained automatically, it is possible to use this information for pre-training. After the pre-training, the network can be trained to recognize specific characters using a few labeled samples. This study has shown that the proposed approach performs better than standard supervised learning and better than graph matching if only few training samples are provided. Furthermore, it performs as well as standard supervised training when a sufficient amount of training data is available.

In preparation of a larger literature review on different iML techniques, which will not be part of my thesis, I have reviewed current definitions of iML. I have argued, in a position paper [10], that the currently used definitions are ambiguous and that the term *interactive machine learning* is unintentionally broad. Therefore, I have proposed to use the term *guided machine learning* (gML) instead and have proposed a suitable definition.

1.2 Future Work

As last study of my thesis, I am planning to extend the study on graph based pre-training for character recognition to word spotting by using the graph edit distance to pre-train the CNN base of a PHOCNet [5]. The aim of this study is to combine execution time efficiency with help of the techniques studied for image binarization with training data efficiency using a LUPI based approach. Furthermore, I am planning to explore possible ways to make it feasible for users to provide the privileged information in form of graphs.

References

1. Afzal, M.Z., Pastor-Pellicer, J., Shafait, F., Breuel, T.M., Dengel, A., Liwicki, M.: Document Image Binarization using LSTM: A Sequence Learning Approach. In: Proc. of the 3rd Int. Workshop on Historical Document Imaging and Processing. pp. 79–84. ACM (2015)
2. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016)
3. Howe, N.R.: A Laplacian Energy for Document Binarization. In: 11th International Conference on Document Analysis and Recognition (ICDAR). pp. 6–10. IEEE (2011)
4. Howe, N.R.: Document binarization with automatic parameter tuning. International Journal on Document Analysis and Recognition (IJDAR) **16**(3), 247–258 (2013)
5. Sudholt, S., Fink, G.A.: Attribute cnns for word spotting in handwritten documents. International journal on document analysis and recognition (ijdar) **21**(3), 199–218 (2018)
6. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)
7. Westphal, F., Grahn, H., Lavesson, N.: Efficient document image binarization using heterogeneous computing and parameter tuning. International Journal on Document Analysis and Recognition (IJDAR) **21**(1-2), 41 – 58 (2018)
8. Westphal, F., Grahn, H., Lavesson, N.: User feedback and uncertainty in user guided binarization. In: International Conference on Data Mining Workshops (ICDMW). pp. 403–410. IEEE (2018)
9. Westphal, F., Lavesson, N., Grahn, H.: Document image binarization using recurrent neural networks. In: 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 263–268. IEEE (2018)
10. Westphal, F., Lavesson, N., Grahn, H.: A case for guided machine learning. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 353–361. Springer International Publishing, Cham (2019)
11. Westphal, F., Lavesson, N., Grahn, H.: Learning character recognition with graph-based privileged information. In: 15th International Conference on Document Analysis and Recognition (ICDAR). IEEE (2019), to appear

A Novel HL-OSV Dataset and Depthwise Separable CNN based Online Signature Verification

Student's name: Chandra Sekhar Vorugunti[1]
Supervisors of the thesis: Prof. Viswanath Pulabaigari[1] & Prof. Prerana Mukherjee[1]
1 Indian Institute of Information Technology-SriCity, Chittoor-Dt, A.P, India.
Ph.D. Thesis will be submitted at: Dept of CSE, IIIT SriCity, India- 517 646.
Starting date of the Ph.D.: 03th August, 2016
Expected finalization date of the Ph.D.: April, 2020
Chandrasekhar.v@iiits.in

Abstract. This paper briefly explains my Ph.D. research direction towards Online Signature Verification (OSV) is a real time challenging problem which is used across domains, e.g., online banking, m-payment, etc. Hence, our contribution is twofold. One: The current OSV datasets provide only basic information like x, y co-ordinates, pressure, azimuthal angle at each sampling point of online signature. This basic information doesn't deliver a deeper understanding of the user signatures, e.g. variations in user signing velocity, pressure etc. between the sampling points. To address this, we have developed a novel dataset named HL-OSV with 44800 user profiles (2D images) for OSV by computing 28 higher level features. Second, in the literature, the OSV frameworks proposed based on traditional machine learning or deep learning algorithms requires a minimum number of signature samples for training to achieve a reasonable amount of classification accuracy. Acquiring minimum number of samples is not feasible in all the scenarios, e.g. online e-commerce transactions, etc. Thorough experimental analysis confirms that the combination of the HL-OSV dataset and a DWSCNN OSV framework achieves few shot learning and superior classification accuracies compared to state-of-the art OSV frameworks.

Keywords: Few shot learning; Domain Adaptation; Online Signature Verification; Depthwise Separable Convolution; User profiles.

1 Introduction and Dataset Preparation

Our proposed framework and the HL-OSV dataset is prepared on top of SVC dataset [1], a widely used online signature dataset. The details of the SVC dataset are illustrated in Table 1. In svc dataset, an online signature is sampled along the online trace of a signature. Each discrete sampling point is a collection of attributes such as x and y co-ordinates, pressure p, azimuth φ and inclination angle θ, which are the time ordered. The features at the pth sampling point is represented as {x(p), y(p), P(p), φ(p), θ (p)}. Generally, for each signature, the number of sampling points varies between 80 to 90. Based on these marginal and low-level dynamic information of the signature trace, a deeper and thorough signature analysis is not feasible. Hence, for deeper analysis, we have computed high level features which are first order differences of the base features. We performed below steps to generate our proposed HL-OSV dataset from SVC dataset.

Step 1) for each user, for each signature of the base dataset SVC, the x, y co-ordinates $x(i), y(i)$ are projected into 2D plane, where $1 \leq i \leq 80$.

Step 2) for each user, for each signature of the base dataset SVC, the pressure $P(i)$ values at each sampling point $(x(i), y(i))$ are projected into 2D plane, where $1 \leq i \leq 80$.

Step 3) for each user, for each signature of the base dataset SVC, we have computed first order difference between sampling points: For 'n' sampling points S.P $= 1, 2, \ldots, n$, we have computed first order differences as follows:

$$\Delta x_s(i) = x(i+s) - x(i),$$
$$\Delta y_s(i) = y(i+s) - y(i),$$
$$V_s(i) = \sqrt{\Delta x_s(i)^2 + \Delta y_s(i)^2},$$
$$A_s(i) = \sqrt{\Delta V x_s(i)^2 + \Delta V y_s(i)^2}$$
$$\text{where } 1 \leq s \leq 6, 1 \leq i \leq n - s$$

The values $V_s(i), A_s(i)$ are projected into 2D plane to convert the numerical values into images of user profiles in 2D space.

Step 4) We have computed the hybrid profiles by combining {P+V_s}, {P+A_s} where s = 1 to 6. The details about the base dataset, i.e. SVC, the developed dataset HL-OSV are illustrated in Table 1 and II. Few example signatures of our dataset are illustrated in Fig 1 and 2.

2 Proposed OSV Framework Architecture

Here we present our framework for online signature classification.

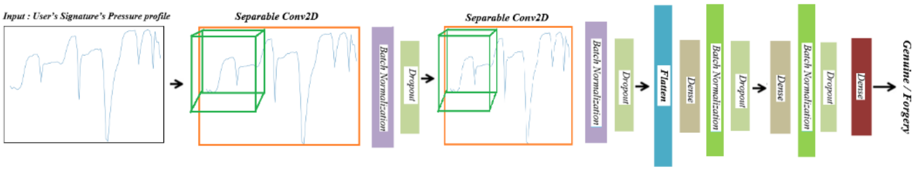


Fig. 1. Overview of the Proposed separableConv2D based OSV framework architecture used in this work.

3 Experimentation and Results

We have conducted our experiments on Nvidia, Titan X Pascal 20 GB GPU. Due to space limitations, we briefly discuss the experimentation analysis. A brief description of the experimentation results is illustrated below.

Method	S_01	S_05	S_10	S_15	R_01	R_05	R_10	R_15
Pressure	18.60	16.19	15.25	12.8	12.21	9.94	7.45	4.52
Velocity : 1 step (V1)	18.73	19.16	16.45	13.6	11.71	10.26	7.73	4.43
Acceleration : 1 step (A1)	18.37	17.78	14.09	12.82	12.05	10.32	7.62	4.21
Pressure +V1	17.97	17.30	14.90	13.60	12.14	10.76	7.84	4.25
Pressure +A1	17.87	18.37	15.65	14.70	12.31	10.65	8.04	4.39
x, y co-ordinates in 2D	15.92	15.17	13.40	12.30	11.47	10.02	7.41	4.20
Only X-co-ordinate	**14.10***	14.8	13	12.2	11.58	9.81	7.61	4.25
Only Y-co-ordinate	15.18	**13.5****	11.55	**10.8****	11.21	9.68	**7.28****	**4.14****
DTW based [1]	-	-	7.80	-	-	-	-	-
Stroke Point Warping [2]	-	-	**1.00***	-		-	-	-
SPW+mRMR+SVM(10-Samples) [2]	-	-	**1.00***	-	-	-	-	-
Variance selection [3]	-	-	13.75	-	-	-	-	-
Relief-1 [3]	-	-	8.1	-	-	-	-	-
PCA [3]	-	-	7.05	-	-	-	-	-
Relief-2 [3]	-	-	**5.31****	-	-	-	-	-
Probabilistic-DTW(case 1)[4]	-		-	-	-	**0.0025***	-	-
Probabilistic-DTW(case 2)[4]	-	-	-	-	-	**0.0175****	-	-
Target-Wise [5]	18.63	-	-	-	**0.50***	-	-	-
Stroke-Wise [5]	**18.25****	-	-	-	**1.90****	-	-	-

Future Work : **Synthetic Signature Generation** :To solve the above problem of signature generation by GAN, it requires large number of training samples for efficient signature generation. Our goal is to generate synthetic versions of a dataset with binary classes using different variations of GAN architectures. We use all the public datasets with different GAN models to generate its synthetic copies. We will show that using GANs it is possible to produce synthetic copies of our chosen dataset that are close to the original dataset for a variety of real-world applications. Finally, the real and synthetic online signatures are used to train the GAN model and used for test signature interpretability.

4 Conclusion and Future Scope

We identified several drawbacks in existing OSV datasets, in contrast to existing datasets, we have developed an online signature dataset named HL-OSV, considering 28 higher level features from the low-level features and projecting these low-level features into 2D plane. Our dataset consists of 44800 2D images, representing 28 profiles of 40 users.

REFERENCES

1. Sharma, A., Sundaram, S.: 'On the Exploration of Information From the DTW Cost Matrix for Online Signature Verification', IEEE Transactions on Cybernetics, vol 48, (feb. 2018).
2. Kar, B., Mukherjee, A., Dutta, P.K.: Stroke Point Warping-Based Reference Selection and Verification of Online Signature, IEEE Transactions On Instrumentation and Measurement, vol. 67, (Jan 2018).
3. Yang, L., Cheng, Y., Wang, X., Liu, Q.: Online handwritten signature verification using feature weighting algorithm relief, Soft Computing, Vol 22, (December 2018).
4. Al-Hmouz, R., Pedrycz, W., Daqrouq, K., Morfeq, A., Al-Hmouz, A.: Quantifying dynamic time warping distance using probabilistic model in verification of dynamic signatures, Elsevier-Soft Computing, vol 23, vol 23, pp 407–418, (Jan 2019).
5. Diaz, M., Fischer, A., Ferrer, M.A., Plamondon, R.: Dynamic Signature Verification System Based on One Real Signature, IEEE Transactions On Cybernetics, vol 48, (Jan 2018).

Detection and analysis of weak signals. Development of a digital investigation framework for a hidden alert launcher service

Julien Maitre
Advisor : Michel Ménard
Start: Oct 1, 2016 - End: Sept 30, 2019
L3i, Faculty of Science and Technology
La Rochelle University
La Rochelle, France
julien.maitre@univ-lr.fr

I. Research statement

A. Problematic

In the context of information mass explosion, the detection of *weak signals* has become an important tool for decision makers. *Weak signals* are the precursors of future events. Ansoff proposed the concept of *weak signal* in a strategic planning objective through environmental analysis. Typical examples of *weak signals* are associated with technological developments, demographic and environmental change, etc. We also find examples of weak signals in whistleblower revelations that represent our study context : detecting weak signal and issuing alerts. The information carried by the latter will have to be correlated with a broader informational context through exploration phases on the networks.

Our goal is therefore the detection of precursor signals whose contiguous presence in a given space of time and places anticipates the occurrence of an observable fact. This detection is facilitated by the early information provided by a whistleblower in form of documents which expose proven, unitary and targeted facts but also partial and relating to a triggering event.

We therefore retain for our study several qualifiers for *weak signals*. We propose the definition below.

Definition. A *weak signal* is characterized by a low number of words per document and in a few documents (rarity, abnormality). It is revealed by a collection of words belonging to a same and single theme (unitary, semantically related), not related to other existing themes (to other paradigms), and appearing in similar contexts (dependence).

B. Plan

For the development of the investigation framework, three actions are therefore undertaken:

- Action 1: Automatic content analysis with minimal *a priori* information. Identification of relevant information. Indicator of coherence of the obtained themes;
- Action 2: Aggregation of knowledge. Information enrichment. Detection of *weak signals*;
- Action 3: Analytical visualization. Putting information into perspective by creating visual representations and dynamic dashboards.

II. Progress to date

This is therefore a difficult problem since the themes carried by the documents are unknown and the collection of words that make up these themes also. In addition to these difficulties of constructing document classes in an unsupervised manner, there is the difficulty of identifying, via the collections of words that reveal it, the theme related to the weak signal. The analysis must therefore simultaneously make it possible to: 1) discover the themes, 2) classify the documents in relation to the themes, 3) detect relevant keywords related to themes, and finally, 4) it's the main purpose of the study, to discover the keywords related to a *weak signal* theme possibly present. Figure 1 illustrates the processing chain. We focus our work on the multidimensional clustering problem.

In previous work, we proposed an approach to searching for common topics in a corpus of documents and detecting a topic related to a *weak signal* characterized by a small number of words per document and present in few documents. The combination *LDA / Word2Vec* as we proposed to implement it allows us to free ourselves from the arbitrary choice of the K parameter (number of clusters) during partitioning. Two directions were explored: 1) the first algorithm aims to find the number of topics leading to a partitioning by *LDA* as consistent as possible; 2) the second algorithm which, in a more advanced way, combines the best topics returned by *LDA* on the whole tree structure built when K is varied.

III. Future Work

The detected words of the topic "*weak signal*" will be used in a further phase. The idea is to feed, with those keywords a attracting/repulsing-based multi-agent system able 1) to augment by quering the networks, 2) to reorganize documents and clusters on the fly and 3) to offer both

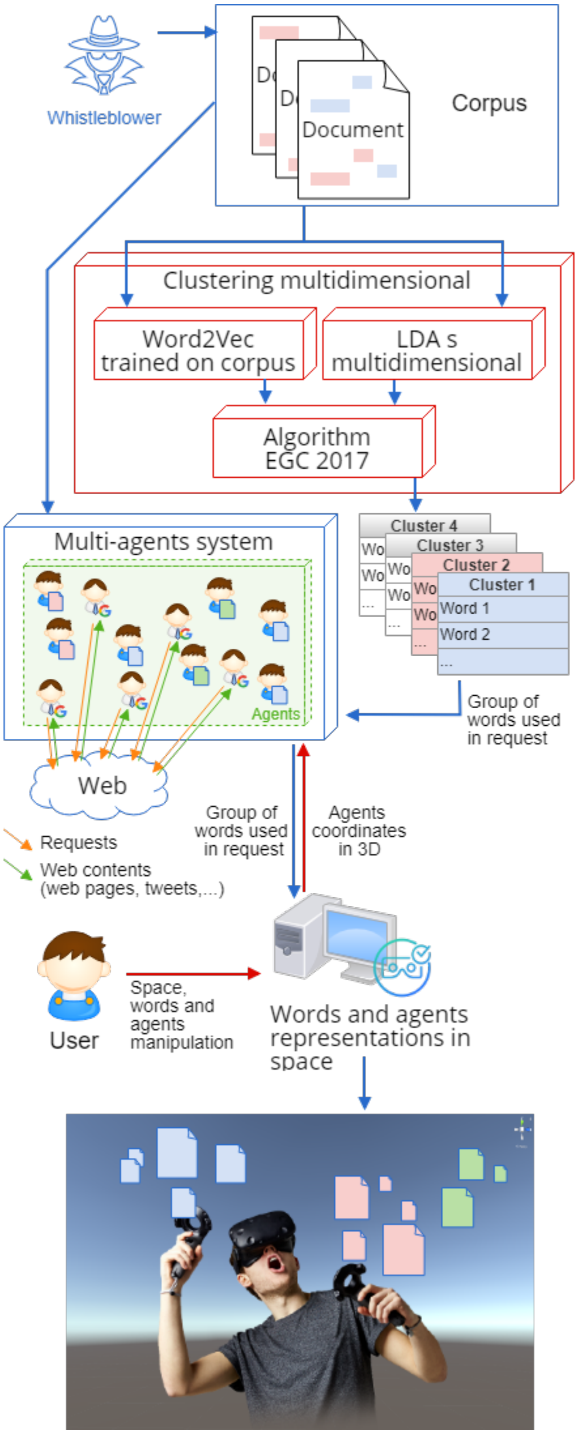

Whistleblower
Corpus
Document
Clustering multidimensional
Word2Vec trained on corpus
LDA s multidimensional
Algorithm EGC 2017
Multi-agents system
Agents
Cluster 4
Cluster 3
Cluster 2
Cluster 1
Word 1
Word 2
...
Group of words used in request
Web
Requests
Web contents (web pages, tweets,...)
Group of words used in request
Agents coordinates in 3D
User
Space, words and agents manipulation
Words and agents representations in space

Figure 1. The system automatically extracts and analyzes the information provided by the whistleblower. The system builds indicators that are put in dashboards for recipients who can also visualize the dynamic evolution of information provided by a multi-agent environment system. This one is used for navigation and document retrieval. Each document is represented by an agent which moves in a 3D environment.

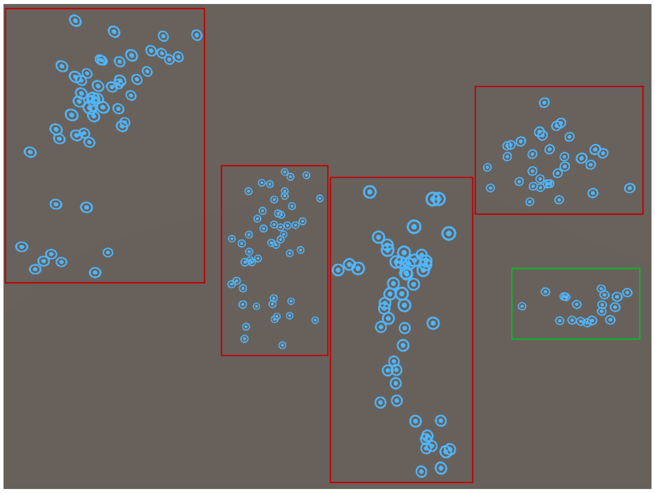


Figure 2. Experimentation of multi-agent system with documents represented by points in 3d space. The red box represents some main topics and green box the topic "*weak signal*"

advanced interaction and visualisation of the manipulated data. The Figure 2 shows our preliminary work on this idea. The objective is to continue searching for documents related to this topic, to increase the corpus of documents on this topic and to discover other related words. The methodological approach is intended to be consistent with that adopted, for example, by journalists, who first rely on unitary and targeted facts/documents, then attempt to consolidate them and assess their relevance by exploring other sources. These make it possible to open up to a broader informational context.

Figure 2 shows MAS in action which actively searches for new documents while it is spatially reorganizing the existing document agents into clusters. This model simplifies the problem of mapping a high-dimensional feature space onto a 3D space in order to facilitate the visualization and allows an intuitive user interaction. By forcing the position of agents in space, the agents become automatically some kind of query-agent, letting no choice to the others free agents to rearrange the positions around the fixed one(s).

References

[1] J. Maitre, M. Menard, G. Chiron, and A. Bouju, "Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique," in *Extraction et Gestion des Connaissances 2017*, Grenoble, France, Jan. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01449911

[2] J. Maitre, "A Wikipedia dataset of 5 categories," jun 2019. [Online]. Available: https://zenodo.org/record/3260046

[3] J. Maitre, M. Menard, G. Chiron, A. Bouju, and N. Sidere, "A meaningful information extraction system for interactive analysis of documents," in *2019 15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, Australia: IEEE, sep 2019.

Exploring Attention Models for Multimodal Description Tasks

Student name: Viviana Beltrán, University of La Rochelle
Collaborators of the thesis: Antoine Doucet, La Rochelle Université,
Nicholas Journet, Université Bordeaux, Mickael Coustaty, La Rochelle Université,
Juan Caicedo, Broad Institute of MIT and Harvard
Starting date of the PhD: May, 2018, Expected finalization date of the PhD: April, 2021
Email: viviana.beltran@univ-lr.fr

Abstract—**Attention has become an essential component of neural networks as it facilitates interpretability. We want to analyse the impact of using attention mechanisms at different configurations and stages of the learning process, and how these configurations affect the main fusion step between the networks representing different modalities. On account of this, we propose to work with a multimodal framework that learns combined visual and textual representations, within the novel task of Scene Text via Visual Question Answering (ST-VQA), which aims to recognize some target text present in wild pictures and that answer to a specific question. The first attention mechanism that we included is the bounding boxes containing the target visual text represented as an additional input that gives the framework an explicit way to learn the correct patterns, on the contrary to the state-of-the-art that includes the OCRs founded. The system may recognize some characters of the word, but failing in retrieving all of them, thus, our objective task predicts scores based on a n-gram representation for the target task, being more suitable than a traditional bag of words representation used in media description systems.**

I. Research plan

Attention has become an essential part of deep learning, and the way it is implemented affects the performance of the model. Our model consists of different networks or components in charge of specific tasks. For the visual modality, we tested well-known architectures, having better results with ResNET [1], where we take the features from the last hidden layer as input to the visual network. For the textual modality, we are evaluating context-free and contextual word embeddings methods to get input features for the textual network, including GloVe [2] and BERT [3], followed sometimes for a specific Long Short Term Memory (LSTM) network [4]. To fuse the networks from both modalities, we are using the functions of point-wise multiplication, concatenation and average between vectors from the last layer of each network (trained to have the same dimensionality) however, we are exploring better fusion techniques, as the current ones can leave out relevant information of each modality. After the fusion, we have a multimodal network in charge of learning a common space.

As we can use the framework to evaluate different tasks, we adapt the objective function as a prediction of scores over a set of candidate items that represent the objective task. We have evaluated two tasks, the first one, in the context of information retrieval, where we didn't include any attention mechanism, but we were able to analyse the components of the whole framework. The second application, ST-VQA represents the objective function as a prediction over a set of n-grams extracted from the set of training answers. Scene Text Recognition has been widely studied to extract the text in the image [5] using methods mainly based on OCR techniques, however, when the extraction of some target text is required via visual question answering, the system needs to include a mechanism of reasoning in which current models fail. This specific application has not received the required attention, because of the lack of databases targeting this specific task, as well as for the challenges imposed [6].

We addressed the problem by adapting our previous strategy for the Robust Reading Challenge on Scene Text Visual Question Answering (ST-VQA) [7]. There were 6 participant teams, leaving our method in a 4th position, with our performance very far from the best method [8], which includes attention mechanisms and a bigger ensemble of models. Our strategy implemented for the challenge does not include any attention mechanisms, which is one of the main reasons our model has a lower performance than the winner. Another reason is regarding the extraction of visual features, while we are using a ResNET as the unique method, the winner method implements an elaborated framework including a network for object recognition and bottom-up and top-down attention mechanisms.

As the principal focus is the design and implementation of attention over the multimodal network, we started by analysing an initial proposal: We train our model by filtering the data by those samples that have metadata regarding the bounding box containing the ground truth answer, and include the bounding box as another network and fuse it with the visual and textual networks, contrary of the state-of-the-art strategy of including OCR string information as an attention mechanism. The hypothesis behind is that the model needs to learn the correct patterns when searching for some target text, those patterns are included in the images that present enormous variability, and thus, if we explicitly send the target patterns, the model will be able to leverage them. For the target task, our objective function

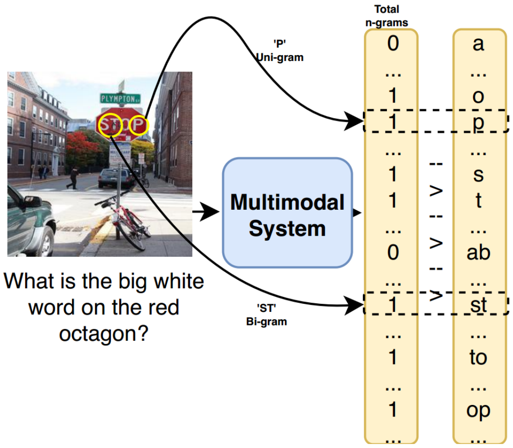


Figure 1. Overview of the n-gram representation for the answers in the database. The representation contains the uni-gram and bi-gram levels. 1 means the image contains the n-gram associated with the position in the vector representation, 0 denotes its absence.

predicts scores for representation based on n-grams extracted from the answers from training set 1, we believe this representation is more suitable because the system can learn to recognize at least at a character level, and therefore, the learned representation will be closed to the target one.

We are evaluating our methodologies in the new dataset, TextVQA, presented in the work of Singh et al. [6], that allow us to do a better comparison with the state-of-the-art for results reported over validation set. By taking into account the state-of-the-art result, we can see that the task is still very hard to solve, with the described strategies, we are having results of 20% vs 26% of accuracy reported in the paper [6]. Although further evaluation is required, clearly, our current methodology is naive to achieve higher results, if we compare it against the methodology proposed by [8], which elaborates in bottom-up and top-down attention mechanisms getting richer representations, specially in the visual modality.

II. Future work

Our focus is to determine better attention strategies used in the current state-of-the-art models to solve the task of ST-VQA. We want to analyse the impact of the design of attention over the multimodal component in a multimodal framework since the performance seems to increase. We are also evaluating fusion techniques that can be seen as a special attention from one modality to the other (currently, we use a point-wise multiplication). We want to analyse the 'direction' of attention in the problem of ST-VQA, is the performance affected if we apply the attention over the text instead of the images? are skip-connections special attention mechanisms studied mainly for the visual modality? and finally, taking into account these points, how to define attention in multimodal frameworks?. Another point to address is to determine better data augmentation strategies for different modalities that allow the model to learn more variations in the data (we have tested transformations to the images suitable for the target task of recognizing the text, thus, we apply transformations that did not change the order of the characters, however we did not have improvements over the performance than training the model using the original data).

III. Curriculum Vitae

A. *Education*

- B.Sc., Systems Engineering and Computing
  August 2008 – August 2014
  Universidad Nacional de Colombia
  Departament of systems and industrial Engineering, School of Engineering
  Bogotá D.C.
- M.Sc., Master Program in Systems, Engineering and Computing
  Aug 2014 – August 2016
  Universidad Nacional de Colombia
  Departament of systems and industrial Engineering, School of Engineering
  Bogotá D.C.
  Thesis: On-line Supervised Nonlinear Dimensionality Reduction
  Advisor: Fabio A. Gonzalez O., Ph.D
- Ph.D student in Computer Science
  May, 2018 Present
  La Rochelle Université,
  Laboratoire Informatique, Image et Interaction (L3i)
  La Rochelle, France

B. *Profesional experience*

- Java Developer from 2012 to 2014 (Academic experience): Object-Oriented Analysis, Design and Development, Relational Database Systems: Design and development of standalone Java information system applications.
- Research assistant from January of 2015 to November of 2015: Participation in the development of the prototype "MultiMedSearch" as part of the project "Multimodal image retrieval to support medical case-based scientific literature search".
- Researcher from August 2014 to December 2016 Modeling and implementation of machine learning algorithms using Python/GPU frameworks such as Theano and Pylearn2. Preprocessing and analysis of data using standard machine learning libraries such as scikit-learn, numpy, etc.
- Contractor at Instituto Colombiano para la Evaluacin de la Educacin ICFES from May 2016 to April 2018: I worked a system engineer, optimizing existing processes related with the different tests, databases ad-

ministration and designing new processes to improve efficiency in the company.
- Ph.D student at university of La Rochelle from 1st May to present.

C. International events

- Research visit: INAOE - Project ”Multimodal image retrieval to support medical case-based scientific literature search”. (Info in https://sites.google.com/site/mirmedicalsearch/) August (12-21/ 2015)
- Assistance to conference ICPRAM 2015 as a speaker of the paper: ”Two-way Multimodal Online Matrix Factorization for Multi-label Annotation”.
- Assistance to conference CIARP 2015 as a presenter of the poster for the paper: ”Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization ”.
- Observation visit as a part of the collaboration with the project: ”Observacion proyecto Diagnostico Bolivia - 2017” and ICFES.
- Symposium International Francophone sur l'Ecrit et le Document, SIFED - 2018 (Participant)
- Symposium International Francophone sur l'Ecrit et le Document, SIFED - 2019 (Oral presentation)

D. Publications

- Vanegas J., Beltrán V. and A. González F. (2015). Two-way Multimodal Online Matrix Factorization for Multi-label Annotation. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, pages 279-285. DOI: 10.5220/0005209602790285
- Beltrán V., Vanegas J., and A. González F. (2015). Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization. In Proceedings of Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings
- Pellegrin, L., Vanegas, J. A., Ovalle, J. E. A., Beltrán, V., Escalante, H. J., Montes-y-Gómez, M., and González, F. A. (2015). INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation. In CLEF (Working Notes).
- Pellegrin, L., Vanegas, J. A., Arevalo, J., Beltrán, V., Escalante, H. J., Montes-y-Gómez, M., and González, F. A. (2016, September). A two-step retrieval method for Image Captioning. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 150-161). Springer, Cham.
- Beltrán V., Journet, N., Coustaty, M., Doucet, A. (2019). Semantic Text Recognition via Visual Question Answering, ICDAR WML 2019 (In progress)
- Beltrán V., Caicedo, J., Coustaty, M., Journet, N., Doucet, A. (2019). Cross-modal document retrieval via multimodal deep learning (Submitted)

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

[3] H. Xiao, “bert-as-service.” https://github.com/hanxiao/bert-as-service, 2018.

[4] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[5] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, “What is wrong with scene text recognition model comparisons? dataset and model analysis,” *arXiv preprint arXiv:1904.01906*, 2019.

[6] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, “Towards vqa models that can read,” *arXiv preprint arXiv:1904.08920*, 2019.

[7] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiñol, M. Mathew, C. Jawahar, E. Valveny, and D. Karatzas, “Icdar 2019 competition on scene text visual question answering,” *arXiv preprint arXiv:1907.00490*, 2019.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Semantic Understanding of Floor Plan Images through Machine Learning Techniques

Student’s Name: Shreya Goyal
Supervisors of the thesis: Dr. Chiranjoy Chattopadhyay & Dr. Gaurav Bhatnagar
University: Indian Institute of Technology Jodhpur, India
Starting date of PhD: 22 July 2016
Expected finalization date of PhD: 31 July 2020
email: goyal.3@iitj.ac.in

***Abstract*—Due to the rapid urbanization, there is a requirement of new designs and plans of buildings which are inspired by existing models and suit user needs. Also, there is a massive demand in understanding heritage buildings, as well as making them accessible for the community using new technologies. There is a requirement to build tools which minimize the gap between architects and user to understand their conditions and fit best in architectural rules as well. In this digital era, a user may want to understand the technicalities of floor plan design using some multimedia description or a self rendered floor plan to experience architecture. Hence, the motivation of this work is to generate a natural language description given the floor plan image and generate floor plans from the story or indoor scene. The synthesized story may also be used for navigation in buildings by visually impaired, tourists, or robots.**

***Keywords*-Floor plans, Textual description, Segmentation, Symbol spotting**

I. Introduction

In the field of architecture and building engineering, the floor plan is a drawing of the house, apartment, or any other building. These are the graphical documents which aid architects to show the interior of a building along with components. It shows the rooms, doors, windows, furniture, and the properties. Floor plan image analysis involves detection of the rooms, walls, doors, other entities, and identifying a connection between them. Tasks such as annotation detection, wall characterization, doors and window detection, room, and sub-room detection are various aspects in floor plan image analysis. This low-level information is used for further processing (high-level analysis).

In the literature, researchers have worked on various aspects of floor plan analysis mentioned in the previous paragraph. However, in this thesis, we have introduced some novel approaches which are important in terms of research as well as real-world applications. Figure 1 depicts the summary of the steps taken so far and the future plan. The direction of arrow on the axis shows the input and output into consideration. In the following sections, a brief description of the works carried out during the Ph.D. are presented.

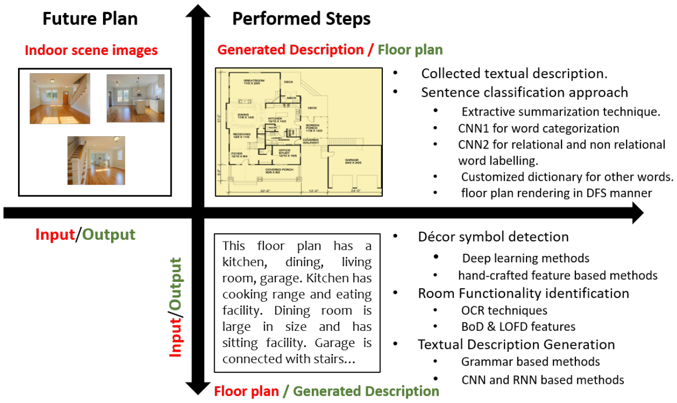

Future Plan
Performed Steps
Indoor scene images
Generated Description / Floor plan
• Collected textual description.
• Sentence classification approach
• Extractive summarization technique.
• CNN1 for word categorization
• CNN2 for relational and non relational word labelling.
• Customized dictionary for other words.
• floor plan rendering in DFS manner
Input/Output
This floor plan has a kitchen, dining, living room, garage. Kitchen has cooking range and eating facility. Dining room is large in size and has sitting facility. Garage is connected with stairs...
Input/Output
• Décor symbol detection
• Deep learning methods
• hand-crafted feature based methods
• Room Functionality identification
• OCR techniques
• BoD & LOFD features
• Textual Description Generation
• Grammar based methods
• CNN and RNN based methods
Floor plan / Generated Description

Figure 1. Steps taken so far and the future plan

II. BRIDGE-dataset

With the advent of deep learning models, there is a requirement of large scale data in order to train these models efficiently. However, there was a lack of large scale data in the context of floor plans and related annotations for the purpose of symbol detection, classification and understanding the functionality of the rooms and plan. The BRIDGE-dataset (Building plan Repository for Image Description Generation, and Evaluation) was constructed, which contains $\sim$ 13000 floor plan images, textual descriptions, region wise captions, dcor symbol annotations [1]. The floor plan images and descriptions contained by BRIDGE were collected from two websites and annotations for region wise captions and objects were done by volunteers. The available description can be used for generating textual description from images and evaluation purposes.

III. Decor symbol detection and classification

For the purpose of understanding floor plans from the symbols they contain, we detected and classified the decor symbols using blob detection technique after pre-processing the images using morphological techniques [2] [3]. Our decor characterization method calculates the signature of a decor item in a given room and compares it with the signatures present in decor template library, and the label of the closest signature is assigned to it. In another work [3], we also classified wall symbols representing different material by segmentation and LBP (local binary pattern) feature matching. Currently deep learning solutions for object detection are achieving benchmark accuracy and hence we experimented with state of the art techniques of object detection in natural images for example YOLO, Fast RCNN, Faster-RCNN for decor symbol characterization using BRIDGE dataset [1]. It was observed that deep learning techniques are much more efficient then traditional methods using hand crafted features. In future we plan to include more decor symbol for detection and classification.

IV. Room characterization

In order to extract information from floor plans, two novel features were proposed, BoD (Bag of Decor) and LOFD (Local Orientation and Frequency descriptor) for extracting room functionality. BoD features is a sparse histogram of frequency count of the decor items present where each cell of the vector represents one decor item taken from standard dataset [4]. LOFD feature also contains spatial information along with the frequency count of the decor items. A machine learning model was trained using these features and rooms were classified in 5 classes, bedroom, bathroom, kitchen, living room, entry [2][5]. In the previous work rooms were identified using OCR techniques [3].

V. Narration generation from floor plans

Low-level information, such as furniture type, location, room type, etc. extracted from a given floor plan images are utilized for textual description synthesis. In [3], ego-centric description was generated from floor plans by parsing a region adjacency graph in DFS manner. Room information was obtained using OCR techniques and decor information was generated by signature matching method. In the later stage textual description was generated using grammar based methods by extracting information from floor plans using machine learning methods (BoD and LOFD) [2], [5]. Information regarding door to door navigation is also included in the [5] by developing algorithm for door to door navigation by obstacle avoidance.The proposed work can be useful for visually impaired for navigating within a building using a text reader software. In future we plan to generate textual description using sophisticated CNN and RNN models to have a more human like descriptions.

VI. Floor plan synthesis from text and images

In [6], we focus on building an artificially intelligent framework for providing support to the architects or designers in the early-stages of the architectural design. To generate floor plans from textual description, natural language processing with deep learning techniques were employed. The textual data collected from volunteers was summarized using extractive summarization technique after a pre-processing. The words in the sentences were classified using two CNNs. First CNN was used to classify words depicting a room on the basis of its functionality, for example, kitchen, bedroom etc. The other CNN was used for labeling relational sentences. From sentences labeled with room tag, we extract information like shape, dimensions, architectural objects inside room, door connectivity and wall sharing for each tag using techniques like tokenization, regular expression matching. For extracting rooms and architectural objects, we have built our own custom dictionaries. Floor plan was rendered using the information extracted in previous steps. In the further steps we plan to generate floor plan images taking indoor images into consideration, since textual description is very user dependent and may not contain all the information.

VII. Applications

The generated textual description from the floor plan images has applications in culture and heritage. They can be used by a tourist to navigate inside a historical monument. It also can be used in robotics for a robot to navigate. Also it has applications for rental websites to describe their posted ads in text automatically. The automatic floor plan generation can be used for real estate renting and selling. If the generation is taken to the 3D level it has application in virtual/augmented reality by giving user a virtual experience of his/her desired home. As the architectural plans/designs are rapidly accumulating in the form of big data, professional architects can use these tool for retrieving existing plans which they might want to refer for new designs. In future we plan to generalize the developed tools for more general floor plan images which could be used in the form of a mobile app by a common user.

References

[1] **Shreya Goyal**, Vishesh Mistry, Chiranjoy Chattopadhyay and Gaurav Bhatnagar, "Bridge: Building plan repository for image description generation, and evaluation," in *ICDAR*. IEEE, 2019.

[2] **Shreya Goyal**, Chiranjoy Chattopadhyay, and Gaurav Bhatnagar, "Asysst: A framework for synopsis synthesis empowering visually impaired," in *MAHCI-ACM MM*. ACM, 2018, pp. 17–24.

[3] **Shreya Goyal** and Chiranjoy Chattopadhyay and Gaurav Bhatnagar, "Plan2text: A framework for describing building floor plan images from first person perspective," in *CSPA*. IEEE, 2018, pp. 35–40.

[4] M. Delalandre, E. Valveny, T. Pridmore, and D. Karatzas, "Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems," *IJDAR*, vol. 13, no. 3, pp. 187–207, 2010.

[5] **Shreya Goyal**, Satya Bhavsar, Shreya Patel, Chiranjoy Chattopadhyay and Gaurav Bhatnagar, “Sugaman: Describing floor plans for visually impaired by annotation learning and proximity based grammar,” *arXiv preprint arXiv:1812.00874*, 2018.

[6] Mahak Jain, Anurag Sanyal, **Shreya Goyal**, Chiranjoy Chattopadhyay and Gaurav Bhatnagar, “Automatic rendering of building floor plan images from textual descriptions in english,” *arXiv preprint arXiv:1811.11938*, 2018.

Layout Analysis in historical documents using machine learning methods - The Baseline Extraction case

Xenofon Karagiannis, PhD Candidate
Visual Computing Group, Department of Electrical and Computer Engineering
Democritus Univesity of Thrace
Xanthi, Greece
Email: xkaragia@ee.duth.gr
Thesis Supervisor: Ioannis Pratikakis
Starting/Expected finalization date of the PhD: Dec 2018 / Dec 2022

***Abstract*—The goal of this research topic is to develop machine learning methods that support individual tasks in a layout analysis framework with an initial focus on text line extraction. These tasks are necessary for the success of text recognition. Our research will focus on manuscripts with minuscule writing dated from 9th to 17th century.**

I. Introduction

Text line detection and segmentation is an essential preprocessing step for handwritten text recognition (HTR) [1] [2] and still presents challenges especially for handwritten historical documents. Some of theses challenges are induced by the heterogeneous nature and the physical degradations (such as bleed-through, stains, faded away characters) that historical documents exhibit. Recent competitions [3] cover many of these challenges by evaluating the performance of state-of-the-art methods for detecting baselines in archival document images.

Current methods that achieve state-of-the-art results on text line segmentation tasks are based on deep learning and they make use of deep learning models either in an end-to-end manner, or as part of a multi-stage approach. Inspired by the work of Tobias Gruning et al. [4], we proceed on a two-stage approach that includes a variant of U-Net which performs pixel-level classification and a second stage that estimates superpixels and outputs the final baselines. Our contribution to their work is the development of an alternative second stage that results in the final baseline extraction.

Our research will focus on manuscripts with minuscule writing dated from 9th to 17th century. The final goal of our research team is the transcription of these manuscripts and as part of my PhD, my efforts will focus towards developing methods for layout analysis.

II. METHODOLOGY - CURRENT WORK - PRELIMINARY RESULTS

As a first step, we have studied previous work on the task of baseline extraction. We focused on the work of Gruning et al. [4] as their approach achieves state-of-the-art results on a challenging dataset [5]. Specifically, we have used the first stage of the two-stage method they proposed that includes the deep learning model ARU-Net - an extension of the popular U-Net. Based on that, we developed our own second stage. After extracting the superpixels (SPs) of the first stage's output and applying the Delaunay triangulation, we follow an alternative approach to build the final baselines. This relies upon grey level statistics and geometric constaints. This way, we end up with coarse baselines. Based upon the assumption that a baseline can be approximated by a second degree polynomial and we apply polynomial regression to each coarse baseline (see Fig. 1).

A. Datasets

So far, we have used two datasets. The first dataset is the Track A of the ICDAR2017 Competition on Baseline Detection (cBAD) dataset [5] and the second is our own proprietary dataset, called Eparchos. Track A subset consists of 216 training images and 539 test images and includes only pages with simple page layouts.

Eparchos dataset is a document collection that originates from the handwritten code British Museum Additional No. 6791, circa 1530, and includes texts by Michael Cello (De Omnifaria Doctrina) and Matthew Blastari (Collectio Alphabetica) handwritten by Antonios Eparchos and Camilo Giannetto. In Fig. 2 we present two sample images from Eparchos dataset. Two of the main challenges are the segmentation of the drop caps and the initial letters, as well as the correct baseline extraction of the paragraph subtitles (red letters). According to the baseline Definition in [5], we have produced ground truth for 50 images in our collection, using the Transkribus platform [6].

B. Preliminary experimental results

For comparative purposes we use the evaluation scheme introduced in [5] on the ICDAR 2017 and ICDAR 2019 Competitions on Baseline Detection (cBAD). For both

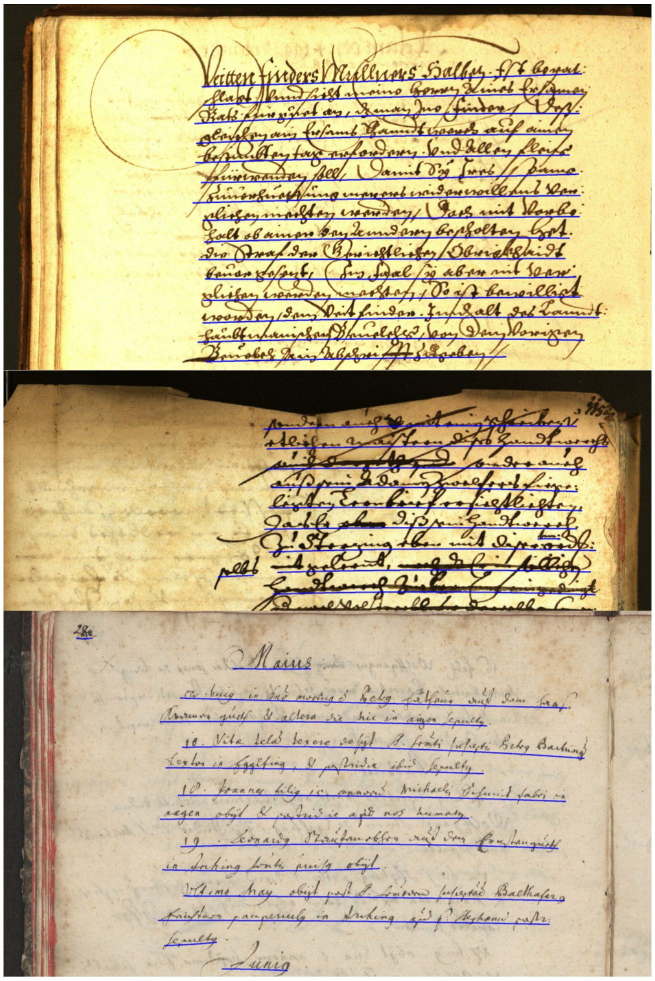


Figure 1: Sample results on cBAD - Track A

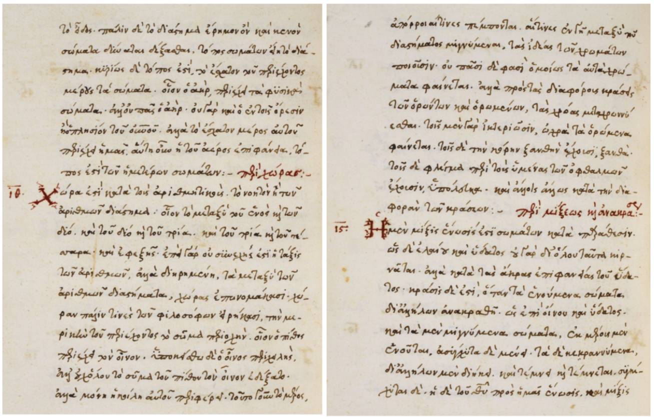


Figure 2: Document images - Eparchos dataset

datasets, we used the pretrained ARU-Net without any fine-tuning. Even though our document images have special characteristics that only a handful documents in the cBAD dataset have, we applied the pretrained first stage along with our second stage on our dataset as well. The preliminary results on both datasets are presented in Table I.

Table I: Results for cBAD-Track A and Eparchos datasets

Dataset	Precision	Recall	F value
cBAD - Track A (Gruning et. al [4])	0.977	0.98	0.978
cBAD - Track A (proposed second stage)	0.942	0.967	0.955
Eparchos (Transkribus platform)	0.978	0.947	0.963
Eparchos (proposed second stage)	0.964	0.928	0.946

Input scale	Precision	Recall	F value
(a)	0.945	0.924	0.934
(b)	0.933	0.91	0.921

III. Challenges and Future Work

The current performance of our method, although being satisfactory, is still behind the state of the art. Thus, further work is needed to improve performance .

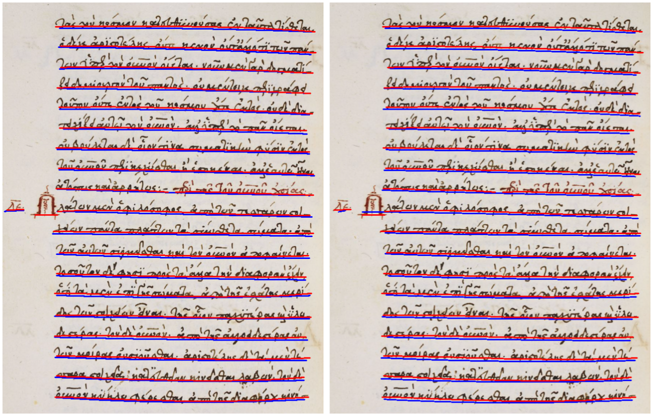
One of our goals is to examine the method's properties of equivariance over scale. Even though the attention mechanism is developed to allow the ARU-Net to focus on content at different scales, we have observed irregularities when we extract baselines of a document image for which the input is considered at different scales. An example of this behavior is presented in Fig. 3. The image at Fig. 3(a) has greater scale than the image at Fig. 3(b). The evaluation for each input scale is shown in Table **??**. Motivated by this observation, we also plan to study the effects of different attention mechanisms.

Furthermore, bleed-through text affects our method's results. We believe that the main part of this problem is that the pretrained ARU-Net cannot discriminate between regular and bleed-through text, producing incorrect probability maps and propagating this error through the second stage. An explanation for this behavior could be the limited number of documents that exhibit bleed-through text. As we plan to retrain ARU-Net from scratch, we will take this into consideration, and try to augment the training data either by using transformations on available documents or by creating artificial document images with bleed-through text.

Finally, both recent experiments and similar work [7] imply that it would be helpful to incorporate a document layout analysis task when extracting baselines. We will work towards identifying a meaningful way to incorporate the document's layout information either as an addition to the model used in the first stage of the method, or as a separate model.

Acknowledgement

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-01939)



(a)

(b)

Figure 3: Baseline extraction results on a document image for which the input is considered at different scales. The input at (a) is twice as larger than the input at (b).

References

[1] Sanchez, J.A., Romero, V., Toselli, A.H. and Vidal, E. ICFHR2016 competition on handwritten text recognition on the READ dataset (2016). In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 630-635). IEEE.

[2] Romero, V., Sanchez, J.A., Bosch, V., Depuydt, K. and de Does, J. Influence of text line segmentation in handwritten text recognition (2015). In 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 536-540). IEEE.

[3] Diem, M., Kleber, F., Fiel, S., Gruning, T. and Gatos, B. cbad: Icdar2017 competition on baseline detection (2017). In 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1355-1360). IEEE.

[4] Tobias Grning, Gundram Leifert, Tobias Strau, Roger Labahn, A Two-Stage Method for Text Line Detection in Historical Documents (2018), Computing Research Repository (CoRR).

[5] Tobias Grning, Roger Labahn, Markus Diem, Florian Kleber, Stefan Fiel, READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents, May 2017, arXiv:1705.03311. [Online]. Available: https://arxiv.org/abs/1705.03311

[6] Kahle, P., Colutto, S., Hackl, G. and Mhlberger, G. Transkribus-a service platform for transcription, recognition and retrieval of historical documents (2017). In 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 4, pp. 19-24). IEEE.

[7] Lorenzo Quiros, Multi-Task Handwritten Document Layout Analysis, Dec. 2018, arXiv:1806.08852. [Online]. Available: https://arxiv.org/pdf/1806.08852

Historical big-data: modelization of strategies to analyze collections of documents

Students name: Camille Guerry

Supervisors of the thesis: B. Couäsnon, A. Lemaitre, S. Adam
University: INSA Rennes
Starting date of the Ph.D.: October 2018
Expected finalization date of the Ph.D.: October 2021
Email: camille.guerry@irisa.fr

Abstract. To recognize collections of document images, method of state of the art often process each image of the collection independently. Consequently, logical links that could exist between data of different pages are under-exploited. The goal of this Ph.D. is to propose a recognition system able to use the context given by the collection. To validate our method, we will test our system on financial documents in the late 19th and early 20th centuries. However, our system will be generic enough to be adapt on other kinds of collections.

Keywords: document images analyzes · historical big data · structure recognition · collection · iterative strategy

1 Short research plan

1.1 Overview of the Ph.D. topic

This Ph.D. thesis is part of the ANR HBDEX project. The ANR HBDEX project is carried out in collaboration with the DFIH team of the Paris School of Economics and the LITIS Laboratory of Rouen. It aims to make a major advance in understanding the long term-dynamics of financial markets. In this context, we are developing a system for automatic recognition of Stock Exchange daily lists collections from different financial markets in the late 19th and early 20th centuries.

Stock Exchange daily lists (see Fig. 1) are tabular structured and contain all securities traded the day before, the exchange prices and various information concerning each security. OCR and recognition tools such as can be found in commerce, do not allow a recognition that is reliable enough. Therefore, the challenge is to propose a reliable recognition system that minimizes the interventions of human users. For this, we explore different strategies that exploit the context brought by the collection. We first focused on documents from La Coulisse (Paris unofficial market). We will then adapt the strategy we develop to other document collections.

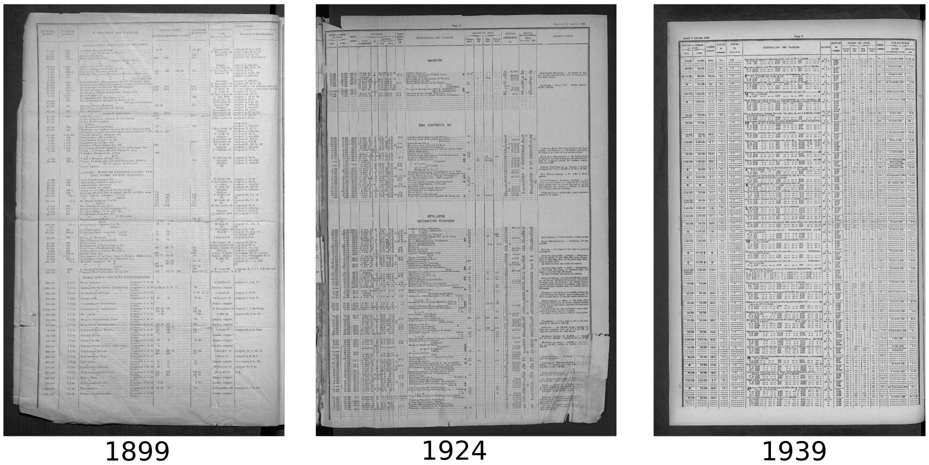


Fig. 1. Exemple of Stock Exchange daily lists

1.2 Work done so far

The work done so far has consisted of designing a strategy for global recognition (see Fig. 2)of the collection and developing a first system that recognizes the tabular structure of daily list documents. Our first system recognizes the structural organization of daily lists documents thanks to a grammatical description written in the EPF language of the generic DMOS-PI method [2]. The terminals of our grammar are the text lines (obtained through deep learning [3]), vertical and horizontal segments. We will then integrate this system into an iterative global recognition process. Each iteration will validate a type of data in the hierarchical order given by the document structure: columns, sections, securities, other fields. Each iteration will consist of two steps. During the first step, the structural recognition system extracts elements in the image. The second step is a transversal validation phase of the extracted information.

For the validation phase, we model the context of the collection. To this end, we use the specifications of the documents produced by our economist partners. We propose a way to exploit these different rules for: provide context to the OCR used and make the extracted information more reliable. To put in place our strategy, we notably rely on the thesis of Chazalon [1].

1.3 Future Work

The next step of this Ph.D. is the implementation of the validation phase. For this phase we explore work on times series. We will then measure the improvement that we can obtain with our system compared to an individual analyze of each image. We will also consider the possibility to extract information from the Financial Industry Business Ontology (FIBO). The aim of this is to integrate

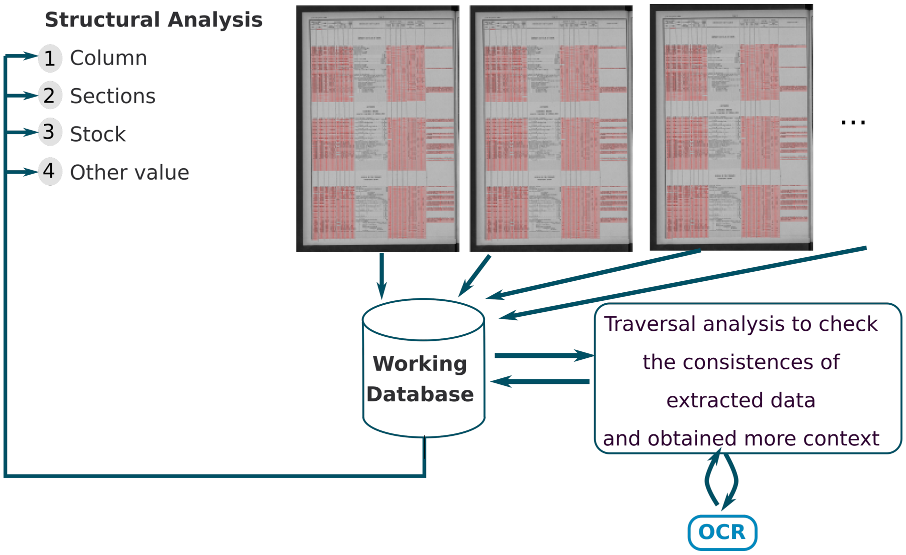

Structural Analysis
1 Column
2 Sections
3 Stock
4 Other value
...
Working Database
Traversal analysis to check the consistences of extracted data and obtained more context
OCR

Fig. 2. Overview of the iterative process.

other financial rules in our system for the validation of extracted data. Finally, an important step of the Ph.D. will be the generalization of our method on other documents collections such as yearbooks of companies or stock exchange lists from other financial markets.

References

1. Couasnon B. Lemaitre A. Chazalon, J. Iterative analysis of pages in document collections for efficient user interaction. International Conference on Document Analysis and Recognition, 2011.
2. B. Couäsnon. Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way. International Journal of Document Analysis and Recognition (IJDAR), 2006.
3. Seguin B. Kaplan F. Oliveira, S. A. dhsegment: A generic deep-learning approach for document segmentation. 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018.

Automated Lecture Video Summarization via Extraction and Feature Representation of Text Content

Student: Bhargava Urala Kota
Advisor: Venu Govindaraju
Start Date: August 2015
Expected End Date: May 2020
University at Buffalo, New York, USA

***Abstract*—Lecture videos are a useful resource for students and educators across the world. Our goal is to summarize lecture videos by extracting regions of text content from every frame, extracting features from text regions to represent content in each frame. Finally, summaries of videos can be produced by comparing text features across frames locally (within a temporal window) and globally (across the entire video) to obtain a smaller subset of frames (called *keyframes*) which contain all of the text in the video. Extracted frames and text features facilitate content based lecture video search.**

I. Research Problem

The ubiquity of cameras and the internet has resulted in the availability of large amounts of lecture videos. While current search engines primarily support meta-data based search and retrieval of lecture videos, effective *video summarization* techniques are needed to extract key content and condense this data into an easily searchable form, to facilitate content based search.

Lecture content is often loosely structured and exhibits large variances in semantic grouping. Examples include sentences, multi-line phrases, sketches, plots and mathematical expressions. Further, background noise, illumination changes and occlusions are also present. Lectures could have handwritten content - either on white/blackboards or digitally rendered, which adds to the challenge for extraction and feature representation of the content.

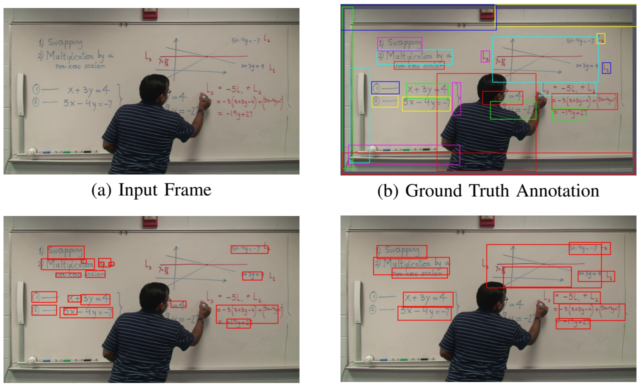
Text content in lecture videos needs to robustly represented in order to distinguish between instances of different unique text regions. Techniques to learn vector embeddings for words when transcripts are available have been shown [1]. However, for lectures this needs to be extended for structured text such as equations and matrices, needing extensive transcription annotations. Forms of lecture video summaries other than keyframes in the literature include composite frame images and transcription-based summaries.

II. Work so far

Our preliminary work has mostly been concentrated on AccessMath, a publicly available annotated dataset of whiteboard lecture videos [2]. Evaluation metrics used are the number of keyframes produced as well as the recall, precision and f-measure of keyframe content (at the level of binarized connected components) with respect to all unique text content in the lecture video.

In prior art, handwritten content (HC) was extracted by assuming the video was shot with a still camera. Background was estimated for the entire video and removed from each frame followed by specialized binarization technique to extract text content [2]. In our work [3], we used a neural network based handwritten content detector (HCD) to extract text regions from lecture video frames with no assumptions about still camera at this stage.

Out of the box, the detector based on TextBoxes [4] trained on detecting English words in natural scenes [5], did not perform well on lecture video data. Particularly, we observed that the model failed to detect the variety of handwritten content and layouts which are expected in a lecture video. Thus a specialized HC detector was needed.



(a) Input Frame

(b) Ground Truth Annotation

(c) Detection with Raw Model

(d) Detection after Fine-tuning

Figure 1. Out-of-the-box and fine-tuned handwritten content detection results for input frame (a) with ground truth (b) are shown in (c) and (d).

We decided to 'fine-tune' TextBoxes from scene text parameter initializations to a create a dedicated model for detecting HC. It should be noted that AccessMath does not contain annotations of the content in terms of bounding boxes nor the timestamps when it was written/erased, since it was intended to be evaluated at the binary CC level. Therefore, we annotated the frames with bounding boxes around handwritten content regions. Sample frame and its annotations are shown in Figure 1(a) and (b) respectively.

After handwritten content was extracted from all frames of videos, detected text regions were compared across frames at both bounding box and binary connected component (CC) level to find similar content in frames over time. Binarization was achieved by background estimation (using a median filter) and subtraction followed by Otsu's algorithm.

In the next step, frames that contain CCs conflicting in spatial location during non-overlapping time intervals in the entire lecture video are identified. Finally, we segment the videos by greedily selecting frames that resolve the maximum number of spatio-temporal conflicts in content associations across frames [2], thus obtaining the *keyframes* of the lecture video. We found that the performance was comparable to state-of-the-art despite using a relatively simple binarization method (see Method 1 in Table I).

In a more detailed study [6], we proposed a generalized pipeline without assumptions of fixed camera to summarize lecture videos. We use a neural network HCD, as before, and a full frame binarizer to represent text content. We trained a different HCD based on the EAST scene text detection model [7]. We compared the performance of training from just lecture videos with our previous work of finetuning from scene text data and found similar performance.

In this work, we also developed a full frame binarizer using a convolutional-deconvolutional neural network and compared it with Random Forest (RF) binarizer used in the baseline method [2]. We found that the RF binarizer worked best in terms of the AccessMath evaluation metrics, which we found to be very sensitive to the thickness of binary CCs in the summary. When compared to the ground truth keyframes using the metrics of pseudo-recall and pseudo-precision, which are designed to emphasize shape over thickness, we found that our full frame binarizer had better performance. The final performance metrics are reported in Table I (Method 2). The higher recall and lower precision can be attributed to new detector model which was anchor-free and thus more flexible but trained on limited data. In terms of f-measure, we get similar performance with a generalized pipeline when compared to the baseline with heuristic components tailor-made for AccessMath.

Table I
COMPARISON OF DIFFERENT METHODS OF LECTURE VIDEO SUMMARIZATION BY MEASURING RECALL (R), PRECISION (P), F-SCORE (F) AND NUMBER OF FRAMES (N_f).

	AVG	AVG GLOBAL			AVG PER FRAME		
METHOD	N_f	R	P	F	R	P	F
Baseline 1 [2]	18	96.28	93.56	94.90	95.73	92.21	93.93
Baseline 2 [8]	13	95.89	86.28	90.83	94.18	85.15	89.44
Method 1 [3]	20	92.33	94.16	93.23	91.69	93.45	92.56
Method 2 [6]	21	95.80	92.88	94.32	95.40	92.44	93.90

In our most recent work [9], we experimented with feature vector based text representation as opposed to binary CCs. We found that standard visual descriptors like SIFT, SURF etc. were not suitable for text and because we did not have transcription annotations we could use methods like PHOCNet [1] for embedding. Thus, we adapted a triplet-loss based embedding scheme for representing text regions. This method showed some promise in being able to group together similar content and recall unique text content across the video, however when used to generate *keyframes*, we obtained many redundant frames.

List of Publications: [3, 6, 9]

III. NEXT STEPS

Our immediate next step is to develop methods to robustly represent regions of text as feature vectors. With these features, we would be able track changes across the video and accurately segment the video into semantic sections based on content changes while handling camera motion and occlusions. This representation must also be sensitive to structured text such as equations and matrices in order to accurately capture all the variety in text content and provide search and retrieval capabilities. Graph embedding methodologies [10] show some promise in this regard. However adapting these techniques to text and math expressions will be a novel and interesting research challenge.

To test retrieval performance, user studies based on queries constructed from text content within the lecture are planned. Applying our framework to lectures and presentations based on slide decks are also planned to establish generalizability. Exploration of different kind of summaries based on work in general and lectures videos such as skims, text-based summaries is also planned.

REFERENCES

[1] S. Sudholt and G. A. Fink, "Phocnet: A deep convolutional neural network for word spotting in handwritten documents," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2016, pp. 277–282.

[2] K. Davila and R. Zanibbi, "Whiteboard video summarization via spatio-temporal conflict minimization," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[3] B. Urala Kota, K. Davila, A. Stone, S. Setlur, and V. Govindaraju, "Automated detection of handwritten whiteboard content in lecture videos for summarization," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2018, pp. 19–24.

[4] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network." in *AAAI*, 2017, pp. 4161–4167.

[5] 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 *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.* IEEE, 2015, pp. 1156–1160.

[6] B. Urala Kota, K. Davila, A. Stone, S. Setlur, and V. Govindaraju, "Generalized framework for summarization of fixed-camera lecture videos by detecting and binarizing handwritten content," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–13, 2019.

[7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proc. CVPR*, 2017, pp. 2642–2651.

[8] F. Xu, K. Davila, S. Setlur, and V. Govindaraju, "Content extraction from lecture video via speaker action classification based on pose information," in *2019 15th International Conference on Document Analysis and Recognition (ICDAR)*, vol. accepted, 2019.

[9] B. Urala Kota, S. Ahmed, A. Stone, K. Davila, S. Setlur, and V. Govindaraju, "Summarizing lecture videos by key handwritten content regions," in *2019 8th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)*, vol. accepted, 2019.

[10] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.

Multilingual OCR correction for ancient books: Looking at multiple documents to fix multiple words

Student's name: Nguyen Thi-Tuyet-Hai
Supervisors of the thesis: Prof. Antoine Doucet, Prof. Mickael Coustaty
University: University of La Rochelle
Starting date of the PhD: 01/2017
Expected finalization date of the PhD: 01/2020
Email: hai.nguyen@univ-lr.fr

***Abstract*—Substantial efforts have been devoted to optical character recognition (OCR) that translates printed documents to digital ones in order to preserve and make them fully accessible, searchable in digital form. However, the poor quality of the paper input is the main reason for producing OCR errors and decreasing the performance of OCR systems. Various post-processing approaches have been proposed to detect and correct OCR errors. My core research, therefore, focuses on building a multilingual post-OCR processing tool.**

I. Short research plan

In this section, I would like to give an overview of my PhD topic, describe the general approach which I am trying to implement as well as my plan for the next steps.

A. Introduction

Paper-based documents contain valuable knowledge that attracts a lot of attention from researchers and libraries around the world. In order to preserve and make these documents easily accessible, much effort has been dedicated for optical character recognition (OCR) to digitize such documents.

A typical OCR system consists of many components, as illustrated in Fig.1. Firstly, a paper-based document is transformed into an image-based document by an optical scanner. Through preprocessing, the OCR system locates data regions and segments words into isolated characters. Feature extraction is one of the most important steps which extracts various types of features to recognize characters. Classification algorithms are based on similarity measures between the extracted features and the existing ones. Finally, linguistic and/or contextual information can be used to identify ambiguous characters or correct words. Following this procedure, the digital document will be stored in the database and be ready to be exploited digitally [1].

Quality of OCRed text depends on not only on physical quality of documents but also performance of OCR techniques. Therefore, various post-processing approaches have been proposed to detect and correct OCR errors. They can be divided into three categories: dictionary-based error correction and context-based error correction [2].

A typical post-processing approach consists of two steps, detecting and correcting errors. In terms of the detection task, dictionary and character n-gram models are often used to detect *non-word* errors. In terms of the correction task, for each OCR error, the list of candidates are generated based on different sources at character level, word level. The best candidate is the correction in an automatic mode, or the top *n* candidates are suggested to correct the error in a semi-automatic mode.

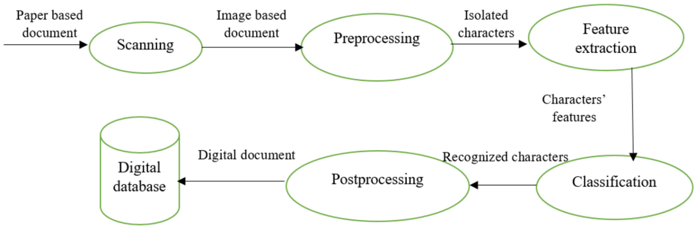


Fig. 1. A typical OCR process.

A wide range of approaches was devoted to OCR post-processing, which can be classified into two main types: dictionary-based and context-based types. The *dictionary-based type* aims to correct isolated-word errors and does not take nearby context into consideration [3], [2], [4], hence this type cannot deal with *real-word* errors.

The *context-based type*, which considers grammatical and semantic contexts of errors, promises to overcome the issues of the first type. Most of the techniques of this type rely on noisy channel and language model [5], [6], [7]. The others explore several machine learning techniques to suggest correct candidates [8], [9], [10].

Jones *et al.* [5] and Tong *et al.* [6] explored several features, including character n-grams, character confusion (or device mapping statistics), and word bi-gram in different ways to detect and correct erroneous OCR tokens. Using similar features, Llobet *et al.* [7] built an error model and a language model, then added one more model built from character recognition confidences, called hypothesis model. Three models were compiled separately into Weighted Finite-State Transducers (WFSTs), then were composed into the final transducer. The best token was the lowest cost path of this final transducer. However, character recognition confidence is often missing at least with the whole competition dataset [11] and Overproof evaluation datasets [12].

Along with the development of machine translation techniques, some approaches considered OCR post-processing as machine translation (MT), which translates OCRed text into the correct one in the same language. Afli *et al.* [8] and

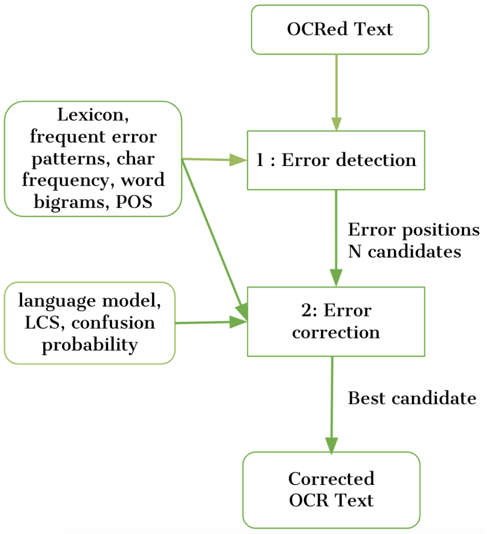

OCRed Text
Lexicon, frequent error patterns, char frequency, word bigrams, POS
1 : Error detection
Error positions
N candidates
language model, LCS, confusion probability
2: Error correction
Best candidate
Corrected OCR Text

Fig. 2. The Post-OCR correction system architect.

some competition approaches of the competition [11] applied machine translation techniques (from statistical MT, neural MT to hybrid MT at word and/or character level) to deal with detecting and correcting OCR errors.

Other approaches [9], [10] explored different sources to generate candidates and then ranked them using a regression model. Several features were extracted such as confusion probability, uni-gram frequency, context feature, term frequency in the OCRed text, word confidence, and string similarity. Then, a regression model was used to predict the best candidate for erroneous OCR token.

In general, in order to deal with OCR errors from a multilingual document, n-gram model is one of the effective techniques. In fact, at the character level, n-gram is useful for detecting errors or learning character confusions. Additionally, n-gram at the word level is a good choice for correcting real-word error because it takes the context into consideration. However, the data sparsity of the word n-gram model is still a big challenge. The other problem is that the n-gram model is not powerful enough to capture a long context [13]. Therefore, some other language models should be considered.

B. Methodologies

In this section, I present two general approach to deal with OCR error detection and correction. The first method utilizes character error model and language model while the second one applies machine translation techniques to detect and correct OCRed errors.

1) Adaptive character error model and language model: The approach consists of three main steps which are detailed in Fig.2. Firstly, the errors are detected by using different methods, for instance, character n-gram models, lexicon techniques, binary classifiers. Secondly, frequent error patterns learned from the ground truth are combined with lexicon, word ngrams for generating correction candidates. In this step, the best N candidates are chosen based on adaptive confusion probability and be used as the input for the next step [12]. Finally, all features from character to word level are utilized to correct errors using ranking mechanisms such as regression models.

2) Machine translation (MT) as post-OCR processing: MT techniques translate OCRed text into GT text in the same language. If all OCRed text is used as input, MT simply copies input to output. Therefore, only erroneous OCRed text and some nearby tokens are used as input and corresponding GT text are used as output. For example, there is an error 'couise' with one previous context token, three following tokens and their corresponding GT text.

Input	OCRed text	This couise was agreed to
Output	GT text	This course was agreed to

Afli *et al.* [8] reported that models at word level outperform those at character level. Otherwise, Amrhein *et al.* [14] suggested to use MT at character level, and they presented that at statistical MT is good for error correction and neural MT work well for error detection. It is the fact that neural MT obtains higher performance than statistical MT. Therefore our approach applies neural MT at character level.

C. Experimental results

1) Dataset: English monograph dataset of the ICDAR 2017 Competition on Post-OCR text correction[1].

2) Evaluation metrics: Our results are evaluated by the same metrics (Precision, Recall, F-measure) and the same evaluation tool as the ones used in the competition.

3) Results: The initial results showed that our methods are competitive with other methods of the competition. Our error detector which applied binary classifier with feature values computed from candidates of each OCRed error positions outperform that others. Performances of our corrections are higher than the single model of neural MT (CLAM in Table. I), but they are still far from the combined model between statistical MT and neural MT (Char-SMT/NMT in Table. I).

TABLE I. INITIAL RESULTS OF OUR DETECTION/CORRECTION APPROACHES, CORRECTION 1 : CHARACTER ERROR MODEL, LANGUAGE MODEL; CORRECTION 2: NEURAL MACHINE TRANSLATION

Approach		Detection			Correction
		P	R	F	(% improvement)
Baseline	Char-SMT/NMT	98	51	67	**43**
	CLAM	94	52	67	29
	EFP	63	77	69	13
	WFST-PostOCR	67	82	73	28
Our approaches	Detection	82	76	**79**	
	Correction 1				30
	Correction 2				33

D. Future Work

I apply neural machine translation to detect and correct OCRed errors. Currently, the performance of our neural MT method is still lower than the combined one. I try to find some solutions to help models pay more attention on erroneous characters as well as nearby context, then suggest more accurate candidates.

Moreover, there are several OCR datasets with poor quality of ground truth, therefore it is worth taking unsupervised approaches into consideration. I keep searching other approaches to deal with post-OCR problem. [15]

[1] https://sites.google.com/view/icdar2017-postcorrectionocr/

II. Curriculum Vitae

A. Education

2017 - present: PhD student, University of La Rochelle

2012 - 2014: Posts and Telecommunications Institute of Technology

- Master of Information Systems
- GPA: 8.56/10
- Title of thesis: The software of identifying the semantic relations between educational legislative documents.

2006 - 2011: Bachelor student, Posts and Telecommunications Institute of Technology

- Bachelor of Information Technology
- GPA: 8.5/10
- Title of thesis: The software of giving and receiving thesis work-flow in IT department based on BPEL (Business Process Execution Language).

B. Experience

2014 - 2016: Posts and Telecommunications Institute of Technology, Information Technology Faculty.

- Lecturer
- Courses: Introduction to Informatics, C++ Programming Language, Service-Oriented Software Development, Object Oriented Programming, and Software Engineering.

2011 - 2014: Posts and Telecommunications Institute of Technology, Information Technology Faculty.

- Teaching assistant
- Courses: Introduction to Informatics, C++ Programming Language, Service-Oriented Software Development, Object Oriented Programming, and Software Engineering.

C. Publications

[C1] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickal Coustaty, Nhu Van Nguyen and Antoine Doucet: Post-OCR Error Detection by Generating Plausible Candidates. ICDAR2019. Accepted.

[C2] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickal Coustaty, Nhu Van Nguyen and Antoine Doucet: Deep Analysis of OCR Errors for Post-OCR Processing. JCDL2019

[C3] Nguyen Thi Tuyet Hai, Mickal Coustaty, Antoine Doucet, Adam Jatowt, Nhu-Van Nguyen: Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. ICADL 2018

[C4] Nguyen Thi Tuyet Hai, Antoine Doucet, Mickal Coustaty, "Enhancing Table of Contents Extraction by System Aggregation", The 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017)

[C5] Nguyen Thi Tuyet Hai, Tan Hanh, "Maximal frequent sequences for document classification", Advanced Technologies for Communications International Conference 2016.

[C6] Nguyen Thi Tuyet Hai, Tan Hanh, Huynh Cong Thinh, "Extracting Semantic Relations Between Vietnamese Legislative Documents", The National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS) 2015.

References

[1] S. Impedovo, L. Ottaviano, and S. Occhinegro, "Optical character recognitiona survey," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 5, no. 01n02, pp. 1–24, 1991.

[2] Y. Bassil and M. Alwani, "Ocr post-processing error correction algorithm using google online spelling suggestion," *arXiv preprint arXiv:1204.0191*, 2012.

[3] H. Niwa and K. Kayashima, "Postprocessing for character recognition using keyword information."

[4] F. Ahmed, E. W. De Luca, and A. Nürnberger, "Multispell: an n-gram based language-independent spell checker," in *Proceedings of Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007)*, 2007.

[5] M. A. Jones, G. A. Story, and B. W. Ballard, "Interating multiple knowledge sources in a bayesian ocr post-processor." *International Journal on Document Analysis and Recognition*, pp. 925–933, 1991.

[6] X. Tong and D. A. Evans, "A statistical approach to automatic ocr error correction in context," in *Proceedings of the fourth workshop on very large corpora*, 1996, pp. 88–100.

[7] R. Llobet, J.-R. Cerdan-Navarro, J.-C. Perez-Cortes, and J. Arlandis, "Ocr post-processing using weighted finite-state transducers," in *2010 International Conference on Pattern Recognition.* IEEE, 2010, pp. 2021–2024.

[8] H. Afli, L. Barrault, and H. Schwenk, "Ocr error correction using statistical machine translation." *Int. J. Comput. Linguistics Appl.*, vol. 7, no. 1, pp. 175–191, 2016.

[9] J. Mei, A. Islam, Y. Wu, A. Moh'd, and E. E. Milios, "Statistical learning for ocr text correction," *arXiv preprint arXiv:1611.06950*, 2016.

[10] I. Kissos and N. Dershowitz, "Ocr error correction using character correction and feature-based word classification," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on.* IEEE, 2016, pp. 198–203.

[11] G. Chiron, A. Doucet, M. Coustaty, and J.-P. Moreux, "Icdar2017 competition on post-ocr text correction," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1423–1428.

[12] J. Evershed and K. Fitch, "Correcting noisy ocr: Context beats confusion," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage.* ACM, 2014, pp. 45–51.

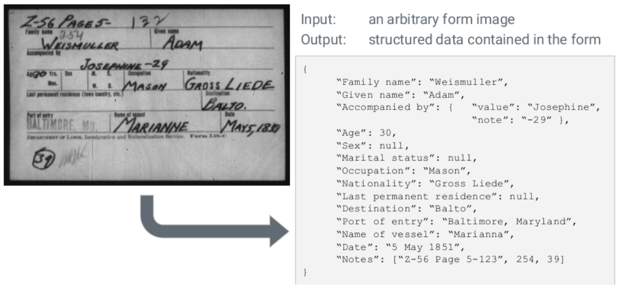
[13] T. Mikolov, "Statistical language models based on neural networks," *Presentation at Google, Mountain View, 2nd April*, 2012.

[14] C. Amrhein and S. Clematide, "Supervised ocr error detection and correction using statisti-cal and neural machine translation methods," *JLCL*, p. 49.

[15] K. Taghva and E. Stofsky, "Ocrspell: an interactive spelling correction system for ocr errors in text," *International Journal on Document Analysis and Recognition*, vol. 3, no. 3, pp. 125–137, 2001.

Template-Free Information Extraction From Arbitrary Form Images

Students name: Brian Davis
Supervisor of the thesis: Dr. Bryan Morse
University: Brigham Young University
Starting date of the PhD: May 2018
Expected finalization date of the PhD: December 2021
Email: briandavis@byu.edu



Z-56 PAGE 5- 132

Family name 254 WEISMULLER | Given name ADAM

Accompanied by JOSEPHINE -29

Age 30 Yrs. Mos. | Sex | M. S. W. D. | Occupation MASON | Nationality GROSS LIEDE

Last permanent residence (town country, etc.) | Destination BALTO.

Port of entry BALTIMORE, MD. | Name of vessel MARIANNE | Date MAY 5, 1851

Department of Labor, Immigration and Naturalization Service. Form 548-C

39

Input: an arbitrary form image
Output: structured data contained in the form

```
{
      "Family name": "Weismuller",
      "Given name": "Adam",
      "Accompanied by": {   "value": "Josephine",
                            "note": "-29" },
      "Age": 30,
      "Sex": null,
      "Marital status": null,
      "Occupation": "Mason",
      "Nationality": "Gross Liede",
      "Last permanent residence": null,
      "Destination": "Balto",
      "Port of entry": "Baltimore, Maryland",
      "Name of vessel": "Marianna",
      "Date": "5 May 1851",
      "Notes": ["Z-56 Page 5-123", 254, 39]
}
```

Figure 1. An example of our goal. Novel form image in and data out.

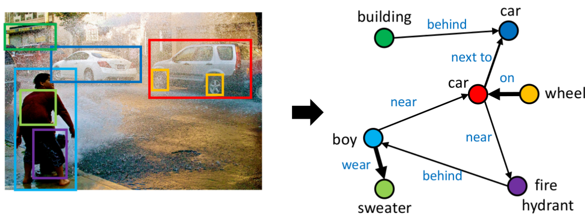


Figure 2. An example of a scene graph, from [1].

***Abstract*—Semi-reliable, end-to-end, template-free, form recognition can be achieved leveraging current deep learning techniques. Currently most research has focused on extraction of from forms with a template or other means of alignment. Being able to extract information from a form which is only seen once is a more challenging problem. We aim to do this leveraging several deep learning techniques as a coherent solution.**

I. Introduction

A. Problem Definition

Our problem is the extraction of information from form images (which requires what some term "form understanding") in a template-free manner. See Fig. 1.

A solution can be given an image of a form, whose field types and layout may have never been seen in developing/training the solution, extract the pre-printed text, input text, and the relationships between these entities such that the *meaning* of the input text may be inferred. *Meaning* is somewhat ambiguous, so we define it as a piece of information (like a label) that would allow the information to be incorporated into a database or other upstream task. In the end, almost all input information should be formulated into label-value format.

For our work we define a form to be a paper document which has pre-printed text (instructions, labels, prompts), pre-printed structure (layout, boundaries, boxes, blank lines) and input text (handwriting, stamps, type-writer entry, checks, notes etc.).

B. Use Cases

There are several companies which provide automatic data extraction from form images as a service to automate data collection/entry, but these are template-based solutions (e.g. Parascript, SoftWorks AI, ABBYY FlexiCapture). The system must be updated for new forms to be handled. While this is very practical where a large number of each type of form are needing processed, this is inefficient if there are a few documents each of many form types. A template-free system can remove the template creation overhead.

II. Proposed Method

We are formulating the extraction of information from forms very similarly to how scene graphs are created in computer vision literature (e.g. [1], [2], see Fig. 2 for an example scene graph). For us the objects are text lines. Understanding the form is identifying the relationships between them. The forms problem has the added difficulty that discerning the precise class of the objects is related to automatic text recognition (ATR).

We have several basic tasks:

1) Detect text lines: this is the object detection in scene graph generation
2) Read information: Perform ATR on the text lines
3) Detect relationships: this builds the graph

As we'll discuss later, these are not necessarily linear tasks.

A. *Text Detection*

This is a foundational step in the overall extraction process and it is also probably the step with a strongest prior work. Both [3] and [4] have excellent results and both use fully convolutional networks. We will likely follow in their footsteps. We use a fully covolutional approach in [5], but it is suboptimal compared to these other works.

B. *Automatic Text Recognition*

To actually extract the information we need to read what's on the page. OCR of printed text we consider a solved problem, and we don't intend to make any innovations in this area. Handwriting recognition has come a long way in the past 5 years. However, it still has room for improvement. Particularly we would like to leverage transfer learning as our forms datasets will be more limited than handwriting databases. We have already begun work on creating synthetic examples of handwriting, conditioned on the desired text and style for dataset augmentation. We hope that this will aid in allowing our system to have fairly reliable handwriting recognition even in datasets we have few handwriting examples for.

C. *Detect Relationships*

Here we build the "scene graph" by finding the relationships between the detected text lines. We hypothesized in [5] that most label-value relationships could be determined using only visual features. We did this by having a heuristic to determine possible relationships and had a classifier determine which possible relationships were real. The classifier received a context window around the potential relationship.

A purely local classification of relationships is suboptimal as relationships may be in contest over the same text line (e.g. a label generally only has one value, so if there are two potential relationships with values, one should be rejected). In [5] we formulated an optimization which satisfies a number of neighbors value predicted for each text line. We feel that a more intelligent method would be to follow in the footsteps of works like [1] and use a graph network to predict. A graph network would allow richer features to be used in the decision making.

D. *Putting It Together*

We've described a modular and linear process for extracting information from forms. However, we believe that the best solution will not be so simple.

1) Using language: It should be obvious that our pairing method in [5] is suboptimal as it does not read the text. A line of pre-printed text which reads "First name" and a line of input text which reads "John" will have a high probability of forming a *label-value* relationship; in contrast if the input text read "Ireland" the relationship would be very unlikely. We would like to include this idea into our relationship detection. It could have benefit in both the relationship proposal component and relationship classification.

But how do we include it? An approach appealing to us is the idea behind word/sentence embeddings (e.g. word2vec, sentence2vec). These could be learned from a larger text corpus than one specifically of forms. This is more likely to get us more vocabulary coverage than just using data from forms. Once a text line is detected it can be transcribed using ATR and then be embedded into a vector. This vector becomes part of the detection's features for later processes. An alternate approach, which may work better for broken text lines (multiline text), is to use unsupervised language models like BERT [6] or GTP-2 [7] to predict the likihood of a given text line continuing into another. Further, language models will be able to correct ATR errors.

2) Iterative refinement: While text detection is the first step in the extraction process, information obtained from reading the text and using a language model can indicate detection errors such as a merged text line or a split line. It should also be seen that just as using language will help relationship predictions, it can also help in reading (e.g. if a label reads "first name" then the value should read as a name). Thus the best results should come not as treating the extraction as simply a linear process, but a cyclic one where we revisit prior steps with our downstream predictions.

E. *The Hard Stuff*

There are a couple of structures we've observed in forms which will present a challenge to how our method operates.

1) Fill-in-the-blank prose: Some forms have sections where there is running prose (sentences, paragraphs), with blanks where the filler of the form is to write in information. We will extract the sentences with the blanks filled in, but will end our processing there. Extracting the appropriate labels from this text becomes a NLP tasks which is beyond the scope of what we want to address.

2) Tables: Tables have unique label-value relationships that are structured in a very predictable way. These relationships would be exceptionally complicated to capture using our scene-graph-like approach, but are easily captured if the geometry is understood and a table specific method is used. We will limit ourselves to table detection, excepting another method to be more effective.

III. The Work Ahead

There is very little prior work in template-free form processing. We already have a preliminary work on identifying label-value pairs [5] and solid work on ATR on non-form documents [3]. We next need to include ATR in the from parsing process. Following this, the items in II-D will need addressed, along with other unforseen problems, to get our method to work well.

References

[1] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph r-cnn for scene graph generation,” in *The European Conference on Computer Vision (ECCV)*, September 2018.

[2] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, “Relationship proposal networks,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] C. Wigington, C. Tensmeyer, B. Davis, W. Barrett, B. Price, and S. Cohen, “Start, follow, read: End-to-end full-page handwriting recognition,” in *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] T. Grüning, G. Leifert, T. Strauß, J. Michael, and R. Labahn, “A two-stage method for text line detection in historical documents,” *International Journal on Document Analysis and Recognition (IJDAR)*, Jul 2019.

[5] B. Davis, B. Morse, S. Cohen, B. Price, and C. Tensmeyer, “Deep visual template-free form parsing,” in *2019 15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Sep 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp. 4171–4186.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, 2019.

Transcription and Indexing of Archival Documents Using Deep Architectures

Student's name: Olfa Mechi
Supervisor/s of the thesis: Maroua Mehri, Rolf Ingold
and Najoua Essoukri Ben Amara
University: University of Sousse
Starting date of the PhD: December 01, 2017
Expected finalization date of the PhD: December 01, 2021
Email: olfamechi@yahoo.fr

Abstract. Developing efficient and robust multilingual document image transcription, indexing and retrieval systems has been considered challenging research topics for the research community working in historical document image analysis (HDIA). Recently, the deep approaches have become an interesting alternative to the classical image processing-based methods to tackle many challenges related to HDIA. Text line segmentation remains one of the most important preliminary task for document image transcription, indexing and retrieval systems. Thus, a novel deep method for text line segmentation is proposed in my first thesis work. To illustrate the effectiveness of the proposed text line segmentation method, qualitative and numerical experiments are given using a large number of historical document images collected from the Tunisian national archives and different recent benchmarking datasets provided in the context of ICDAR and ICFHR competitions.

Keywords: Text transcription · Document indexing · Historical document images · Deep architectures.

1 Introduction

During the past few decades, numerous initiatives through many research projects and studies have taken place to exploit and preserve cultural heritage. Moreover, due to the large digitization programs conducted by the Tunisian National Archives (ANT) [1], an important need of a robust and accurate text transcription system has been emerged. Nevertheless, the conventional optical character recognition tools used by ANT archivists are hindered by many issues related to the idiosyncrasies and particularities of the ANT collections. The digital collections of the ANT encompass more than six centuries (1500-2000) and compose primarily of printed and manuscript image documents written in Arabic and Latin [1]. In this context, the work conducted in my research project, which is in collaboration with the ANT, aims at providing effective tools, that assist ANT

[1] http://www.archives.nat.tn/

archivists to transcribe their heritage documents automatically on the one hand, and to develop content-based retrieval and indexing tools on the other hand. On most document image transcription, indexing and retrieval systems, text line segmentation remains the most fundamental preliminary tasks. Thanks to the computer hardware and software evolution, several methods based on using deep architectures continue to outperform the existing classical state-of-the-art methods, which are proposed to meet the pattern recognition issues and particularly those related to HDIA [2]. Thus, our thesis work aims to propose a novel deep framework able to segment text lines and recognize words in Arabic and Latin historical document images.

2 Methodology

In my first thesis work, a method based on using an adaptive U-Net architecture for text line segmentation in historical document images is proposed. One of the main advantages of choosing the U-Net-based architecture is its effectiveness in medical image analysis. Besides, images of variable sizes can be fed as input of the U-Net architecture. Moreover, the training phase does not require large volumes of images. In the proposed architecture, we use the "Conv2DTranspose" operation for the decoder phase in order to keep the same resolution on both the input and output of the network architecture. Another optimization has been proposed in relation to the number of filters in the contracting path of the U-Net architecture. This optimization has led to the reduction of the memory requirements, the processing time and the numerical complexity of the network on the one hand, and to the elimination of the over-fitting issues in the training phase on the other hand. We have shown the effectiveness of the proposed architecture for segmenting text lines using a large number of historical document images collected from the ANT and different datasets provided in the context of recent open competitions at ICDAR and ICFHR conferences.

3 Future work

The first aspect of our future work will be to refine the text line segmentation results. Moreover, we will propose a thorough performance benchmarking of the most recent and widely used deep architectures for text line segmentation.

Acknowledgment

The authors would like to acknowledge the ANT for providing access to their digital collections.

References

1. Elhedda, W., Mehri, M., Mahjoub, M.A.: A comparative study of filtering approaches applied to color archival document images. ACIT (2017)
2. Zayene, O., Essefi Amamou, S., Essoukri Ben Amara, N,: Arabic Video Text Recognition Based on Multi-Dimensional Recurrent Neural Networks. ICDAR (2017)

Table Information Extraction from Business Documents

PhD Student: Clément Sage*†
Supervisors: Alexandre Aussem*, Haytham Elghazel*, Véronique Eglin* and Jérémy Espinas†
*Univ Lyon, CNRS, LIRIS UMR 5205, F-69100, VILLEURBANNE, France
{clement.sage, alexandre.aussem, haytham.elghazel, veronique.eglin}@liris.cnrs.fr
†Esker, F-69100, VILLEURBANNE, France
{clement.sage, jeremy.espinas}@esker.fr

Starting date of the PhD: March 2018
Expected finalization date of the PhD: March 2021

***Abstract*—This PhD thesis is about extracting predefined sets of tabular information from business documents having unconstrained layouts. To develop generic extraction models, we resort to Machine Learning techniques and especially Recurrent Neural Networks (RNN) taking as input the sequence of document words. Words are segmented by an external OCR engine if needed and represented by textual and spatial features.**

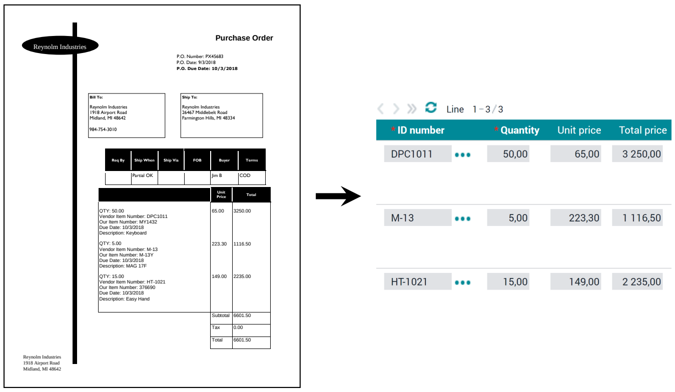
We proposed a RNN classifier based approach tagging each word with one of the possible information type to extract. Business specific post processing heuristics then reconstruct the desired table entities from the word classifier predictions. In our ICDAR 2019 paper, we empirically demonstrated the usefulness of recurrent connections in the word classifier for extracting table fields from documents with unknown layouts.

In the rest of the thesis, we plan to use encoder-decoder RNN architectures and particularly pointer-generator networks in order to directly output structured table entities from the document words. This new approach will be evaluated against the word classifier based model on a large dataset of business documents.

I. Introduction

Business documents, whose main exchanged classes are invoices and purchase orders, contain valuable information that companies want to retrieve for further processing such as integration in their information system and structured archiving. Even if Electronic Data Interchange (EDI) of documents is progressively spreading, a large part of daily issued business documents is still printed on paper or generated in digital format such as PDF, thus requiring an information extraction step. If performed manually, this additional task is time-consuming for employees in charge of document processing, especially in companies facing huge incoming document flows. Yet, automating information extraction from business documents is challenging. Even if the set of information types to extract is known and fixed for a given document class, the positioning and textual representation of the information to retrieve are unconstrained [1]. Indeed, every document issuer is free to generate business documents with a specific layout (also called template) and change it when desired. Therefore, the developed extraction model must be template agnostic, i.e. able to extract information even if the system processes a document template for the first time. Moreover, the method should be generic enough to be applicable to a broad range of document classes, information types to extract and languages at minimal configuration cost.

In this thesis, we are particularly interested in retrieving information contained in tables of the business documents. This data type is more challenging to extract than non tabular information as tables contain entities that are structured, unlike the flat entities found outside tables such as document date or number. For example, extracting information from purchase orders may require to retrieve each ordered item with its field values (see Fig. 1).



Reynolm Industries

Purchase Order

P.O. Number: PX45683
P.O. Date: 9/3/2018
P.O. Due Date: 10/3/2018

Bill To:

Reynolm Industries
1918 Airport Road
Midland, MI 48642

984-754-3010

Ship To:

Reynolm Industries
26467 Middlebelt Road
Farmington Hills, MI 48334

Req By	Ship When	Ship Via	FOB	Buyer	Terms
	Partial OK			Jim B	COD

	Unit Price	Total
QTY: 50.00 Vendor Item Number: DPC1011 Our Item Number: MY1432 Due Date: 10/3/2018 Description: Keyboard	65.00	3250.00
QTY: 5.00 Vendor Item Number: M-13 Our Item Number: M-13Y Due Date: 10/3/2018 Description: MAG 17F	223.30	1116.50
QTY: 15.00 Vendor Item Number: HT-1021 Our Item Number: 376690 Due Date: 10/3/2018 Description: Easy Hand	149.00	2235.00
	Subtotal	6601.50
	Tax	0.00
	Total	6601.50

Reynolm Industries
1918 Airport Road
Midland, MI 48642

Line 1 – 3 / 3

* ID number	* Quantity	Unit price	Total price
DPC1011	50,00	65,00	3 250,00
M-13	5,00	223,30	1 116,50
HT-1021	15,00	149,00	2 235,00

Fig. 1. Illustration of our table information extraction objective for the purchase order class. We aim at retrieving the ordered products and their related fields, i.e. ID number, quantity, unit and total prices.

So far, we have considered that text recognition is out of the scope of the thesis. When dealing with scanned documents, we use an external commercial OCR engine to retrieve their words and do not perform any post corrections on its results.

II. Related work

Methods for extracting information from business documents were historically based on significant domain specific knowledge either encoded in hand-crafted patterns and rules [2], [3], in constraints of a optimization problem [4] or included in a morphological approach [5]. These approaches work well for extracting information in rather structured documents with small layout variability but are not easily adaptable to other information types or languages than ones for which they were designed for.

Our information extraction task is also closely related to Named Entity Recognition (NER) problem where Recurrent Neural Networks (RNN) currently constitute the state of the art methods [6]. So, Palm et al. [7] resort to this neural architecture for extracting main non tabular information from invoices. To that end, the RNN iterates over the document words which are represented by textual and spatial features in order to tag each word with one of the possible information type to retrieve. Then, word class predictions are refined by post processing heuristics, such as verification that the extracted words are compliant with the corresponding expected syntax (e.g. a word identified as a total must be parsable as an amount) and with business constraints (e.g. verify that tax total = sub total x tax percentage).

III. Completed work

Inspired by the works of Palm et al., we used for our extraction objective a word level RNN classifier identifying individual table fields. We validated our approach on a private dataset of 28,570 English purchase orders for which we wanted to retrieve the ID number and quantity of the ordered products. See our paper accepted to ICDAR 2019 [8] for more details about this word classifier. Main findings of this paper are that recurrent connections are helpful to capture dependencies between word labels resulting in greater extraction performances than a baseline feedforward neural network and that our model is able to extract rather well table fields for templates not seen during classifier training. Since then, we've also shown that using a character level RNN for generating textual components of word feature vectors is more effective than word level learnable embeddings with respective micro F1 scores of 0.847 and 0.821 on unknown templates for the resulting extraction model.

Structured table entities, i.e. the ordered products, are then constructed by domain specific post processing heuristics. To that end, ID number and quantity field instances detected by the word classifier are paired by solving a linear sum assignment problem with vertical distances on the document between instances as matching costs. Hungarian algorithm [9] is used for the solving. This pairing approach is optimal if table field extraction is perfectly performed by the word classifier but can cause serious inconsistencies in the opposite case, e.g. matching of ID number instances with the quantity instances of the products above/below in the table. Moreover, post processing operations must be consequently modified if we want to extract tables entities that are not structured the same way as ordered products.

IV. Perspectives

Future works will explore models able to directly output structured table entities from the document words, thus avoiding a post processing step and making our approach more easily adaptable to other structured information extraction tasks. To do this, we could resort to encoder-decoder RNN architectures augmented with attention mechanisms [10] that currently constitute the state-of-the art methods for sequence to sequence learning. Particularly, we could use pointer-generator networks [11] which would generate the field XML-like tags structuring the output and point to the sequence of input words in order to retrieve the actual field values. To validate the effectiveness of this new approach, product recognition performances will be compared with results of the word classifier based model. Finally, we want to evaluate the behaviour of our extraction models when confronted with a multilingual dataset.

Acknowledgements

This is an industrial thesis supported by the company Esker, a software editor of cloud-based business processes automation solutions.

References

[1] M. Cristani, A. Bertolaso, S. Scannapieco, and C. Tomazzoli, "Future paradigms of automated processing of business documents," *International Journal of Information Management*, vol. 40, pp. 67–75, 2018.

[2] D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, and A. Hofmeier, "Intellix–end-user trained information extraction for document archiving," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 101–105.

[3] M. Rusinol, T. Benkhelfallah, and V. Poulain dAndecy, "Field extraction from administrative documents by incremental structural templates," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1100–1104.

[4] F. Deckert, B. Seidler, M. Ebbecke, and M. Gillmann, "Table content understanding in smartfix," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 488–492.

[5] Y. Belaïd and A. Belaïd, "Morphological tagging approach in document analysis of invoices," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1. IEEE, 2004, pp. 469–472.

[6] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2145–2158.

[7] R. B. Palm, O. Winther, and F. Laws, "Cloudscan-a configuration-free invoice analysis system using recurrent neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 406–413.

[8] C. Sage, A. Aussem, H. Elghazel, V. Eglin, and J. Espinas, "Recurrent neural network approach for table field extraction in business documents," 2019.

[9] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[11] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1073–1083.

Multimodal analysis and reconstruction of ancient papyrus fragments using image processing and deep learning

Antoine Pirrone
antoine.pirrone@labri.fr
University of Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800
Bordeaux, France
Thesis supervised by Marie Beurton-Aimar and Nicholas Journet
marie.beurton@labri.fr, journet@labri.fr
Thesis title : Multimodal analysis and reconstruction of ancient papyrus fragments using image processing and deep learning

I. PhD Dates

- PhD Started in December 2018
- Expected to be finalized in December 2021

II. Short research plan

The goal of this PhD is to develop methods for dealing with a particularly difficult puzzle solving problem : reconstructing ancient papyrus fragments. This puzzle is difficult because some fragments are missing, most are very damaged and the contour of the fragment is not a reliable information as torn papyri do not form a regular tear as conventional paper do.

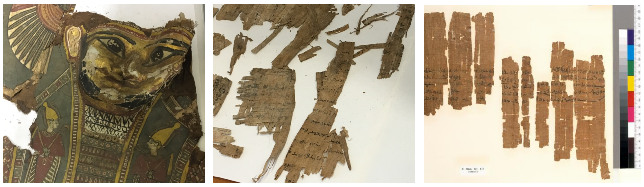


Figure 1. Overview of the fragment extraction and reconstruction process (third image is from the University of Michigan papyrus collection : https://quod.lib.umich.edu/a/apis)

The PhD is part of the GESHAEM Project, which is founded by the ERC (European Research Council), which's goal is to study the content of the Jouguet collection of the Sorbonne. The papyri of this collection come from funeral ornaments that were made to decorate mummies (see first image on fig. 1) about 2000 years ago. At the time, papyrus was an expensive resource, so they actually recycled already used papyrus (covered in writing) to make the ornaments. The funeral workers would tear the papyrus to fit the needed shapes and sizes. Today, when archeologists find such documents, they are a very rich and interesting source of information on the economic and administrative life of the era. This is why they must be carefully extracted from the ornaments (see second image on fig. 1) and reconstructed in order to be studied (see third image on fig. 1).

We worked with the Papyrologists of the project to define the acquisition protocol of the collection, taking photos both sides of the fragments in color and infra-red with a reference ruler and aruco tags. Additionnaly, meta-data is produced manually by the papyrologists for each digitized fragments.

The PhD was started about 9 months ago. So far experiments on segmentation, line extraction and simple contour matching methods have been conducted, leading to the preliminary conclusions that using only one modality of the data is not enough to solve the problem. We then started working on a subproblem, which is sorting the fragments. A paper [1] has ben submitted and accepted to the HIP2019 (Historical Document Imaging and Processing) workshop. In this article, we describe the first version of a tool that aims to help the Papyrologists to sort the fragments according to an image similarity criterion. Our main contribution is the creation of a deep siamese network architecture to determine if two fragments belong to the same papyrus. We obtained encouraging results, and this solution can already be used to perform a significant filtering on the database. We believe this is a good first step towards the ultimate goal of full reconstruction, may it be automatic or semi-automatic.

Over the next months, the goal is to improve the results by trying different architectures variations (like triplet networks) and experimenting with data augmentation. We are also in the process of building a larger dataset (from the University of Michigan Collection[1]) with sufficient ground truth data to perform thorough testing of our algorithms. This is a tedious process as there are no pre-made databases formatted in a way to be directly useful to us.

[1] https://quod.lib.umich.edu/a/apis

As of now, we only tackled the subproblem of sorting the database. The next step is to actually try to solve the puzzle, meaning aligning the fragments together to reconstruct the original papyrus. Currently, our most promising leads would be to use a combination of meta-data analysis (content of the text written on the fragment, presence of margins indicating that the fragment is located on an end of the papyrus . . .), layout analysis (if the text is cut on the border of a fragment, we could use the local boundary information to find matching candidates for example) and maybe texture analysis (papyrologists use the direction and pattern of the fibers of the papyri to find matching candidates). Our first experiments on contour matching made us think that contour alone is not sufficient as a feature. Indeed, the borders of two fragments next to each other in a papyrus don't necessarily match as they are often damaged by the original tearing process and time.

In the end, the goal is to provide the papyrologists with useful piece of software that they can use to process their data. To do that, we believe that the most promising approach to solve this complex problem is to combine multiple complementary approaches and to involve them as much as possible. A fully automatic solution may not be feasible, so we are also investigating on semi-automatic solutions that would prompt expert users for specific inputs that could provide rich and sementical information in some cases.

References

[1] Antoine Pirrone, Marie Beurton Aimar, and Nicholas Journet. Papy-S-Net : A Siamese Network to match papyrus fragments (to be published).

Recognize Scene Text and Handwritten Text with Segmentation-free Methods

Student's Name: Qingqing Wang
Supervisors of the Thesis: Yue Lu, Xiangjian He
Co-supervisor of the Thesis: Michael Blumenstein
University: East China Normal University, China & University of Technology Sydney, Australia
Starting Date of the PhD: September, 2013
Expected Finalization Date of the PhD: December, 2019
Email: Qingqing.Wang-1@student.uts.edu.au

Abstract—**Text recognition is of great importance in the field of computer vision. Traditional solutions to scene/handwritten text recognition are usually segmentation-based or over-segmentation-based, where hand-crafted features and well-designed classifiers play critical roles. However, structures of these methods are very complicated and their performance is seriously limited by the quality of extracted features, classifiers, segmentation results, path searching strategies, etc. Recently, with the development of deep learning techniques, segmentation-free models have been the dominated solution to the recognition of both scene text and handwritten text. In this thesis, we study segmentation-free recognizers that utilize deep learning techniques like Convolutional Neural Networks (CNN), Fully Connected LSTM (FC-LSTM), Convolution LSTM (ConvLSTM), etc.**

I. Short Research Plan

A. Introduction

Text presents everywhere in our daily life to convey us important information and knowledge, such as notes on books, license plates on cars, product descriptions on bags and road signs on direction boards etc. Automatically reading text from images is of great application potentials with the dramatic development of techniques, especially those on mobile devices.

Traditional scene text recognition pipeline is usually composed of three main components, i.e., pre-processing, character segmentation and single character recognition. Here, character segmentation, namely precisely discriminating pixels belonging to characters from those belonging to backgrounds, is the most challenging part of scene text recognition due to the complicated backgrounds and unconstrained imaging conditions (skew, perspective skew, uneven illumination, etc). According to literature, though many efforts have been made from the perspectives of binarization, graph-based prediction and word matching, the inaccurate segmentation results are still the bottleneck of scene text recognition. Compared with scene text recognition, Handwritten text recognition does not suffer from problems like complex background and unconstrained imaging conditions, but the diverse writing styles and serious torching-character problems have posed huge challenges. Therefore, traditional solutions to handwritten text recognition usually segment input text images into components corresponding to single characters or portion of characters, and then form a candidate lattice with these components, followed by searching optimal paths with considering single character recognition results.

In literature, these traditional recognizers are called segmentation-based or over-segmentation-based models. However, structures of these models are usually very complicated since many operations are required to pre- or post-process the inputs and outputs of the classifiers, especially those related to segmentation. To avoid the intractable segmentation problem, inspired by speech recognition and machine translation, the community also turns to segmentation-free methods for text recognition, especially after the introduction of deep learning. Nowadays, as clarified in [1], almost all of the state-of-the-art text recognizers are on the base of deep networks. Therefore, in this thesis, we study segmentation-free solutions to scene/handwritten text recognition.

Existing state-of-the-art approaches regard text recognition as a sequence-to-sequence prediction problem and widely adopt techniques like LSTM [2] and attention mechanism [3] in their sequential transcription module. However, the LSTM used in these recognizers is the fully-connected-LSTM (FC-LSTM) that only takes stream signals like sentences or audio as inputs and connects them in a fully connected way, while text recognition generates sequential outputs from 2-D images. To adapt FC-LSTM to text recognition, the most straightforward way is pooling 2-D feature maps to a height of one or flattening them into 1-D sequential feature vectors. Unfortunately, such operations could severely disrupt the valuable spatial correlation relationships among pixels, which is essential to computer vision tasks, especially to text recognition, where the structures of strokes are the key factors to discriminate characters. To retain such important spatial and structural information, in this work, we proposed three solutions. More details will be presented below.

B. Current Progress

In this work, we argue that scene text recognition is essentially a spatiotemporal prediction problem for its 2-D image inputs, and propose a convolution LSTM (ConvLSTM)-based scene text recognizer, namely, FACLSTM, i.e., Focused Attention ConvLSTM, where the spatial correlation of pixels is fully leveraged when performing sequential prediction with LSTM. Particularly, the attention mechanism is properly incorporated into an efficient ConvLSTM structure via the convolutional operations and additional character center masks are generated to help focus attention on right feature areas. Experimental results demonstrate that the proposed FACLSTM can achieve promising performance on regular text, low-resolution text and noisy text, and outperforms other state-of-the-art approaches significantly on the more challenging curved text. Moreover, we also propose a recognizer named ReELFA, namely scene text Recognizer with Encoded Location and Focused Attention, to improve the recognition performance of traditional FC-LSTM models. Our proposed ReELFA consists of two modules, *i.e.,* an encoder-decoder feature extraction module, which is the same as the one used in FACLSTM, and an attention-LSTM-based sequence transcription module.

In addition, handwritten string recognition has been struggling with connected patterns fiercely. Segmentation-free and over-segmentation frameworks are commonly applied to deal with this issue. For the past years, RNN combining with CTC has occupied the domain of segmentation-free handwritten string recognition, while CNN is just employed as a single character recognizer in the over-segmentation framework. The main challenges for CNN to directly recognize handwritten strings are the appropriate processing of arbitrary input string length, which implies arbitrary input image size, and reasonable design of the output layer. In this work, we propose a sequence labeling convolutional network for the recognition of handwritten strings, in particular, the connected patterns. We properly design the structure of the network to predict how many characters present in the input images and what exactly they are at every position. Spatial pyramid pooling (SPP) is utilized with a new implementation to handle arbitrary string length. Moreover, we propose a more flexible pooling strategy called FSPP to adapt the network to the straightforward recognition of long strings better. In particular, in the proposed network, Four convolutional layers and two mean pooling layers are designed before an Inception module, which is followed by a SPP or FSPP layer. Then the fixed-length features extracted by the SPP or FSPP layer are fed into two fully-connected layers that assembled before the final classifier layer. If we regard SPP or FSPP layer as extracting global features at multiple scales, and Inception module as extracting local features at multiple scales, setting a SPP or FSPP layer after an Inception module will be beneficial for the model to extract richer information. The goal of the target network is to predict the order concerned sequential characters. Intuitionally, information from two aspects needs to be extracted: (1) how many characters are contained in the input images, and (2) what exactly it is at each position. So we equip the proposed network with two modules named CM, which is short for ‘counting module’, and PM, which is short for ‘prediction module’, to predict corresponding information. Apparently, the proposed network is supposed to have a cluster of classifiers rather than a single one. Therefore, we need to blend prediction errors of these classifiers into one structure when we design the cost function for the network. Experiments conducted on handwritten digital strings from two benchmark datasets and our own cell-phone number dataset demonstrate the superiority of the proposed network.

C. Future Work

Recently, Neural Network Search (NAS), i.e., automatically searching network structures with a super network, shows great potentials in the community of deep learning. According to literature, networks searched by NAS achieve better efficiency and effectiveness than manually designed networks. Therefore, we are now seeking more powerful deep models via NAS for scene text detection and scene/handwritten text recognition. In addition, data scarcity is a common problem in the field of computer vision. Domain adaption and one/few shot learning are the most widely researched solutions to this problem. Hence, we will also conduct research in this direction.

II. Reference

1 S. Long, X. He, and C. Yao, “Scene text detection and recognition: the deep learning era,” CoRR, vol. arXiv preprint arXiv: 1811.04256v3, 2018.

2 S. Hochreiter and J. Schmihuber, “Long short-term memory,” Neural Computation, vol. 9, pp. 1735–1780, 1997.

3 D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in ICLR, 2015.

Interactive Systems for Reading Texts in Indian Streets and Documents

Student's name: Rohit Saluja
University: IITB-Monash Research Academy, Mumbai, India
Title of the thesis: Interactive Systems for Reading Texts in Indian Streets and Documents
Supervisors of the thesis: Dr. Ganesh Ramakrishnan, Dr. Parag Chaudhuri and Dr. Mark Carman
Starting date of the Ph.D.: 31 December 2014
Expected finalization date of the Ph.D.: 31 December 2019
Email: rohitsaluja22@gmail.com

Abstract—**It has been an integral part of humans' journey to try that machines mimic us to ease our jobs such as reading, hearing, and typing. Optical Character Recognition (OCR), the process of converting document or scene text images to editable electronic format, is one of the outcomes of such human desires. We have developed a novel methodology for labeling a large amount of training data in Indian traffic videos. We then present the first results of multi-head attention models on the task of text recognition. For reading license plates in chaotic Indian streets, we observe gains as large as 7.18% over the baseline model by incorporating multi-headed attention. Our models also outperform state-of-the-art results on French Street Name Signs dataset (FSNS) and IIIT-ILST Devanagari dataset by 1.1% and 8.19% respectively. We additionally release a new multi-lingual data-set of 1000 videos (with text in Hindi, Marathi, and English) each covering an Indic street board from different orientations. We then present StreetOCRCorrect: a novel framework for OCR corrections in chaotic street videos. We further used Long Short Term Memory (LSTM) networks to correct OCR errors in Indian documents for four different Languages with varying inflections. Our LSTM model when tuned with appropriate delay, that depends on OCR confusion patterns, outperforms the state-of-the-art results. We further show that sub-word embeddings, derived from language or training data itself, can help to correct OCR errors more effectively in Indic Languages. We finally present OpenOCRCorrect: an interactive framework for end-to-end corrections of errors in Indic OCR. The system updates a domain vocabulary and the document-specific OCR confusions on the fly and helps reduce the human efforts for documents in different Indian Languages with different inflections. We aim to progress forward in direction of reading scene text in videos via text saliency or attention map that update through trajectory in the temporal domain for stable and efficient systems.**

I. Short research plan

Obtaining a high-quality OCR output in smart cities, with human-in-the-loop, is an interesting problem for surveillance, record keeping, and other similar applications. Achieving high accuracy while reading the street text in videos is cumbersome due to complexities like multiple vehicles, high-density traffic in spatial and temporal domains, occlusions and resolutions.

We first leverage state-of-the-art text spotters to generate a large amount of noisy labeled training data [1]. The data is filtered using a pattern derived from domain knowledge. We augment the training and testing data with interpolated boxes and annotations that make our training and testing robust for reading the text in videos. Further use of synthetic data increases the coverage of the training process. We train two different models for street text recognition. Our baselines include black box detectors such as Convolution Neural Network (CNN) and humans, followed by the Recurrent Neural Network (RNN) based recognizers. As our first contribution, we bypass the detection phase by augmenting the baseline with an Attention mechanism in the RNN decoder. Next, we build in the capability of training the model end-to-end on scenes containing license plates by incorporating inception based CNN encoder that makes the model robust to multiple scales. A salient point of our framework is that our models, when trained only on a combination of noisy labeled data and clean synthetic data and when appropriately tuned, set new benchmarks for the task. Moreover, we present the first results of using multi-headed attention models on end-to-end text recognition in images and illustrate the advantages of using multiple heads over a single head. We illustrated the application of multi-head attention in two scenarios: (i) recognizing license plates automatically in chaotic traffic conditions, a task for which we curated our dataset and (ii) the existing publicly available FSNS and IIIT-ILST Devanagari datasets.

We then present StreetOCRCorrect: a modular framework for OCR corrections in the chaotic Indian traffic videos [2]. The patterns used in our framework are obtained from the outputs of a state-of-the-art deep learning model. To ease the correction process, our human-interactive framework i) breaks down the multi-vehicle videos into multiple clips, each containing a single vehicle from the video and ii) provide suggestions for an individual vehicle using consensus in the temporal domain. We perform the breakdown of videos in the spatial domain and the temporal domain using an object detector and a video tracker respectively. Our framework then selectively presents these extracted clips to the user to verify/correct the predictions with minimal human efforts via interactive suggestions. Such high-quality output can be used to continuously update a large database for surveillance as well as training or improving deep models on video data. We use StreetOCRCorrect to generate license plate dataset which we further use to improve the recognition accuracy of license plate recognition model used in OCR-on-the-go [1]. The model was trained on the video data obtained from 3 different sources at 3 different weather conditions. Since the deep learning techniques require a large amount of data, we additionally collected 100 hours of traffic video data from 15 different sources. With 5 annotators working for a total of 15 hours each and 3 reviewers working for 8 hours each, we generate 2.67 million high-quality image-level labeled dataset. Thus we generate a large amount of dataset in less than 100 man-hours, 80% of which we used for domain adaptation of the pre-trained OCR-on-the-go model. The dataset helps us in improving the sequence accuracy (exact match) of attention-ocr model from 41% to 81% on the test set (20 hours of video). We also observe that it takes 4 hrs to manually annotate the sample 1 hour video whereas we annotate the same 1-hour video in 55 minutes using the StreetOCRCorrect. This demonstrates the effectiveness of multi-frame consensus used in our framework. The source code of our framework is available at https://github.com/rohitsaluja22/StreetOCRCorrect. As future work, we would like to extend this work to general scene text recognition in videos. We would also like to leverage error detection to actively improve the models. The OCR text in videos can be stabilized with the help of text saliency or attention map that updates through trajectory in the temporal domain. This can also help in fast multi-frame consensus.

Investigations in the field of document OCR demonstrate that texts in Indic Languages contain a large proportion of out-of-vocabulary (OOV) words due to frequent fusion using conjoining rules (of which there are around 4000 in Sanskrit). OCR errors in the documents further accentuate this complexity for the error correction systems. Using Open Source and Commercial OCR systems, we have observed the Word Error Rates (WER) of around 20-50% on typewriter printed documents according to our experiments. Moreover, developing a highly accurate OCR system with accuracy as high as 90% is not useful unless aided by the mechanism to identify errors. We train Long Short Term Memory (LSTM) with a fixed delay to jointly learn the language and correction patterns [3]. The model corrects the wrong OCR words, and abstain from changing the correct words. We use the dataset of around 100k words for four Indian languages with varying inflections. Our model outperformed the previous results for error detection [4]. Our model jointly performs error detection as well as error correction. We also set a new benchmark for correcting Out of Vocabulary (OOV) errors using the LSTM model, achieving a decrease in overall WER by at least 26.7% & at least 63.3% of the erroneous words were corrected by our model for the four languages. We further improve the error correction results by using sub-word embeddings [5]. As a baseline, we use sub-words within a context window around the OCR characters to be corrected. We append the OHE input of the LSTM model with the frequency of such sub-words in the ground truth of training data. Such a baseline outperform the state-of-the-art models for three Indic Languages. We further experiment with the concatenation of fastText embedding vectors, pre-trained on different datasets, with the OHE input of LSTM. We present that our baseline model works similar to the fastText embeddings when pre-trained on the same data from which we derive the frequencies for the baseline model. We also present a better procedure of training fastText with all possible substrings of the desired length. Our models set new benchmarks for the task of Indic OCR Correction.

We finally present OpenOCRCorrect, an adaptive framework for interactively correcting OCR errors in the Indian documents [6]. In our framework, we incorporate the ability to identify, segment and combine partially correct word forms which are obtained from corrected parts of the document itself as well as auxiliary sources such as dictionaries and common OCR character confusions. Our framework also leverages consensus between outputs of multiple OCR systems on the same text as an auxiliary source for dynamic dictionary building. The framework updates a domain dictionary and learns OCR specific n-gram confusions from the human feedback on-the-fly. Experimental evaluations confirm that for highly inflectional Indian languages, matching partially correct word forms an result in a significant reduction in the amount of manual input required for correction. Furthermore, significant gains are observed when the consolidated output of multiple OCR systems is employed as an auxiliary source of information. We have also presented the benefits of reduction in human efforts due to OpenOCRCorrect for four different languages. In Figure 1 we have shown the overall decrease in average time for correcting the OCR errors in the documents of four different Indian languages. This is possible due to adequate error detection, using adequate color coding (backed off by word conjoining rules) for the incorrect words, updating the auxiliary sources on-the-fly and providing adequate suggestions using such auxiliary sources. We have corrected over 12000 document images at our institute using OpenOCRCorrect. The source code of our framework is available at https://github.com/rohitsaluja22/OpenOCRCorrect.

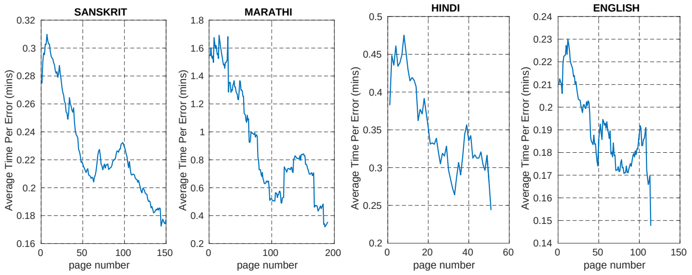


Fig. 1. System analysis of documents in different Languages. For Indian Languages, there is overall decrease in time per error as the user progresses in page number. The system also works well for the English document.

References

[1] R. Saluja, A. Maheshwari, G. Ramakrishnan, P. Chaudhuri, and M. Carman, "OCR On-the-Go: Robust End-to-end Systems for Reading License Plates & Street Signs," in *15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[2] P. Singh, B. Patwa, R. Saluja, G. Ramakrishnan, and P. Chaudhuri, "StreetOCRCorrect: An Interactive Framework for OCR Corrections in Chaotic Indian Street Videos," in *2nd ICDAR Workshop on Open Services and Tools for Document Analysis (ICDAR-OST)*, 2019.

[3] R. Saluja, D. Adiga, P. Chaudhuri, G. Ramakrishnan, and M. Carman, "Error Detection and Corrections in Indic OCR using LSTMs," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[4] V. Vinitha and C. Jawahar, "Error Detection in Indic OCRs," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016.

[5] R. Saluja, M. Punjabi, M. Carman, G. Ramakrishnan, and P. Chaudhuri, "Sub-word Embeddings for OCR Corrections in Highly Fusional Indic Languages," in *15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[6] R. Saluja, D. Adiga, G. Ramakrishnan, P. Chaudhuri, and M. Carman, "A Framework for Document Specific Error Detection and Corrections in Indic OCR," in *1st ICDAR Workshop on Open Services and Tools for Document Analysis (ICDAR-OST)*, 2017.

II. Curriculum Vitae

Fig. 2. Student's picture.

A. Areas of Interests

Interactive Learning Systems, Indic OCR, Scene Text Recognition, Automatic License Plate Recognition, Video OCR, Machine Learning/ Deep Learning, Never Ending Learning.

B. Education

- Ph.D. (2014 - current), IITB-Monash, CGPA: 8.13.
- B.E. Hons. (E.E.E. 2006 - 2010), B.I.T.S. Pilani K.K. Birla Goa Campus, CGPA: 9.14.
- XII (2006), C.B.S.E, Percentage: 86.6.
- X (2004), C.B.S.E, Percentage: 82.

C. Professional Experience

- Engineer (2010-2014), Bharat Heavy Electricals Ltd. (BHEL), Project Engineering Management (PEM) Division, Noida, India.
  Completed Improvement Projects Rewards Scheme System (IMPRESS) project on cable routing in 2014. Developed a lisp program (AUTOCAD) to extract cable lengths interactively from drawings and a windows host script to enter them into routing software. The project saves 400 man-hours & 1500 papers in every power plant project @ BHEL-PEM.
  Skills developed: Control System Design, System Design Studies.
- Engineer Trainee (2010 January - July), LSI Logic (now Avagotech), Pune, India.
- Research Trainee (2008 May-July), Central Electronics Engineering Research Institute (CEERI), Pilani, Rajasthan, India.

D. Publications

- "OCR On-the-Go: Robust End-to-end Systems for Reading License Plates and Street Signs", Rohit Saluja, Ayush Maheshwari, Ganesh Ramakrishnan, Parag Chaudhuri, and Mark Carman, International Conference on Document Analysis and Recognition (ICDAR) 2019, Sydney, Australia.
- "Sub-word Embeddings for OCR Corrections in highly Fusional Indic Languages", Pankaj Singh, Bhavya Patwa, Rohit Saluja, Ganesh Ramakrishnan and Parag Chaudhuri, International Conference on Document Analysis and Recognition (ICDAR) 2019, Sydney, Australia.
- "StreetOCRCorrect: An Interactive Framework for OCR Corrections in Chaotic Indian Street Videos", Rohit Saluja, Mayur Punjabi, Mark Carman, Ganesh Ramakrishnan and Parag Chaudhuri, 2nd International Workshop on Open Services and Tools for Document Analysis (ICDAR- OST) 2019, Sydney, Australia.
- "Error Detection and Corrections in Indic OCR using LSTMs", Rohit Saluja, Devaraj Adiga, Parag Chaudhuri, Ganesh Ramakrishnan and Mark Carman, International Conference on Document Analysis and Recognition (ICDAR) 2017, Kyoto, Japan.
- "A Framework for Document Specific Error Detection and Corrections in Indic OCR", Rohit Saluja, Devaraj Adiga, Ganesh Ramakrishnan, Parag Chaudhuri, and Mark Carman, 1st International Workshop on Open Services and Tools for Document Analysis (ICDAR-OST) 2017, Kyoto, Japan.
- "A Framework for Error Detection and Corrections in Sanskrit", Rohit Saluja, Devaraj Adiga, Ganesh Ramakrishnan, Parag Chaudhuri and Mark Carman, Research and Innovation Symposium in Computing (RISC) 2017 (Most Admiring Poster Presentation Award), IIT-Bombay, India.
- "Analysis of bluetooth patch antenna with different feeding techniques using simulation and optimization", Rohit Saluja, A.L. Krishna, P.K. Khanna, D. Sharma, P. Sharma and H.C. Pandey, Recent Advances in Microwave Theory and Applications, International Conference on IEEE, 2008, Jaipur, India.

E. Software Experience

- Developed OpenOCRCorrect - An Adaptive Framework for correcting mistakes in OCR output. The source code can be downloaded from https://github.com/rohitsaluja22/OpenOCRCorrect.
- Helped in developing Shobhika - A Devangar font for scholars, The source code can be downloaded from https://github.com/Sandhi-IITBombay/Shobhika.
- Programming Languages: C++, Lua, Python, Matlab.
- Engineering Softwares: tesseract-ocr, Qt, Torch, Tensorflow, Kaldi (Automatic Speech Recognition).

F. Extracurriculars

- Taekwondo - Black belt, Wushu and Judo (Won 5 gold medals and 2 best fighter trophies at the district level, 3 - silver medals at the state level and a bronze medal at the national level). I also like to write poems and play musical instruments like Harmonium and Flute.