ICDAR2017 Doctoral Consortium

(in conjunction with ICDAR 2017)

Kyoto, Japan

Date: 9:00 – 13:10, November 12 2017
Location: Large Conference Room, 3rd Floor

Chairs: Véronique Eglin and Rafael Dueire Lins

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 ICDAR 2013 (in Washington D.C., USA) and with ICDAR 2015 (in Lyon, France). The Doctoral Consortium at ICDAR 2017 gives continuity to such a tradition, creating an 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) was 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 2017 Doctoral Consortium has accepted 24 students, which have been mentoring by 20 senior active researchers 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.

PROGRAM

09:00 - 09:15	Opening - Introduction to ICDAR Doctoral Consortium 2017 (Raphael Dueire Lins & Veronique Eglin)
09:15 - 10:30	Brief presentation of the projects from tutees
10:30 - 11:00	*Coffee Break and Setting-up of Posters*
11:00 - 11:30	Talk given by **Dan Lopresti** : "How to succeed in your Ph.D. degree"
11:30 - 13:00	Poster session and discussions
13:00 - 13:10	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 2017 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 24 DC projects may be clustered in seven groups:

PRINTED TEXT RECOGNITION AND MULTILINGUAL OCR

- Rohit Saluja. Indic OCR with Font and Layout Preservation
- Thi Tuyet Hai Nguyen. Multilingual OCR correction for ancient books: Looking at multiple documents to fix multiple words

HANDWRITING RECOGNITION

Off-line handwriting analysis, identification and recognition

- Chandranath Adak. Automated Handwriting Analysis on Unconventional Documents
- Paul Maergner. Graph-based Signature Verification
- Hussein Mohammed. Computational Analysis of Writing Style in Digital Manuscripts
- Daniel Wilson-Nunn. A Path Signature Approach to Online and Offline Arabic Handwriting Recognition
- Martin Schall. Segmentation-free multi-line offline handwriting recognition using LSTM networks
- Jianshu Zhang. Deep Learning Based Approach to Handwritten Mathematical Expression Recognition

One-line / dynamic shapes and /or symbols recognition

- Momina Moetesum. Deformation Estimation and Classification of Graphomotor Impressions-An Application to Neuropsychological Assessements
- Amir Ghodrati. Grouping and Recognition of Digital Ink Diagrams
- Vivek Venugopal. Exploration of novel strategies for Online Writer Identification
- Alexander Pacha. Optical Music Recognition with Deep Learning

DOCUMENT ANALYSIS AND PHYSICAL STRUCTURE

- Made Windu Antara Kesiman. Document Image Analysis of Balinese Palm Leaf Manuscripts
- Pau Riba. Graph-based representations for Document Analysis
- Bastien Moysset. Detection and localization of text lines in heterogeneous document images with deep neural networks
- Christopher Tensmeyer. Deep Learning for Document Binarization and Segmentation

INFORMATION RETRIEVAL

- Divya Sharma. Content Based Architectural Floor Plan Retrieval
- Ahmed Sabir. Enhancing Text Spotting with Visual Context Information

LARGE SCALE DOCUMENT IMAGE PROCESSING AND USER ACCESSIBILITY

- Florian Westphal. Efficient Processing of Large Document Image Repositories
- Axel Jean-Caurant. Analysis of heterogeneous documents and user behavior to improve accessibility
- Minghui Liao. Phd Research Work of Scene Text Detection

SECURITY AND COUNTERFEIT

- Héloïse Alhéritière. Securisation of hybrid documents by content-based physical analysis
- Albert Berenguel Centeno. Document counterfeit detection through background texture printing analysis

DEEP NEURAL MACHINE LEARNING TECHNIQUES FOR DOCUMENT RECOGNITION TASKS

- Michele Alberti. Understanding Deep Neural Networks Learning Behaviour

List of mentors

Olivier Augereau	Osaka Prefecture University, Japan
Najoua Benamara	ENIS, Tunis, Tunisia
Jean-Christophe Burie	L3I, Université de La Rochelle France
Florence Cloppet	LIPADE, Université Paris Descartes, France
Bertrand Coüasnon	IRISA, INSA de Rennes, France
Mickael Coustaty	L3I, Université de La Rochelle France
Chawki Djeddi	Larbi Tebessi Uinversity, Tebessa, Algeria
Uptal Garain	CVPR Unit, Indian Statistical Institute, India
C.V. Jawahar	International Institute of Information Technology, India
Christopher Kermorvant	TEKLIA, Paris, France
Bart Lamiroy	Université de Lorraine - LORIA UMR 7503, France
Angelo Marcelli, UNISA	Université de Salerne, Italy
Simone Marinai	University of Florence, Italy
Harold Mouchere	IRCYYN, Univeristé de Nantes, France
Umapada Pal	CVPR Unit Indian Statistical Institute, India
Ioannis Pratikatis	Department of Electrical and Computer Engineering, Greece
Oriol Ramos	CVC Barcelona – UAB, Spain
Marçal Rossinyol	CVC Barcelona – UAB, Spain
Ernest Valveny	Computer Vision Center – UAB, Spain
Nicole Vincent	LIPADE, Université Paris Descartes, France

Short bio of the DC Chairs

Veronique Eglin is full professor in computer science at INSA de 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 12 Ph.D theses in computer vision and document image analysis and recognition, ten papers in international journals, five books chapters, more than sixty 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).

Rafael Dueire Lins is full professor in computing at Centro de Informática at the Universidade Federal de Pernambuco (Brazil) since 2010 and at the Universidade Federal Rural de Pernambuco (Brazil) since 2016. He holds a B.Sc. degree in Electrical Engineering (Electronics) from the Federal University of Pernambuco, Brazil (1982) and a Ph.D. degree in Computing from the University of Kent at Canterbury, UK (1986). Lins published 10 books, amongst them the best-seller "Garbage Collection: Algorithms for Dynamic Memory Management", (John Wiley & Sons, UK, 1996) translated into Chinese (Mandarin) and published by ChinaPub in 2004. His pioneering contributions encompass the creation of the Lambda-Calculus with explicit substitutions, the first general and efficient solution to cyclic reference counting in sequential, parallel and distributed architectures. Lins was one of the pioneer researchers in document engineering and digital libraries in Latin America. In this area, he was the first to address the problem of back-to-front interference (bleeding) in documents. Lins supervised 51 M.Sc dissertations and 15 Ph.D. theses in computer science and electrical engineering. Lins published 44 papers in refereed journals and over two hundred articles in international conferences. Lins was the vice-chair of AIPR TC-10 (Graphics Recognition) from 2011 to 2015 and chair from 2015 to 2017.

Automated Handwriting Analysis on Unconventional Documents

Student's name: Chandranath Adak[1]
Supervisors of the thesis: Prof. Michael Blumenstein[2] & Prof. Bidyut B. Chaudhuri[3]
[1] Griffith University, Australia & [2] University of Technology Sydney & [3] Indian Statistical Institute
Ph.D. Thesis will be submitted at: School of ICT, Griffith University, Australia-4222.
Starting date of the Ph.D.: 27th April, 2015
Expected finalization date of the Ph.D.: April, 2019
adak32@gmail.com

***Abstract*—This paper briefly explains my Ph.D. research direction towards handwriting analysis, more specifically writer identification. My research aims to deal with such handwritten documents those are not quite popular in writer identification literature, but fairly realistic with respect to the practical forensics-need. The initial outcomes were worth publication and published in the proceedings of two international conferences.**

I. Introduction

"Handwriting" is basically a kind of pattern. However, from the pre-historic era, it bears the connotation of human civilization. Although the world is going fast towards paperless e-world, "handwriting remains just as vital to the enduring saga of civilization (− Michael R. Sull)".

For computer scientists, *automated analysis of handwriting* is a recognized field-of-study owing to the ever-increasing complexity of extreme variation and having the positive impacts on the fields of Forensics, Biometrics, Library and Data Science.

From the enormous field of handwriting analysis, our focus is on writer identification [4, 5] from unconventional documents. We use the term "unconventional" to refer the documents those are not used quite often. In the literature, we see the used documents are mostly free from handwriting error. However, it is quite plausible that a handwritten document may contain struck-out/ crossed-out errors [7]. In [1], we have analyzed the influence of such struck-out texts on writer identification.

Sometimes, in a multilingual country, people can read/write in multiple (mostly two) scripts. Now, a situation may arise when we have the knowledge of writer's handwriting only in one script (say, Sc_1) and we need to identify/verify the writer using test data in another script (say, Sc_2). In [2], we have tackled this problem of writer identification/ verification on English script while training has been performed by Bengali script.

In *Sec. II*, we have discussed our initial work-done with some preliminary results. *Sec. III* briefly states our future work plan.

II. Methodology with Initial Outcome

In this section, we have discussed the methodology in brief with initial outcome.

A. Effect of struck-out texts on writer identification

In [1], we have analysed the impact of struck-out text (Fig.1) on writer identification. Here, at first, the struck-out texts are detected using a hybrid classifier of a CNN (Convolutional Neural Network) and an SVM (Support Vector Machine). Then the writer identification process is activated on normal and struck-out text separately, to ascertain the impact of struck-out texts. For writer identification, we use two methods: (a) a hand-crafted feature-based SVM classifier, and (b) CNN-extracted auto-derived features with a recurrent neural model.

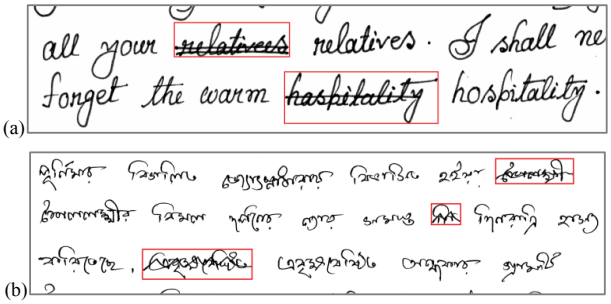

Fig.1. Examples of (a) English and (b) Bengali handwritings containing some struck-out texts marked in red boxes.

Initial results: Experimenting on our English (Bengali) handwritten database [1], we observed that the presence of struck-out texts degraded the writer identification F-Measure by 5.43% (4.92%) and 3.08% (3.23%) while employing the hand-crafted and auto-derived features, respectively.

We show our writer identification performance in a bar-

chart of Fig. 2.

B. Writer identification on English while training is on Bengali:

Our intention to perform such experiment is to study whether there exists any implicit personal characteristic in handwriting styles independent of scripts. In [2], 23 types of structural and statistical features are extracted, and multiple classifiers are employed for such identification.

Mainly KNN, MLP (2 variations: MLP1 and MLP2) and SVM are used as classifiers. On each of these classifiers, 3 training sessions are performed to generate 3 classifiers (*C1*, *C2*, *C3*). Finally, sum rule is applied to combine classifiers [3].

Initial results: We performed our experiment when Bengali handwriting was available for training, and testing was done on English handwriting only. Overall, the best outcome was 71.19% Top-1 F-measure. In Fig.3, we have shown different outcomes of combining classifiers in a bar-chart.

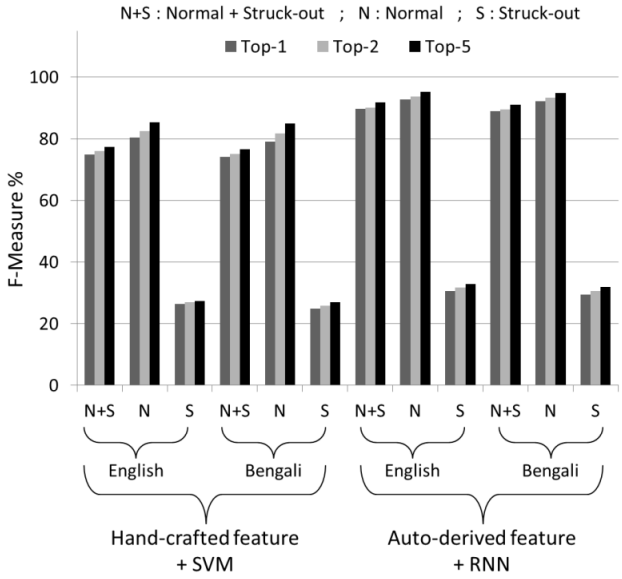

Fig. 2. Bar chart of writer identification performance on handwritten pages containing struck-out texts.

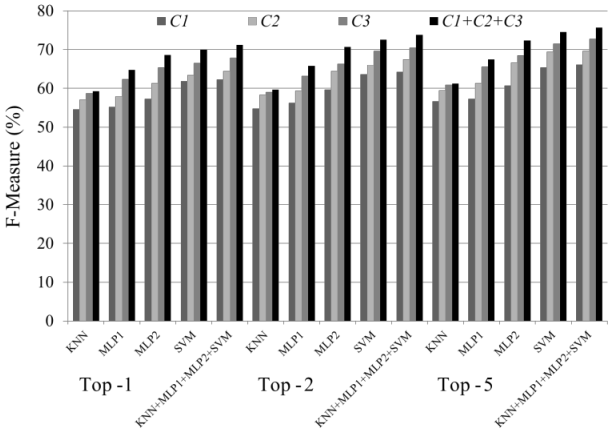

Fig. 3. Bar chart: Top-1, Top-2 and Top-5 writer identification performance while training (Bengali) and testing (English) on different scripts.

III. Future Plan

In future, we will extend our current work of [1] and [2]. We will also focus on challenges of writer identification task while there exist high intra-variability of an individual's multiple handwriting specimens. Some details of our research idea to be pursued can be found in the preprint version of [6].

References

[1] C. Adak, B. B. Chaudhuri, M. Blumenstein, "Impact of Struck-out Text on Writer Identification", Proc. 30th Int. Joint Conference on Neural Networks (IJCNN), pp. 1465-1471, Anchorage, Alaska, USA, 14-19 May, 2017

[2] C. Adak, B. B. Chaudhuri, M. Blumenstein, "Writer Identification by Training on One Script but Testing on Another", Proc. 23rd Int. Conference on Pattern Recognition (ICPR), pp. 1148-1153, Cancun, Mexico, 4-8 Dec., 2016.

[3] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On Combining Classifiers", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.20, no.3, pp.226-239, 1998.

[4] R. Plamondon, G. Lorette, "Automatic Signature Verification and Writer Identification - The State of the Art", Pattern Recognition, vol.22, no.2, pp.107-131, 1989.

[5] L. Schomaker, "Advances in Writer Identification and Verification", Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), vol.2, pp.1268-1273, 2007.

[6] C. Adak, B.B. Chaudhuri, M. Blumenstein, "Writer Identification and Verification from Intra-variable Individual Handwriting", arXiv preprint arXiv:1708.03361.

[7] B.B. Chaudhuri, C. Adak, "An Approach for Detecting and Cleaning of Struck-out Handwritten Text", Pattern Recognition, vol. 61, pp. 282-294, January 2017.

Graph-based Signature Verification

Student's name: Paul Maergner
Supervisor of the thesis: Andreas Fischer
Co-Supervisors of the thesis: Kaspar Riesen and Rolf Ingold
University: University of Fribourg, Switzerland
Starting date of the PhD: August 1, 2016
Expected finalization date of the PhD: July 31, 2019
Email: paul.maergner@unifr.ch

***Abstract*—The task of signature verification is challenging since it usually has to rely on only a few genuine signatures. Automatic methods for signature verification have improved significantly over the past decade, but they are still far from human performance. Most of these methods rely on feature vector representations and statistical classification. These methods have known limitations since they are lacking a convenient way to capture the global structure of the signatures as well as the relations between their subparts. A graph-based representation combined with a structural classification framework could overcome these limitations and we believe such an approach could be beneficial for signature verification. The probable reason for the lack of graph-based pattern recognition is the high computational complexity involved in structural matching. However, in recent years promising approximation frameworks for graph edit distance have been introduced that make graph matching more applicable to a wider range of pattern recognition problems. In my Ph.D. thesis, I want to explore the use of graphs and the potential benefits of structural pattern recognition methods for signature verification. As a first step, we have implemented a graph representation and graph matching framework for signature verification that has been used for other handwriting recognition tasks. The approach showed promising results compared to state-of-the-art. We want to continue our work by comparing different and novel graph representations as well as different graph matching frameworks.**

I. Short research plan

A. Introduction

Handwritten signatures are widely used for personal authentication. Hence, there has always been a demand to verify the authenticity of signatures. Today signature authentication and verification is done by humans as well as machines. However, while signature verification turns out to be a difficult task for humans, the latter is still quite rare as only a few and very specialized automatic systems exist which can be used by forensic experts.

The pattern recognition community defines automatic signature verification as a two-class pattern classification problem [1]. An automated system has to decide whether or not a given signature belongs to a referenced authentic author. If a system can find enough evidence of genuine authorship from the questioned signature, it considers the signature as genuine; otherwise, it declares the signature as forged (or as a signature that belongs to another user).

In the pattern recognition community, the development of novel automatic signature verification systems is still a very active field of research. The community defines two different kinds of signature verification: *Online* signature verification and *offline* signature verification. In the online scenario a user typically signs on an electronic device and thus several dynamic characteristics, such as the writing speed, the pressure, or time information, are available for recognition. Offline systems are based on the 2D image of the signature only and thus no dynamic features are available. Offline signature verification applies to more use cases, but it is also considered as the more difficult task due to the lack of information. The present research project is focused on the offline scenario.

State-of-the-art offline handwriting recognition systems either rely on global information, e.g., using logistic regression to obtain a large number of geometrical features like number of holes, moments, projections, distributions, position of barycenter, number of branches in the skeleton, Fourier descriptors, tortuosities, directions, curvatures and chain codes, and many others [1], [2], or they take local information into account, e.g., Gaussian grid features taken from signature contours or histogram of oriented gradients (HOG) or local binary patterns (LBP) [3].

Most of the state-of-the-art systems for automatic handwriting recognition as well as most of the research activities are based on feature vector representations. Therefore, statistical pattern recognition methods are usually employed in signature verification. Standard statistical classifiers include support vector machines (SVM), Hidden Markov Models (HMM), and dynamic time warping (DTW) [4].

On the other hand, only very few works take graphs into account for representing the underlying handwriting. Consequently, a lack of research on structural pattern recognition can be observed in the field of signature verification. This is likely due to the high computational complexity of structural matching like graph edit distance. However, recently approximations of the graph edit distance have been introduced that make graph matching applicable to a wider range of pattern recognition tasks [5].

The basic idea of our research project is to employ graphs and in particular graph matching and graph embedding techniques for the task of signature verification. Clearly, the overall question to be answered is, whether or not graph-based representation and structural pattern recognition methods can be beneficially employed for this task. We believe that the high representational power of graphs could be beneficial for signature verification.

From the current point of view we identify the following

three major lines of research to be pursued in the present research project:

1) Creating graph representation from signature images
2) Adapting and developing graph matching techniques for signature verification
3) Combining structural classifier ensembles for signature verification

B. Completed Work

The first goal has been to create an initial offline signature verification system that uses a graph representation for signature images and a graph matching framework. We have implemented our first graph matching framework for signature verification based on the bipartite approximation of the graph edit distance. The first graph representation is based on keypoint graphs, which have been successfully used for historic handwriting recognition [6] and keyword spotting [7]. Keypoint graphs use various keypoints on the skeleton image of the signature as nodes and connect them with edges if they are connected on the skeleton. The initial signature verification system was then evaluated on the publicly available MCYT-75 dataset. The preliminary results have been promising and we presented our findings at the International Graphonomics Society Conference (IGS) in June 2017 (see [8]).

In the next step, we have experimented with several ways of creating, normalizing, and comparing signature graphs built from keypoints in the skeleton images. The performance was evaluated on three publicly available benchmark datasets (GPDSsynthetic, MCYT-75, and SigComp2011). Our parameters have been optimized and analyzed on the GPDSsynthetic dataset and then a system using the best parameters has been applied on the MCYT-75 and SigComp2011 dataset. We have demonstrated the effect of the different parameters and the importance of a good normalization. The results on the different datasets look promising. This latest work will be presented at ICDAR 2017 in Japan (see [9]).

C. Future Work

There are several ways to further explore structural approaches to signature verification. We want to research the three parts of our signature verification system: (1) graph representation, (2) matching framework, and (3) combining classifiers.

Regarding the graph representation, we want to research novel graph-based representations of signatures images. We want to extent the skeleton-based keypoint graphs with additional keypoints. Additionally, we want to try representations that focus on stroke primitives instead of keypoints. Lastly, we are intrigued by exploring the usefulness of matching graphs as a graph representation.

We want to look into improvements to the graph matching framework, both in terms of approximation accuracy and computational complexity. For example, there is an approximation of the graph edit distance based on Hausdorff matching that runs in quadratic time [10]. This approximation could allow us to work efficiently with even larger graphs. Additionally, we want to look into the integration of a signature stability measure into the matching process.

In term of combining classifiers, we think that the combination of both structural and statistical verification systems could benefit from their complementary perspectives on signature images. It is likely that a system that combines these perspectives could increase its robustness of signature verification and therefore improve biometric authentication.

References

[1] D. Impedovo and G. Pirlo, "Automatic signature verification: The state of the art," *IEEE Trans. on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 5, pp. 609–635, 2008.

[2] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification - the state of the art," *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989.

[3] M. B. Yilmaz, B. Yanikoglu, C. Tirkaz, and A. Kholmatov, "Offline signature verification using classifier combination of HOG and LBP features," in *Proc. Int. Joint Conference on Biometrics*, 2011, pp. 1–7.

[4] M. Ferrer, J. Alonso, and C. Travieso, "Offline geometric parameters for automatic signature verification using fixed-point arithmetic," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 993–997, 2005.

[5] K. Riesen and H. Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," *Image and Vision Computing*, vol. 27, no. 7, pp. 950–959, 6 2009.

[6] A. Fischer, K. Riesen, and H. Bunke, "Graph similarity features for HMM-based handwriting recognition in historical documents," in *Proc. 12th Int. Conf. on Frontiers in Handwriting Recognition*, 2010, pp. 253–258.

[7] M. Stauffer, A. Fischer, and K. Riesen, "Graph-based keyword spotting in historical handwritten documents," in *Proc. Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 2016, pp. 564–573.

[8] P. Maergner, K. Riesen, R. Ingold, and A. Fischer, "Offline signature verification based on bipartite approximation of graph edit distance," in *Proc. of International Graphonomics Society Conference (IGS)*, June 2017.

[9] ——, "A structural approach to offline signature verification using graph edit distance," in *Proc. of 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, November 2017 (to appear).

[10] A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, and H. Bunke, "Approximation of graph edit distance based on Hausdorff matching," *Pattern Recognition*, vol. 48, no. 2, pp. 331–343, 2 2015.

Deformation Estimation and Classification of Graphomotor Impressions-An Application to Neuropsychological Assessements

Student Name: Momina Moetesum
Supervisor of the thesis: Imran Siddiqi
Starting date of the PhD: Spring 2016
Expected finalization date of the PhD: Fall 2019
University: Bahria University, Islamabad, Pakistan
Email:reach.momina@gmail.com

I. Overview of Research

Neuropsychological assessments are designed to assess various cognitive and perceptual abilities of an individual [1]. These assessments, in correlation with other clinical findings, are used for the purposes of early detection and diagnosis of various neurological and neuropsychological disorders (such as dyspraxia, visuo-spatial neglect and Parkinson). They can also be used to monitor progress of prescribed treatment and rehabilitation. Most of these tests include psychometric analysis of graphomotor impressions (e.g. handwriting and drawings etc.) as they are a direct product of complex cognitive, perceptual and motor skills [2]. Employing these easily administered and non-intrusive, 'pencil-and-page' based tasks, the subjects are required to reproduce a stimulus either by copying it or by memory. The measurement of deformations helps the practitioners to determine various aspects of the neuropsychological state of the subject. Some popular tests include Bender Gestalt Visual Motor Test [3], Clock Draw Test [4] and Rey Osterrieth Complex Figure Test [5]. Conventional assessment of these tests involves manual identification and scoring of errors in the responses, which is a time consuming activity and subject to high inter-scorer variability. Studies [6], [7], [8], suggest that development of a computerized framework for automated analysis and result reporting for a range of drawing and writing based tests, will not only allow standardization but will also facilitate the practitioners to focus more on diagnosis and future test development.

The key objective of our research is to design and implement a conceptual model of a clinical decision support system based on image processing and pattern recognition techniques which can assist clinicians in analysis of hand drawn figures produced by subjects. The intended system can provide useful assistance to clinical practitioners in behavior profiling, early detection of certain disorders and to check the validity and effectiveness of these common tests and practices. The specific objectives necessary to achieve the research aims are identified as:

- Designing an effective technique for localization and recognition of the component(s) of interest from a multi-object image.
- Development of techniques to model the expected prototypes.
- Determining techniques for estimation and classification of deformations created by the subjects.
- Development of a framework to score deviations in samples from expected model.

As a case study, we are primarily considering Bender Gestalt Visual Motor Test (BGT), a popular neuropsychological screening test used to evaluate the visual-motor maturity and perceptual distortions associated with various neurological disorders [9], [10]. The test comprises of showing a set of drawings (Figure 1-a) to the subjects, which are required to reproduce them on a sheet of blank paper (Figure 1-b) . The drawings progress from simple patterns to complex ones to capture all dimensions of mental maturation. The produced drawings are then analyzed according to the standardized scoring guidelines, Lack's scoring manual [11] being the most popular of these. Lack suggested 12 discriminators of brain damage based on the Gestalt psychology, which are determined by some or all of the nine figures of the BGT test as shown in Figure 2.

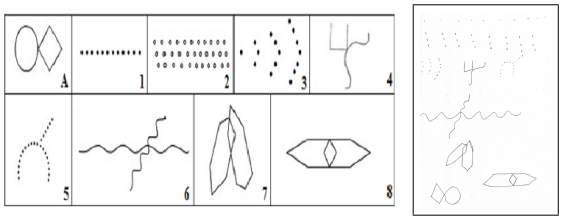

Fig. 1. (a) BGT Drawing Cards (b) Sample drawn by the subject

A. Challenges

Prime challenges that need to be addressed during the research are outlined in this section. To the best of our knowledge, presently there is no digitized dataset available for the complete samples of this test. Therefore, in addition to the primary objectives mentioned above, our proposed research will also contribute in the compilation of a digitized dataset of offline images. From system design point of view, the first challenge is the localization and segmentation of individual BGT figures from multi-figure samples drawn by the subjects.

Errors/ Discriminators to be Analyzed	Brief Description	BGT Figures							
		A	1	2	3	4	5	6	7	8
Rotation	Whole drawing rotated 80 to 180 degrees	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cohesion	Changes in size within a figure by 1/3 **OR** changes in size between figures by 1/3	✓	✓	✓	✓	✓	✓	✓	✓	✓
Simplification	Circles for dots on figure 1 **OR** don't overlap (more than 1/8 inch) oversimplified (not curves for angles)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Motor Incoordination	Irregular, tremored lines, very heavy pressure	✓	✓	✓	✓	✓	✓	✓	✓	✓
Collision	One figure overlaps another or comes with 1.4 inch	✓	✓	✓	✓	✓	✓	✓	✓	✓
Impotence	Tries repeatedly but can't successfully reproduce	✓	✓	✓	✓	✓	✓	✓	✓	✓
Closure Difficulty	Slight separation or overlap (less than 1/8 inch) **OR** significant overlap, contact in wrong place, distorted or overworked at contact point	✓				✓			✓	✓
Retrogression	Loops for circles **OR** dashes for dots, must be extreme and persistent **OR** triangle, square, **OR** rectangles substituted for a hexagon	✓	✓	✓		✓		✓	✓	✓
Overlapping Difficulty	Contact point is omitted, simplified, reworked, or distorted **OR** overlap in wrong place **OR** don't overlap (less than 1/8 inch)							✓	✓	
Perseveration	Dots from Fig 1 onto 2 **OR** circles from Fig 2 onto 3 or 5 **OR** more than 13 dots on Fig 1 **OR** more than 12 columns of dots on Fig 2 **OR** extra row of circles or dots		✓	✓	✓		✓			
Angulation	Can't reproduce the angulation **OR** whole figure turned 45 to 80 degrees **OR** variable angulation of half of columns on Fig 2		✓	✓						
Fragmentation	Figure is broken apart with destruction of the gestalt **OR** 6 or fewer items on figs 1 and 2, or missing row	✓	✓	✓	✓	✓	✓	✓	✓	✓

Fig. 2. Lacks's Discriminating Errors for Brain Damage Determined by BGT

Problems arise from variations in placement of figures in each document e.g. figures being too close or overlapping at times as well as the parts of the same figure being drawn far apart etc. Similarly classification of segmented BGT figures itself is characterized by challenges of a typical sketch recognition system where variations introduced by the subject deviate the figure from the expected prototype (Figure 3). Finding similarity between an intended pattern and distorted version itself is a challenging task.

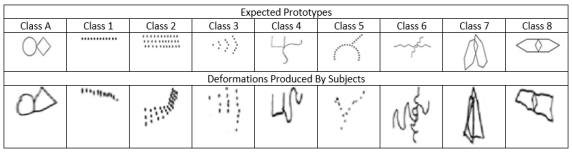

Fig. 3. Examples of Deformations Produced By Subjects

The most important yet most challenging part of our research is the measurement of the dissimilarity of figures drawn by the subjects and their intended prototypes. Contrary to a human expert, measurement of unconstrained deformations of the expected prototype is extremely difficult task for a machine. There are several conditions that determine the presence/absence of an error in a particular figure, making it hard to hand craft features for each variation. For instance, Rotation error, which is measured across all figures, is scored only if the whole figure (not its parts) is moved along its axis from 80 degrees to 180 degrees. In some cases (e.g. BGT figure 1 and 2), if the figure is moved between 45 degrees and 80 degrees, then Angulation is scored instead of Rotation. Furthermore, characteristics of same error are measured differently across different BGT figures. For instance Retrogression for BGT figure 1 is substitution of loops for circles and dashes for dots in BGT figure 2 as shown in Figure 4-a and Figure 4-b respectively. On the other hand it is substitution of triangle or square for diamond in BGT figure A and rectangle for hexagon in BGT figure 8 as shown in Figure 4-c and Figure 4-d respectively. An added complexity of the problem is that some errors are scored on page level like Cohesion and Collision, while multiple errors can exist in same figure. All these instances make deformation estimation and classification one the most challenging part of this research.

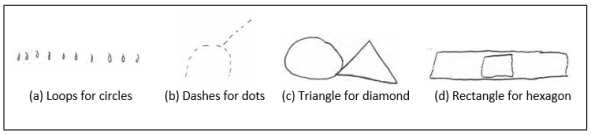

Fig. 4. Examples of Retrogression Scoring in four BGT Figures

B. *Proposed Methodology*

Keeping Lack's scoring for BGT as our prime target study, we aim to design a generic approach to system implementation for different conditions of interest like other figure based tests and handwriting as well. The conceptual model of the proposed system is presented in Figure 5.

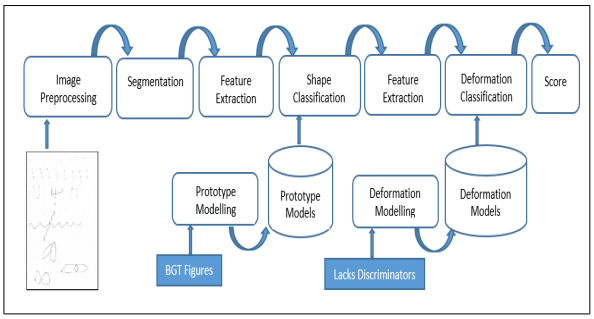

Fig. 5. Conceptual Model of Proposed System

During the training phase, modeling of the expected prototype and possible deformations will be done. These models will later be used in the testing or analysis phase to classify both the figures and the deformations. Literature suggests various structural representation techniques like polygonization [12] and graph theory [13] etc. to represent shapes [14], [15] as well as handwriting [16], [17]. On the other hand, statistical approaches (like Hidden Markov models [18]) and deep learning approaches (Convolutional Neural Networks [19]) have also been employed to train models. We will be exploring these techniques and select the most suitable one for modeling our ground truth patterns. During the second phase, offline samples drawn by subjects are presented to the system. In order to analyze each gestalt separately, localization, segmentation and classification of each BGT response will be done. Once classified, the system will estimate the type of deformation present in the resultant impression. These deformations are considered errors and their presence usually determines the diagnosis. Type of deformation will be determined by the comparison of created impression with the model of the expected prototype.

II. Current Progress and Preliminary Findings

Samples from almost 200 subjects varying from 12 to 18 years of age (mainly targeting children and adolescents) have been collected till now. All the samples were acquired under expert supervision of trained psychologists and were later scored. Data collection will be an ongoing process through out our research.

We have already conducted some pilot studies to establish theoretical grounds for the proposed methodology. In a pilot study (Moetesum et al. 2016), we used laws of Gestalt theory for segmentation purposes. We grouped the nine figures of BGT test into three groups based on their closure, continuity and similarity and realized promising segmentation performances. In the same study, we employed a well-known structural technique, shape context features [20] to match each segmented shape with the prototype shapes in the reference base and classify the query shape into one of the nine Gestalt classes. The effectiveness of the classification scheme was evaluated by using one image of each of the 9 classes as reference set. Out of the 153 drawings presented to the system, 129 were correctly recognized reading an overall classification rate of 84.31%. In our most recent work accepted at ICDAR 2017, we have employed different combinations of pre-trained CNNs as feature extractor and different classifiers (e.g. SVM & LDA) to distinguish between the nine drawings and achieved classification rates of approximately 93.5%.

As for deformation estimation and classification, in one of our initial pilot studies (Moetesum et al. 2015), we attempted hand crafted features to represent a subset of scoring properties. A heuristic approach consisting of various morphological operations was applied to detect the presence of different errors. Although the results were promising but the procedure is too tailored for specific problem with less room for generality.

A. Relevant Publications

- **2017**: Nazar, H., **M. Moetesum**, Ehsan, S., Siddiqi, I., Khurshid, K., Vincent, N. and McDonald-Maier, K.D. Classification of graphomotor impressions using convolutional neural networks an application to automated neuropsychological screening tests, In Proc. of 14th International Conference on Document Analysis and Recognition (ICDAR), 2017 (To Appear).
- **2016**: **Moetesum, M.**, Siddiqi, I., Masroor, U., Vincent, N., and Cloppet, F. Segmentation and Classification of Offline Hand Drawn Images for the BGT Neuropsychological Screening Test. In Proc. of 8th International Conference on Digital Image Processing (ICDIP), Chengdu, China, May. 2016.
- **2015**: **Moetesum, M.**, Siddiqi, I., Masroor, U. and Djeddi, C. "Automated Scoring of Bender Gestalt Test Using Image Analysis Techniques. In Proc. of 13th IAPR International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, Aug. 2015.

III. Future Plan

The work carried out so far segments and classifies the drawings into one of the 9 shape classes but does not currently measure the deformations. Currently, we are trying to model the deformations using deep learning techniques. This will allow automated identification and scoring of errors. This may require designing and training of a CNN architecture from scratch. The major challenge in developing such a system is to have training data with instances of all possible errors that the subjects can make. This may require generating synthetic data or applying data augmentation techniques to allow sufficient training data. In addition to deep learning based solutions, modeling of drawings using graph theory is also planned to be investigated.

References

[1] A. M. Poreh, *The quantified process approach to neuropsychological assessment*. Psychology Press, 2012.

[2] J. Ziviani and M. Wallen, *The development of graphomotor skills*. Mosby Elsevier, 2006.

[3] R. Walrath, "Bender visual motor gestalt test," in *Encyclopedia of Child behavior and Development*. Springer, 2011, pp. 233–234.

[4] B. J. Mainland and K. I. Shulman, "Clock drawing test," in *Cognitive Screening Instruments*. Springer, 2013, pp. 79–109.

[5] M.-S. Shin, S.-Y. Park, S.-R. Park, S.-H. Seol, and J. S. Kwon, "Clinical and empirical applications of the rey–osterrieth complex figure test," *Nature protocols*, vol. 1, no. 2, pp. 892–899, 2006.

[6] C. Rémi, C. Frélicot, and P. Courtellemont, "Automatic analysis of the structuring of children's drawings and writing," *Pattern Recognition*, vol. 35, no. 5, pp. 1059–1069, 2002.

[7] M. C. Fairhurst, T. Linnell, S. Glenat, R. Guest, L. Heutte, and T. Paquet, "Developing a generic approach to online automated analysis of writing and drawing tests in clinical patient profiling," *Behavior Research Methods*, vol. 40, no. 1, pp. 290–303, 2008.

[8] N. Renau-Ferrer and C. Remi, "A method for visuo-spatial classification of freehand shapes freely sketched," *arXiv preprint arXiv:1305.1520*, 2013.

[9] A. A. A. d. Santos and L. M. d. Jorge, "Bender test with dyslexics: comparison of two systems of punctuation," *Psico-USF*, vol. 12, no. 1, pp. 13–21, 2007.

[10] R. B. Ferreira, C. F. Feil, and M. L. T. Nunes, "Bender visual-motor gestalt test in the children's clinical assessment," *Psico-USF*, vol. 14, no. 2, pp. 185–192, 2009.

[11] P. Lacks, *Bender Gestalt screening for brain dysfunction*. John Wiley & Sons Inc, 1999.

[12] B. R. De Araújo, D. S. Lopes, P. Jepp, J. A. Jorge, and B. Wyvill, "A survey on implicit surface polygonization," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 60, 2015.

[13] L. Lin, X. Wang, W. Yang, and J.-H. Lai, "Discriminatively trained and-or graph models for object shape detection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 959–972, 2015.

[14] L. Prasad, A. N. Skourikhine, and B. R. Schlei, "Feature-based syntactic and metric shape recognition," in *International Symposium on Optical Science and Technology*, 2000, pp. 234–242.

[15] W. Lee, L. B. Kara, and T. F. Stahovich, "An efficient graph-based recognizer for hand-drawn symbols," *Computers & Graphics*, vol. 31, no. 4, pp. 554–567, 2007.

[16] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," *Pattern Recognition*, vol. 43, no. 11, pp. 3853–3865, 2010.

[17] P. Wang, V. Eglin, C. Garcia, C. Largeron, J. Lladós, and A. Fornés, "A novel learning-free word spotting approach based on graph representation," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, 2014, pp. 207–211.

[18] A.-L. Bianne-Bernard, F. Menasri, R. A.-H. Mohamad, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem, "Dynamic and contextual information in hmm modeling for handwritten word recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 2066–2080, 2011.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

Securisation of hybrid documents by physical layout extraction

Student's name: Héloïse ALHERITIERE [1,2]
Supervisor/s of the thesis: Camille KURTZ [1], Florence CLOPPET [1], Nicole VINCENT [1] and Jean-Marc OGIER [2]
University: [1]University Paris Descartes, LIPADE (EA 2517), SIP team
[2]University La Rochelle, L3i (EA 2118)
Starting date of the PhD: October 2015
Expected finalization date of the PhD: September 2018
Email: heloise.alheritiere@parisdescartes.fr

***Abstract*—The objective of my thesis is to define a methodology to secure a document, either in a paper or digital format (what we call hybrid document), by analyzing and hashing its content. We consider that the text and the image represent the information in various ways. Thus, the first essential step of this project is to extract the layout of the document. The difficulty of this step is directly related on the nature of hybrid documents. This implies that the method needs to be stable regarding alterations made to the document.**

I. Short research plan

My thesis is funded by a project of the French National Research Agency (ANR) "Semantic Hash for Advanced Document Electronic Signature" (SHADES) which seeks to secure paper and digital documents.

A document is not defined by its support but by its content. Thus if a document is created in Paris and thereafter printed in Kyoto, we wont have the same support though the document is the same, only its support can change. However if the support has been modified (physically or numerically) we will have a different document. Thus we define the concept of hybrid document.

The first step of pre-processing is to dematerialize the document if it is a paper document. We work on images of documents, a document consisting of a single page. On this document image we will associate a digest (keycode) that will be calculated from the elements that compose it. This stamp will allow us to identify the document thus allowing securization. In this way, the stamp can be used as a basis for comparing documents, even if the document is lost or destroyed as long as the stamp has been stored in a protected place, the document entity can always be compared with its copies.

So regardless of the time t where the digest has been calculated, if the document has not been legally changed, any generation of digest at time $t + x$ will have to provide an identical digest. The thesis objective is to describe the layout of the document in a stable way so that each part can be extracted and hashed in a specific manner in order to optimize the results.

A. Methodology

In a first report called "Bibliographic report related to document layout extraction" [1], we studied the methods already developed in the litterature, and evaluated their stability. We also studied the robustess of different features, and this yielded to the second report called "Report on feature robustness" [2]. The conclusion of the latter was that few features can be considered as stable during the life of the document even though led to good results. These preliminary studies have led us to look for the different layers based on different but complementary points of view. Thus we have chosen to use a global line-based method to extract the layout. This leads to a conference publication [3]. In another research direct use a layer-based extraction for some elements such as the alternate color [4].

B. Method proposed

The employed method for complex layout extraction is based on features of higher level than pixels obtained from a document straight line based segmentation. We propose to capture the straight line segments thanks to a new transform called the Local Diameter Transform (LDT), which integrates the local spatial organization of the segments contained in the document content. Such segments may be used to approximate filled forms, lines and drawings that constitute another level of document primitives from a topological point of view. Furthermore according to the length of the segments, some document parts can be discriminated. The proposed transform is applied simultaneously on the foreground (related to the document content) or the background pixels, in order to take advantage of the duality of information present in both parts of the document in oder to extract its layout. The segmented document regions are then labeled with different classes of the document layout (text areas, images, separators and tables) thanks to a decision-tree procedure involving high-level rules also derived from the straight line segments (cf.Fig. 1).

C. Future Work

Many points remain to be considered for future works as part of my thesis. We plan to use the transform presented in our paper ICDAR 2017 [3] for other use than the extraction of the layout for example using them as a shape descriptor

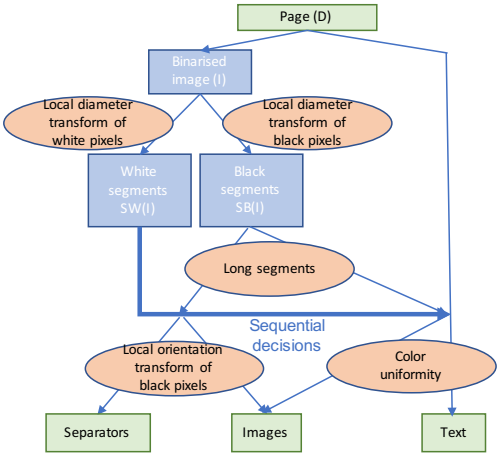

Fig. 1. Overview of the method (input / outputs are depicted in green).

for recognition. Our work is based on numerous rules that have thresholds adapted to the size of the document but fixed (which do not depend on its content). One way to improve the results would be to find a way to automate them. Our method depends on the binarization of the document. In order to get rid of binarization problems, we want to generalize our transform to gray level and perhaps later to color images.

During my PhD, I really liked teaching so after my PhD I am open to the idea of becoming an assistant Professor but im also thinking about working in a research and development department in a company to discover this area.

REFERENCES

[1] H. Alhéritière, C. Kurtz, F. Cloppet, and N. Vincent, "Bibliographic report related to document layout extraction," 2016.

[2] Lipade and L3i, "Report on feature robustness," 2016.

[3] H. Alhéritière, F. Cloppet, C. Kurtz, J. Ogier, and N. Vincent, "A document straight line based segmentation for complex layout extraction," in *Proceedings of the IAPR International Conference on Document Analysis and Recognition – ICDAR 2017*, Kyoto (Japan), November 2017, p. accepted.

[4] H. Alhéritière, C. Kurtz, F. Cloppet, and N. Vincent, "Utilisation de la couleur pour l'extraction de tableaux dans des images de documents," in *Proceedings of the Colloque International Francophone sur lcrit et le Document – CIFED 2016*, Toulouse (France), March 2016, pp. 349–364.

Understanding Deep Neural Networks Learning Behaviour

Student's name: Michele Alberti
Supervisor/s of the thesis: Prof Rolf Ingold, Dr. habil. Marcus Liwicki
University: University of Fribourg
Starting date of the PhD: April 2017
Expected finalization date of the PhD: 2021
Email: michele.alberti@unifr.ch

***Abstract*—In recent years the knowledge on Deep Neural Network (DNN) made huge steps forward, yet there is still no clear and exhaustive understanding of when and why a deep model works. This makes it difficult to discriminate reliable good practices from techniques that might work, but it is not clear under which circumstances they perform well. Too often authors are applying a particular technique just because "once it worked" or because modern Deep Learning frameworks allow to use many tools "out of the box" removing the requirement of understanding how they work. I plan to investigate the fundamentals of learning Deep Neural Networks models with a focus on two main areas.**

First, I will experiment with different types of network initialization and measure how they affect the network learning behaviour and the performance on different tasks, i. e., digit and object recognition, or layout analysis on historical documents. I believe that for achieving the goal of being able to manipulate what and how the network is learning, we need a true and precise understanding on what is happening during the learning phase and the network initialization is the first step towards it.

Afterwards, I will investigate how deep models are able to interact with one another, i.e., if they are still able to perform/learn when put into a group of agents and tasked to jointly perform a task. Over the last years we have seen a huge increase in the capabilities of deep neural networks. The models became more complex and more powerful. Therefore, for me it is just natural to ask: "what if we were to take more than one model?". This however, should not be done in a trivial way and hence I want to delve into understanding how can we combine two or more models in a meaningful way.

I. Short research plan

A. Introduction

Very DNN are now widely used in machine learning for solving tasks in various domains.

Although artificial neurons have been around for a long time [11], the depth of artificial neural networks increased significantly over the last 15 years[1][12]. This is due to two reasons: the returning of layer-wise training methods[2][14] and higher computational power available to researchers.

Because of their success in many domains, a lot of resources have been invested into research and development of DNN. However, they still suffer from two major drawbacks: The first being that despite the computational power of new processors and GPUs, the training of DNN still takes time. Especially for large networks, the training time becomes a crucial issue, not only because there are more weights to use in the computations, but also because more training epochs are required for the weights to converge. The second drawback is that initializing the weights of DNN with random values implies that different networks will converge to different local minima.

B. What I have done so far

In recent years it has been shown that neural networks with many layers could be trained successfully if their weights are initialized in a meaningful way rather than pure randomly. However, not every meaningful way to initialize a network does necessarily lead to good results. This is the case where each layer is trained to reconstruct its input as an unsupervised pre-training step, e.g in the context of Stacked Convolutional Auto-Encoder (SCAE) [9]. Our initial work [3], supported by additional experiments [4] proves that if a Stacked Convolutional Auto-Encoder (SCAE) is good in reconstructing images, it is not necessarily good in discriminating their classes. More formally, the sub-dimension representation space learned from the SCAE (while being trained for input reconstruction) is not necessarily better separable than the initial input space. Hence, we reject the thought that using auto-encoder sub-dimensional features — learned by reconstructing the input — for classification tasks, is always a good idea.

Continuing the work done in my Master Thesis, we propose to initialize a Convolutional Neural Network (CNN) layer-wise with Principal Component Analysis (PCA) instead of random initialization [1]. We outperform state-of-the-art random weight initialization methods on layout analysis at pixel level on historical documents.

However, one might argue that features obtained by maximizing the variance of the input data — which is what PCA features do — might not be the optimal ones for performing classification tasks. For this, we investigated the performances of initializing a CNN layer-wise with a goal oriented (supervised) algorithm such as Linear Discriminant Analysis (LDA) by performing layout analysis at the pixel level on historical

[1]Note that deep neural architectures where proposed already much earlier, but they have not been heavily used in practice [13].

[2]Referred to as Deep Belief Networks, and often composed of Restricted Boltzmann Machines.

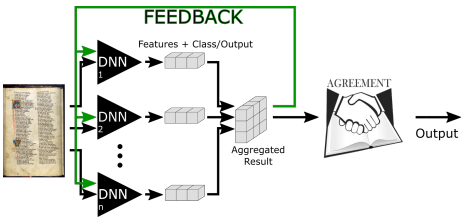

Fig. 1. Visualization of the concept (simplified) of cooperative multi agent system.

documents [2]. We show that such initialization is very stable[3], converge faster and perform better when compared with the same architecture initialized with random weights. Additionally, even before fine-tuning a network initialized with LDA exhibits noticeable results for classification task.

C. Current Status

The current focus of my research is investigating the joint use of multiple statistical tools (such as PCA and LDA) to initialize a network. Additionally an eye is kept on how different optimization algorithms (e.g plain SGD, Adam, RMSProp) behave on this type of data-driven initialization. Preliminary results are very promising, but we don't have enough data to make any statement yet, as these results are very recent and still require further investigation.

D. Future Work

We will continue to investigate the effectiveness of different initialization methods. After that, we plan to move into the context of "multi-task" learning. The idea is to show an input to a collection of agents (which can be neural networks or other algorithms), where each of them is an expert in performing a single task (e.g binarization, layout analysis, script recognition) and then use recurrent connections to make the experts "debate" their output until they agree (converge) to a stable and definitive output. The motivation behind this is that knowledge is created and maintained not only with data but with information as well. When an agent produces an output, it is inherently creating information, which we will then share among his "colleagues" to provide them with more information to update their internal knowledge. Since it is trivial that more information leads to better decisions, we argue that this cooperative scheme will allow us not only to achieve better results in each single task, but also to study the underlying in-domain relationship between the different tasks.

In Figure 1 the concept (simplified) of cooperative multi agent system is visualized. The objective is to obtain a multi-dimensional space which is composed of the input data and of the information produced by each expert. Analyzing this space and which part of it is retained by each agent is of great interest for us.

The idea of multi-task networks is not novel and it is closely related to the concept of generative networks. Here, however, we are more interested into studying the synergy between the different specialized agents rather than to create a super-model that performs all task simultaneously.

[3]It leads to highly similar patterns of weights in networks initialized on different random samples from the same dataset.

E. Tools

To conduct research in a such low level inside the networks, we developed a framework [7] [8] which we kept as simple as possible instead of being fully optimized like its commercial counterparts. Finally, to compare our results with state of art, we developed an open-source evaluation tool for the task of layout segmentation at pixel level which we plan to extend to work with polygons as well [5]. This tool was first used for the ICDAR 2017 competition on Layout Analysis for Challenging Medieval Manuscripts [6].

PUBLICATIONS

[1] M. Seuret, M. Alberti, R. Ingold, and M. Liwicki, "Pca-initialized deep neural networks applied to document image analysis," *Accepted ICDAR 2017*, vol. abs/1702.00177, 2017. [Online]. Available: http://arxiv.org/abs/1702.00177

[2] M. Alberti, M. Seuret, V. Pondenkandath, R. Ingold, and M. Liwicki, "Historical document image segmentation with lda-initialized deep neural networks," *Accepted ICDAR-HIP 2017*, 2017.

[3] M. Alberti, M. Seuret, R. Ingold, and M. Liwicki, "What you expect is NOT what you get! questioning reconstruction/classification correlation of stacked convolutional auto-encoder features," *CoRR*, vol. abs/1703.04332, 2017. [Online]. Available: http://arxiv.org/abs/1703.04332

[4] ——, "Questioning the correlation between reconstruction and classification abilities of deeply learned features," *CoRR*, 2017.

[5] M. Alberti, M. Bouillon, and R. Ingold, "Open evaluation tool for layout analysis of document images," *to appear ICDAR-OST*, 2017.

[6] F. Simistira, M. Seuret, M. Bouillon, M. Würsch, M. Alberti, M. Liwicki, and R. Ingold, "ICDAR2017 competition on Layout Analysis for Challenging Medieval Manuscripts," in *2017 International Conference on Document Analysis and Recognition*, 2017, p. to appear.

[7] M. Seuret, M. Alberti, and M. Liwicki, "N-light-n," Université de Fribourg, Tech. Rep., 2016.

[8] ——, "N-light-N : Read The Friendly Manual," Technical Report 16-02. Department of Informatics, University of Fribourg, Fribourg, Switzerland, Tech. Rep., 2016.

REFERENCES

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[11] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of Math. Biophysics*, vol. 5, pp. 115–133, 1943.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, published online 2014; based on TR arXiv:1404.7828 [cs.NE].

[14] D. H. Ballard, "Modular learning in neural networks." in *AAAI*, 1987, pp. 279–284.

A Path Signature Approach to Online and Offline Arabic Handwriting Recognition

Student's name: Daniel G. Wilson-Nunn
Supervisor/s of the thesis: Prof. Terry Lyons (Oxford), Dr Anastasia Papavasiliou (Warwick) & Dr Hao Ni (UCL)
University: The Alan Turing Institute & The University of Warwick
Email: dwilson-nunn@turing.ac.uk

***Abstract*—The Arabic script is one that has many properties that come together and result in what is commonly cited as one of the most beautiful scripts. Used by over 400 million people worldwide and with a history spanning over 1800 years, the Arabic script remains one of the most important languages in the world. Using tools from the theory of rough paths, combined with state of the art techniques from deep learning, we develop a recognition methodology for Arabic handwriting. Preliminary results using online Arabic handwritten characters show that the methodology developed can result in a significant decrease in error rate.**

I. Short research plan

A. Introduction

Handwriting recognition methods are of two distinct types; online and offline. Online handwriting recognition deals with data recorded "in time", i.e. data that is represented as a function of time. Offline handwriting recognition, on the other hand, deals with recognition of data that is in image format. Online handwriting recognition deals with the spatio-temporal resolution of the input whereas offline handwriting recognition deals with the spatio-luminance of an image [1]. Handwriting recognition has a number of important application areas, from converting handwriting to text on a tablet or touch screen to signature verification for bank fraud to digitisation of ancient manuscripts for historical study [4].

Recent advances in computing and deep learning have resulted in a large amount of interest in the area of handwriting recognition, with over 17,000 results for "handwriting recognition" appearing on Google Scholar since 2013. The popularity of the field has lead to significant developments and advancements in recent years.

The Arabic script is one with a number of challenges that add to the complexity of handwriting recognition:

1) Cursive Nature: The Arabic script comprises of 28 standard letter characters (with additions in regional variants). Arabic is always written cursive, regardless of whether it is handwritten or typed.

2) Changing letter shapes: Each of the 28 characters in the Arabic script can take a number of different shapes. This is a direct result of the cursive nature of the script and adds to the complexity of the script. Other cursive scripts such as Latin or Bengali often have a fixed letter shapes with small variations, however Arabic characters may change drastically depending on their location in the word.

3) Joining and non-joining characters: Of the 28 standard characters, 6 are "non-joining" which means that although the script is cursive, they do not join to the following letter. This will result in a small space within a word without the word actually ending.

4) Delayed strokes: Arabic is heavily reliant on delayed strokes, specifically dots, that are added to letters to distinguish them from each other. Whilst this is less of a concern in offline recognition, for online recognition, this adds significant difficult.

Thus far, my work has concentrated on online Arabic character recognition, and the intent is to continue this work to text recognition and offline recognition for Arabic too. In part *B. Methodology* of this paper, the methodology developed in the first year of my PhD will be outlined and in part *C. Future Work* ideas for future work will be presented.

B. Methodology

Online handwritten characters are sequences of (x, y) coordinates, possibly with multiple sequences per character to indicate multiple strokes in a single character.

$$X_i = \{\overbrace{\left[(x,y)_1, (x,y)_2, \ldots, (x,y)_{\ell_1}\right]_1}^{\text{Arbitrary number of coordinates per stroke}}, \ldots, \underbrace{\ldots, \left[(x,y)_1, (x,y)_2, \ldots, (x,y)_{\ell_N}\right]_N}_{\text{Arbitrary number of strokes per character}}\} \quad (1)$$

The data used in this section comes from the Online KHATT dataset with individual characters segmented. This data has been produced and worked on by al Hilali and Mahmoud [5].

In order to carry out classification on the data, it is first necessary to carry out some pre-processing. Initially work must be done to ensure that each character is of fixed length owing to the fact that most classification tools; including neural networks work best with data of fixed length. The second step is to apply an encoding so as to provide the classification tool with as good an input as possible.

1) Fixing dimensionality: Each character contains a number of strokes and each stroke contains a number of coordinates (x, y). The first step is to change the character from a list of strokes to a single list of coordinates. To do this, a third

dimension is added, and a number of "jump points" are also added. For ease of notation, let $(x,y)_i^k$ be the i-th coordinate of stroke k. The third dimension, P, is added through the following mechanism:

$$(x,y)_i^k \mapsto ((x,y)_i^k, P),$$

where

$$P = 2 \times (k-1),$$

where k is the number of the stroke that contains the coordinate. For example, the coordinate $(0.1, 0.5)_2^2$ (i.e. the second coordinate in stroke 2) would become $((0.1, 0.5)_2^2, 2)$. The data is now of the form:

$$\begin{aligned} X_i = \{&((x,y)_1^1, 0), \ldots, ((x,y)_{\ell_1}^1, 0), \\ &((x,y)_1^2, 2), \ldots, ((x,y)_{\ell_2}^2, 2), \\ &\ldots \\ &((x,y)_1^N, 2N-2)), \ldots, ((x,y)_{\ell_N}^N, 2N-2)\}. \end{aligned} \quad (2)$$

The second step to this stage is to add coordinates to indicate where the pen leaves the paper (tablet). These coordinates are added between coordinates where a new stroke begins and have the exact same x and y coordinate values as the end point of one stroke and beginning point of the next. The P value of these two coordinates that are added is equal to the odd integer that lies between the two even integers of the existing coordinates. For example, the data in (2) would become

$$\begin{aligned} & X_i = \{((x,y)_1^1, 0), \ldots, ((x,y)_{\ell_1}^1, 0), \\ \text{new} \quad \rightarrow \quad & ((x,y)_{\ell_1}^1, 1), ((x,y)_1^2, 1), \\ & \ldots \\ \text{new} \quad \rightarrow \quad & ((x,y)_{\ell_{N-1}}^{N-1}, 2N-3), ((x,y)_1^N, 2N-3), \\ & ((x,y)_1^N, 2N-2)), \ldots, ((x,y)_{\ell_N}^N, 2N-2)\}. \end{aligned} \quad (3)$$

2) Producing Inputs for Classifier: After the data has been converted to a single series of coordinates in $\mathbb{R}^3$, the data extraction method in preparation for the classifier is to be carried out. The method that will be used in this work is know as dyadic signatures., developed based on the rough path signature, invented by Terry Lyons [3].

Definition 1: Given a path $X : [S,T] \to \mathbb{R}^d$, the signature of the path over an interval $[s,t]$ is defined as

$$S_{[s,t]}(X) = \left(1, X_{[s,t]}^1, X_{[s,t]}^2, \ldots, X_{[s,t]}^n, \ldots\right),$$

where $X_{[s,t]}^n$ is defined as

$$X_{[s,t]}^n := \int_{\substack{u_1 \leq \cdots \leq u_n \\ u_1,\ldots,u_n \in [s,t]}} \mathrm{d}X_{u_1} \otimes \cdots \otimes \mathrm{d}X_{u_n}.$$

Note that it is common to only take the signature up to a certain level, e.g.

$$S_{[s,t]}^{(N)}(X) = \left(1, X_{[s,t]}^1, X_{[s,t]}^2, \ldots, X_{[s,t]}^N\right).$$

The first step to using the signature to encode details of the characters is to take a linear interpolation of the sequence of points in 3D that were produced in the first stage. Second is to consider this as a linear path in 3D and to calculate the signature of each character (up to a fixed level, say M).The resulting data is used as the input for a classifier.

In order to use *dyadic* signatures, instead of calculating the signature of the whole character, the character is split into 2^k sections of equal length and the signature of each of these sections is calculated and used as input for the classifier.

C. Results

Using a Long Short Term Memory (LSTM) network, with three hidden layers, each containing 50 nodes, the recognition is carried out. The model is trained using TensorFlow on an NVidia GTX 1080Ti graphics card and the results for different signature levels and number of dyadic intervals are obtained as follows.

Run	Signature Level	Dyadic Level	Recognition Rate (%)
1	2	3	90.44%
2	5	5	92.57%
3	2	5	91.26%
4	5	2	90.57%
5	10	2	90.56%
6	10	4	91.38%

TABLE I. RESULTS OF LSTM NETWORKS TRAINED USING VARIOUS PARAMETERS

The results show that we obtain a recognition rate of over 92.5% when using signature level 5 and 32 (2^5) dyadic intervals of the character. This rate is significantly improved on the 82% achiveed by Hilali and Mahmoud in their original paper, thus showing the capabilities of the method on this data.

D. Future Work

A selection of future work directions are being considered. The aim is to improve the field of Arabic handwriting recognition, particularly using the signature as a tool to transform the data.

- Online Arabic handwriting recognition for phrases as opposed to characters
- Offline Arabic handwriting recognition using a signature based approach
- Generating Arabic handwriting using a signature based approach

REFERENCES

[1] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, Jan. 2000.

[2] Jeremy Reizenstein, "Calculation of Iterated-Integral Signatures and Log Signatures," 2015. [Online]. Available: http://www2.warwick.ac.uk/fac/cross_fac/complexity/people/students/dtc/students2013/reizenstein/logsignatures.pdf

[3] I. Chevyrev and A. Kormilitzin, "A Primer on the Signature Method in Machine Learning," *arXiv:1603.03788 [cs, stat]*, Mar. 2016, arXiv: 1603.03788. [Online]. Available: http://arxiv.org/abs/1603.03788

[4] A. Priya, S. Mishra, S. Raj, S. Mandal, and S. Datta, "Online and offline character recognition: A survey," in *2016 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2016, pp. 0967–0970.

[5] B. M. Al-Helali and S. A. Mahmoud, "A Statistical Framework for Online Arabic Character Recognition," *Cybernetics and Systems*, vol. 47, no. 6, pp. 478–498, Aug. 2016. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01969722.2016.1206768

Efficient Document Image Binarization using Machine Learning

Student's name: Florian Westphal
Supervisor/s of the thesis: Håkan Grahn, Niklas Lavesson
University: Blekinge Institute of Technology, Sweden
Starting date of the PhD: 14.01.2015
Expected finalization date 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, word spotting or transcription, and achieve reasonable performance, is a difficult task. On top of that, there are additional challenges to make these algorithms execute fast enough to be able to process vast collections of images in a reasonable amount of time, and to make them able to deal with the absence of labelled training data specific to the target image collection.**

In my research, I am focusing on improving the execution performance of document image analysis algorithms either by implementing them so that they can be executed on CPU and GPU simultaneously or by identifying alternative algorithms, which display a comparable performance on the respective task, but have a lower execution time. Additionally, I am working on interactive machine learning to enable human users to improve the performance of a prebuilt model for document analysis on a specific target image collection through feedback. This reduces the need for labelled training data and is linked to the overall theme of my thesis in that it requires high execution performance to allow the user to interactively improve the model.

I. Short research plan

A. Introduction

It has never been easier to access historical documents than now, since different companies and government institutions provide access to high resolution colour 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 or image transcription 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 on those tasks.To be truly useful to companies and government institutions which possess 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 focusing on these two challenges, i.e., improved execution performance and dealing with missing training data. The approach I am taking to address the latter challenge is interactive machine learning to make it possible for users to improve the performance of a learned model on a specific document image analysis task by allowing them to provide feedback to the learning algorithm. This approach is connected to the first challenge since interactive machine learning also requires low execution times to make it possible for users to interactively improve the model's performance.

B. Methodology

In a first study, I have focused on improving the execution performance of Howe's binarization algorithm (HBA) [1], whose variations have achieved one of the best binarization performances in the latest Handwritten Document Image Binarization Contests (H-DIBCO).The main approach taken to improve HBA's execution performance was to break up the algorithm into different parts and to map those parts to a heterogeneous platform, i.e., a compute platform combining different processor types. In this case, HBA was split into three parts and implemented, so that each part could be executed on the CPU or the GPU to evaluate which mapping from parts to CPU and GPU would yield the best execution performance. I have presented the general idea of this approach and initial results as work-in-progress report at the Family History Technology Workshop 2016 [2]. The final evaluation of this approach has shown that mapping all parts of HBA to the GPU results on average in a 3.5 times faster execution compared to mapping all parts to the CPU on the available DIBCO and H-DIBCO datasets without affecting the binarization performance. Other variations of mapping HBA parts to CPU and GPU resulted in varying execution performances ranging between the all CPU and all GPU performance.

Another approach taken to improve the execution performance of HBA was to find an alternative algorithm for Howe's parameter tuning algorithm for HBA [3]. The proposed

approach uses random forests based multivariate regression to predict HBA's parameters using a set of 4 image features. This has improved the execution performance of the parameter tuning step on average by a factor of 2.5 on larger images from the available DIBCO and H-DIBCO datasets, while producing binarization results comparable to Howe's tuning algorithm.

The two aforementioned approaches to improve HBA's execution performance and their evaluation results have been summarised in one manuscript, which is currently being revised after a requested minor revision from the International Journal for Document Analysis and Recognition (IJDAR) [4].

A previous study on the use of interactive machine learning for image binarization has been put on hold, due to rather discouraging results. The main idea of this study was to enable users to improve an initial machine learning model trained for image binarization by providing feedback on the model's binarization result by selecting and labelling misclassified pixels either as foreground or background. The goal was then to use those few labelled pixels to update the model, resulting in an improvement in binarization performance on the image the user provided feedback on, as well as on similar images.

Different attempts have been made to achieve this goal. One of them was to increase the number of labelled training examples by assuming that unlabelled pixels in close proximity to user labelled pixels were correctly classified and can be used as labelled training data, additionally to the user provided labels. Another approach was to use ideas from few-shot learning using siamese networks to improve the potential impact user feedback can have on the binarization performance. However, all attempts tested so far have shown that training on the provided user feedback achieves at most minor improvements in binarization performance on the image the user provided feedback on and no improvement on similar images. This may indicate that the currently chosen feedback mechanism is not suitable for the objective of improving binarization performance through feedback.

Currently, I am investigating the impact different design choices have on the binarization performance of a Recurrent Neural Network (RNN) based approach for image binarization. For this, I have extended the RNN based approach by Afzal et al. [5] to take into account the surroundings of the pixels processed at one time step by providing a scaled down version of those surroundings together with the input pixels. The design choices under evaluation are the applied scale factor, the input pixel block size per time step, as well as the loss function used for training. The loss functions considered are unweighted binary cross entropy (BCE), statically weighted BCE and dynamically weighted BCE. In contrast to the statically weighted BCE loss, which discourages either foreground labelling errors or background labelling errors, the proposed dynamically weighted BCE loss uses the idea of pseudo F-Measure [6], weighting labelling errors differently depending on their location relative to the actual text. One version of this approach, trained with statically weighted BCE loss, has been submitted to this year's Document Image Binarization Contest.

C. Future Work

After finishing the current study on RNN based image binarization, I am planning to focus again on interactive machine learning for image binarization. This new attempt will address the identified problems with the previous study on interactive machine learning, while incorporating its worthwhile ideas.

The first change, compared to the previous approach, is to evaluate the performance of the proposed feedback mechanism in terms of its ability to achieve the binarization algorithm's maximum performance on a dataset through as few user interactions as possible. This is in contrast to assessing the overall improvement in binarization performance, which is limited by the algorithm's maximum performance. Evaluating a feedback mechanism close to this upper limit might always produce worse results, since the only remaining room for changing the labelling error is in trading a reduced labelling error on the feedback pixels with an increased labelling error for other pixels.

Furthermore, previously obtained results indicate that the ratio between feedback information provided by the user compared to the amount of parameters to adjust through learning might have been unsuitable for an efficient feedback mechanism. This problem can be addressed in two ways, either by reducing the number of parameters to learn based on the given feedback or by increasing the information content of the provided feedback. In my future research, I am planning to follow these two directions to develop a more suitable feedback mechanism for image binarization.

While the previous attempt to use few-shot learning techniques, such as siamese networks, for interactive binarization was a step into the direction of reducing the number of parameters to learn from feedback, this work could be extended in the future. Another possible step in this direction would be to learn to predict the tuning parameters for a binarization algorithm, such as HBA, from feedback.

On the other hand, the information content of the provided feedback could be increased, for example, by providing the user with explanations for a certain binarization choice, which can then be corrected by the user in case of an error. Possible explanations for binarization choices might be the identification of lines, which belong together or even the classification of characters. Based on these explanations, it might be easier for a user to provide more meaningful feedback.

REFERENCES

[1] N. R. Howe, "A Laplacian Energy for Document Binarization," in *2011 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2011, pp. 6–10.

[2] F. Westphal, H. Grahn, and N. Lavesson, "Efficient Binarization for Historical Document Analysis," 2016, Family History Technology Workshop (FHTW 2016).

[3] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 16, no. 3, pp. 247–258, 2013.

[4] F. Westphal, H. Grahn, and N. Lavesson, "Efficient Document Image Binarization Using Heterogeneous Computing and Parameter Tuning," submitted to IJDAR after requested revision.

[5] M. Z. Afzal, J. Pastor-Pellicer, F. Shafait, T. M. Breuel, A. Dengel, and M. Liwicki, "Document Image Binarization using LSTM: A Sequence Learning Approach," in *Proc. of the 3rd Int. Workshop on Historical Document Imaging and Processing*. ACM, 2015, pp. 79–84.

[6] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance evaluation methodology for historical document image binarization," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 595–609, 2013.

Indic OCR with Font and Layout Preservation

Student's name: Rohit Saluja
Supervisors of the thesis: Dr. Ganesh Ramakrishnan, Dr. Parag Chaudhuri, Dr. Mark Carman,
Dr. David Squire and Dr. Radha Krishna Pisipati
University: IITB-Monash Research Academy, Mumbai, India
Starting date of the Ph.D.: 31 December 2014
Expected finalization date of the Ph.D.: 31 December 2018
Email: rohitsaluja22@gmail.com

***Abstract*—We have developed a framework of interactively correcting the errors in Indic OCR. The system updates a domain vocabulary and the document-specific OCR confusions on the fly and is helpful in reducing the human efforts for documents in different Indian Languages with different inflections. We further used Long Short Term Memory (LSTM) networks to detect and correct errors in Indian Languages with varied inflections. Our model outperforms the state of the art results. We aim to progress forward in research and development of an end-to-end Indic OCR system that preserves layout and font, similar to Adobe pdf editor (which works well for English), for document images in Indian Languages.**

I. Short research plan

A. Introduction

Optical Character Recognition (OCR) is the process of converting the document images into an editable electronic format. This has many advantages like data compression, enabling search or edit options in the images/text, and creating the database for other applications like Machine Translation, Speech Recognition, and enhancing dictionaries and language models.

OCR in Indian Languages is quite challenging due to their richness in inflections. 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. Also, developing a highly accurate OCR system with an accuracy as high as 90% is not useful unless aided by the mechanism to identify errors.

B. Methodology

To develop an end-to-end interactive OCR system for Indian Languages, we started with the research related to error detection and corrections in Indic OCR. We have developed OpenOCRCorrect, an adaptive framework for interactively correcting OCR errors in the Indian documents, the source code of which can be downloaded from https://github.com/rohitsaluja22/OpenOCRCorrect. Our framework updates the domain vocabulary and the document-specific n-gram OCR confusions on the fly. We have also presented the benefits of reduction in human efforts due to OpenOCRCorrect for four different languages in ICDAR-OST workshop [1]. Since this was a system paper, 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,

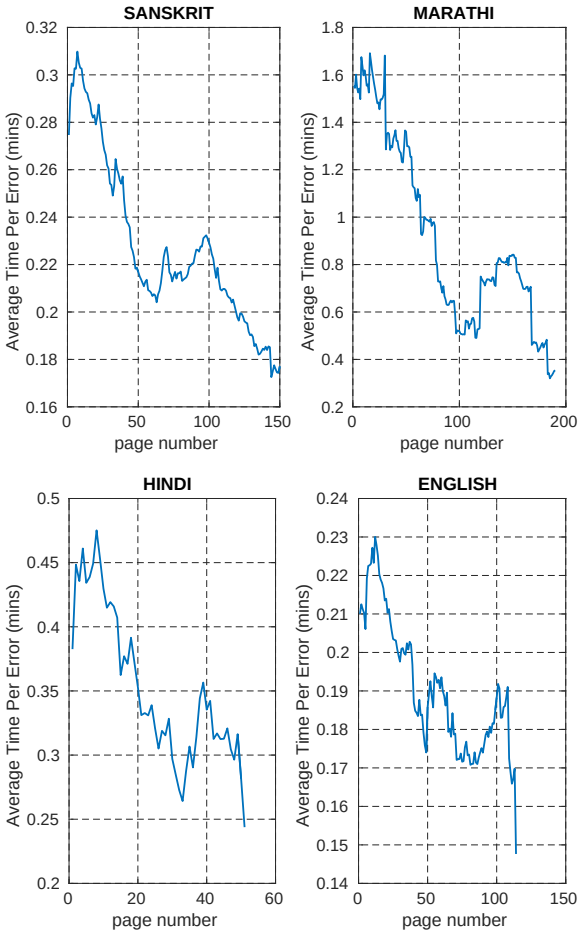

Fig. 1. System analysis of documents in different Languages. For Indian Languages, there is overall decrease in time per error as the user progresses in page number. The system also works well for the English document.

using adequate color coding (backed off by word conjoining rules) the incorrect words, updating the data structures on-the-fly and providing adequate suggestions using various auxiliary sources.

For OpenOCRCorrect, the error detection results are shown

Lang.	TP	FP	TN	FN	Prec	Recall	F-Scr.
Sanskrit							
LB	87.45	39.02	60.98	12.55	30.36	87.45	45.08
UB	91.62	0	100	8.38	100	91.62	95.63
Loglinear	85.13	17.84	82.16	14.87	48.62	85.13	61.89
Marathi							
LB	33.80	25.02	74.98	66.20	36.10	33.80	34.91
UB	15.20	0.03	99.97	84.80	99.49	15.20	26.37
Loglinear	76.93	3.83	96.17	23.07	78.77	76.93	77.84
Hindi							
LB	53.15	19.21	80.79	46.85	49.83	53.15	51.43
UB	44.72	1.55	98.45	55.28	91.18	44.72	60.01
Loglinear	64.34	15.25	84.75	35.66	55.97	64.33	59.86

TABLE I. ERROR DETECTION RESULTS. THE RESULTS OF BOTH DUAL ENGINE AGREEMENT (USED IN FRAMEWORK) AND LOG LINEAR CLASSIFIER ARE BETTER THAN LB (LOWER BASELINE) AND ARE IN-BETWEEN (OR BETTER THAN) LB AND UB (IDEALIZED UPPER BASELINE).

in the Table I. Here we used documents with 16k to 26k words in three different Indian Languages for training and testing a logistic linear classifier, and vary its threshold to optimize F-Scores. The LB(Lower Baseline) and UB(Upper Baseline) results are for dictionary lookup approach with training vocabulary and train + test vocabulary as dictionary respectively. As shown, the F-Scores achieved using logistic linear classifier are in-between (or better than) idealized (as this baseline uses test vocabulary from the ground truth which is not available) upper baseline. For Sanskrit, with the same dataset, we are able to achieve F-Score of above 64 with better features in our recent experiments. Also, we are able to correct 23% to 46% of errors using various auxiliary sources for this dataset.

Lang.	TP	TN	FP	FN	Prec.	Recall	F-Score
San.	92.63	94.54	5.45	7.36	94.84	92.64	93.72
Mal.*	87.56	94.23	5.77	12.44	93.82	87.56	90.58
Mal.	92.62	96.02	3.98	7.38	93.26	92.63	92.94
Kan.	98.51	97.28	2.71	1.48	96.92	98.41	97.66
Hin.*	72.30	90.90	9.10	27.70	89.30	77.22	82.82
Hin.	91.96	93.86	6.14	8.04	92.94	91.95	92.44

TABLE II. ERROR DETECTION RESULTS IN INDIC OCR. *STATE-OF-THE-ART RESULTS [2]

Our model outperforms the state-of-the-art results [2] in "Error Detection in Indic-OCR" in our ICDAR-2017 conference paper [3] as shown in Table II using Long Short Term Memory (LSTM) based Neural Network that corrects the wrong OCR words, and abstain from changing the correct words. Such model is able to learn the language as well as OCR specific error patterns. Here, we have worked on four different Indian languages with varied inflections. We use the dataset of around 1 lakh words for these four languages.

Lan.	Word Error Rate (WER)				%age words corrected by		
	OCR	Baseline		LSTM	Baseline		LSTM
		Lower	Upper		Lower	Upper	
San.	51.20	58.60	20.01	21.41	9.62	66.12	63.34
Mal.	37.28	48.43	10.83	10.59	9.09	58.20	78.30
Kan.	47.44	48.13	27.77	15.73	18.31	54.57	69.66
Hin.	46.80	45.43	34.17	16.71	20.94	27.46	72.47

TABLE III. DECREASE IN WER AND PERCENTAGE OF ERRONEOUS WORDS AUTO CORRECTED BY OUR MODEL.

We have taken a step in solving the Out of Vocabulary (OOV) problem for "Error Correction in Indic-OCR" using the LSTM model in the ICDAR-2017 conference paper [3]. The results are shown in Table III. It can be observed that we achieve a decrease in overall WER by at least 26.7% & at least 63.3% of the erroneous words were corrected by our model for all the languages. Here the lower baseline and upper baseline use the dictionary lookup, in addition to using n-gram confusions for breaking ties, from training vocabulary and train + test vocabulary respectively. The next step is to use such Recurrent Neural Networks adaptively in the back-end of OpenOCRCorrect to enhance the exploitation of human feedbacks.

C. Future Work

While developing an end-to-end OCR system for Indian languages, we are trying to solve the problem of font detection in parallel to OCR through a single model based on CNN and attention. Such approach is expected to avoid inter font differences and utilize inter font similarities. It would be very useful for the rare fonts (used in ancient books) for which huge datasets are not available for training. The recent paper beautifully summarizes the benefit of font detection before OCR, to achieve better OCR accuracy [4]. Here, the well known CNN models are used for font detection. There has been no relevant work on font detection in parallel to OCR through a single model as per our knowledge.

We are also facing issues related to layout segmentation in highly fragmental Indic documents. The recent literature in layout segmentation makes use of OCR text in Multimodal Fully Convolutional Neural Network to get the semantic information about the segment (of layout) class [5]. The problem with this approach is that in highly fragmental document images in Indic OCR, the OCR text of some of the regions is not available. The reason for this is that the text fragments in such documents are generally recognized as images in the layout analysis of the OCR engine. So, we are planning to use deep neural networks to utilize layout analysis and improve the quality of OCR text instead of using OCR text to enhance layout analysis in such documents. We plan to use memory based networks to extract OCR text (from the text segments in the image) in parallel to layout analysis and font identification via Convolutional Neural Network (CNN) based network. We propose to use Neural Turing Machines and/or Differentiable Neural Computers for solving the problem.

REFERENCES

[1] 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.

[2] V. Vinitha and C. Jawahar, "Error Detection in Indic OCRs," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, pp. 180–185.

[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] C. Tensmeyer, D. Saunders, and T. Martinez, "Convolutional neural networks for font classification," *CoRR*, vol. abs/1708.03669, 2017. [Online]. Available: http://arxiv.org/abs/1708.03669

[5] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles, "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Enhancing Text Spotting with Visual Context Information

Ahmed Sabir
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
asabir@cs.upc.edu

***Abstract*—This research proposal addresses the problem of text spotting – being able to automatically detect and recognize text in the wild. In this proposal, we approach this problem by drawing an inspiration from the successful development of machine learning algorithms with natural language understanding. In particular, deep learning and neural networks with extra information models. Due to the computational complexity of training and tuning the hyper-parameters of deep network architecture. Training an end-to-end text recognition system from scratch is a very expensive process. Therefore, our approach focuses on the integration of external prior information. We propose three approaches: first, a trainable language model integration to a pre-trained deep network. Secondly, a visual context bias information model, to exploit the semantic relatedness between word and image context. Thirdly, a combined model of multiple prior knowledge in a fusion based deep model. We also conduct experiments with public datasets, and compare our results to current state-of-the-art benchmarks.**

I. Short research plan

A. Introduction

Research in areas such as text recognition in the wild has not yet reached a mature level in implementation. There are still many challenges due to the many possible variations in textures, backgrounds, fonts, and lighting conditions that are present in such images. Many scene understanding methods successfully recognize objects and regions like roads, trees, sky in the image, but tend to ignore the text on the sign boards. Our goal is to fill this gap in understanding text in the scene. In addition, the automatic detection and recognition of text in natural images, *text spotting*, is an important challenge for visual understanding. There are two stages in text spotting: word detection and word recognition. The detection stage is to generate the bounding box around a candidate text image. The candidate words in the bounding boxes are recognized in the text recognition stage. In this proposal, we focused on improving the text recognition stage.

B. Related Work

The idea of end-to-end text recognition refers to algorithms able to automatically detect and recognize text in images. A text spotting system should be able to read text images that humans can read. The first text spotting system for image recognition which did not require a list or dictionary was proposed by [1]. The system extracted the character candidates via maximally stable extremal regions (MSER) and eliminated non-text candidates through a trained classifier. The remaining candidates were fed into a character recognition module, which was trained using a large amount of synthetic data. Later, [2] presented a new method for text spotting that combined the advantages of sliding windows and component based methods. In deep learning approach, [3] used a convolutional neural network (CNN) with unsupervised pre-training for text detection and character recognition. PhotoOCR is a text spotting system able to read characters in uncontrolled conditions [4]. It is a deep neural network (DNN) model running on HOG feature instead of image pixels. Another deep learning approach, [5] proposed a new CNN architecture, which allows feature sharing. Convolutional Recurrent Neural Networks (CRNN) presented by [6], a novel neural network architecture that integrates both CNN and RNN for image based sequence recognition. Deep learning based methods have two drawbacks. First, they require a huge dataset to train. Secondly, the computational burden of these methods is extremely high. A hybrid approach between deep learning and classical statistical model opens the possibility to overcome these limitations, and lead to introducing simpler models with better results. In this proposal, we investigate both approaches.

C. Methodology

In this work, we propose a simple approaches for introducing prior knowledge to the text spotting system. Additional knowledge, such as language model and visual context information, are essential to understand text in the scene. Our approaches to tackle these problems will be:

The first approach is an independent N-gram language model. The N-grams compute the probability of the candidate word directly from frequency counts taken from a large corpus. For instance, the word 'on' has higher probability than 'Onn'. The second approach consists of integrating a visual context bias. Contextual information is important to understand text in the scene. For instance, we can use word-embeddings to calculate the similarity distance between a candidate word and its visual context from the image. For example, if the output from a visual context model is 'street', 'car', or 'traffic lights', semantically related words such as 'parking' or 'left turn' are more likely to appear. Finally, we combine both visual context bias and language model information into a single model. The combination of both models should produce more informed predictions, that take into account both word context and linguistic probabilities.

D. Language Model

The language model is based on a N-gram model. N-grams are essential in any task in which we have to identify words from a noisy and ambiguous input. In speech recognition for instance, the input signal is noisy and most of the words extremely similar.

In this work, we apply a unigram language model over a deep network with pre-defined dictionary [5]. In particular, we extract the most probable words from a *baseline CNN* [5] and pass it into our language model. The outputs from the baseline are the candidate words and the associated probability based on softmax score $P_0(w) = p(w|CNN)$. The unigram language model (ULM) that was trained on a huge English corpus acts as a second independent dictionary. The main purpose of the unigram language model is to increases the probability of most common words in the pre-defined dictionary of the baseline, reranking the most probable words according to this modified probability $P_1(w) = p(w|CNN) \times p(w|ULM)$ from the language model. This hybrid approach opens the possibility of introducing higher-order trainable language models.

We trained two models on different corpora. The first model was trained on *Opensubtitles*[1], a database based on subtitles for movies. The corpus contains around 3 million words (combination of words and digits). Secondly, we trained a bigger model with *google book n-gram*[2] corpus, that contains around 5 million (only words). The overall results show that the language models have improved the baseline accuracy by 2% on both ICDAR03, ICDAR13 and 2.6% on COCO-Text dataset.

E. Visual Context Bias Information

The relation between text and its surrounding environment is very important to understand text in the scene. We propose to integrate visual context bias knowledge to the text spotting pipeline. In particular, we exploit the semantic relatedness between the spotted text and its image context. For example, the word 'parking' is more semantically related to word 'car' than it is to word 'banana' or 'pencil', thus it will be more likely to appear in a visual context where cars are present rather than where fruits or office appliances do.

We use a pre-trained state-of-the-art object classifier [7], [8] to find objects in the image, and we compute the semantic similarity between the candidate word and the detected elements. We use a word embedding approach to estimate the semantic similarity between the spotted word and its visual context information (VCI) and compute $p(w|VCI) = f(similarity(w, context\_objects))$. As in the language model step, this probability can be used to alter the total word probability and re-rank the best words suggested by the baseline CNN: $P_2(w) = p(w|CNN) \times p(w|ULM) \times p(w|VCI)$. We evaluate the performance of the model on the ICDAR 2017 robust reading challenge COCO-Text dataset [9]. The results show that by understanding the semantic relatedness between the spotted text and its visual context, the model improved the baseline by 4%.

[1] *https://www.opensubtitles.org*

[2] *https://books.google.com/ngrams*

F. Fusion Model

A deep model fusion based architectures have been successful in video action recognition [10], which is a joint classification between two deep networks. Our approach uses a similar architecture, to train two streams combined models, multiple prior knowledge and text recognition models. There are different approaches to combine two deep models, early fusion, late fusion and slow fusion, which we plan to explore.

G. Future Work

There are other lines of research that can be approached with this architecture setting:

1) Using higher order language model, such as bi-gram and tri-gram.
2) Re-train from scratch the combined models, text recognition and visual context bias model on the same dataset.
3) Integrate LM with long short term memory (LSTM) models for better word prediction.

ACKNOWLEDGEMENTS

I would like to thanks my supervisors Lluís Padró, Francesc Moreno-Noguer for guidance and Ernest Valveny for fruitful discussions.

REFERENCES

[1] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 770–783.

[2] ——, "Scene text localization and recognition with oriented stroke detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 97–104.

[3] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text detection and character recognition in scene images with unsupervised feature learning," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 440–445.

[4] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[5] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting." in *ECCV (4)*, 2014, pp. 512–528.

[6] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[7] 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*, 2016, pp. 770–778.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[9] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

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: Professor Antoine Doucet, Professor 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*—With the aim of preserving these documents and making them fully accessible, searchable in digital form, substantial efforts have been devoted to optical character recognition (OCR) that translates printed documents into digital ones. However, the poor quality of the paper input is the main reason for producing OCR errors and decreasing the performance of OCR systems. Therefore, various post-processing approaches have been proposed to detect and correct OCR errors. This is also my core research, which tries to find out how to build a multilingual post-OCR correction tool.**

I. Introduction

Paper-based documents contain valuable knowledge that gets many attentions from researchers and libraries around the world. In order to preserve and make these documents easily accessible, several efforts have been dedicated for optical character recognition (OCR) to digitize such documents.

A typical OCR system consists of five steps. Firstly, the paper-based document is transformed into the image-based document by an optical scanner. Through preprocessing, the OCR system locates the printed or written data regions and segments words into isolated characters. Then these characters are smoothed, and normalized to make the next processing stage easier. Feature extraction is one of the most important steps which extracts various types of features to recognize characters. These 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 into the database and be ready to be exploited digitally [1].

The main reason of decreasing the OCR systems' performance is a poor quality of paper document. Therefore, various post-processing approaches have been proposed to detect and correct OCR errors. They can be divided into three categories: manual error correction, dictionary-based error correction and context-based error correction [2].

The first approach lets human beings manually review and correct the output text. It is not only costly but also time-consuming because it requires a continuous human intervention. In addition, it is still error-prone since the human eyes are not error-proof.

A better approach was suggested, using a lookup dictionary to search misspelled words and correct them automatically. One of dictionary-based algorithms is the string-matching algorithm that weights the words in a text using a distance metric. The correction candidate has the lowest distance with the misspelled word is assumed to be the best choice [3].
Another correction algorithm applies character n-gram model [4], which is a subsequence of n characters of a string. A similarity measure is calculated by the fraction of n-grams which both strings have in common and unique n-grams of each string. However, this approach cannot correct errors related to their grammatical and semantic context.

The context-based error correction approaches were proposed to eliminate these disadvantages. Typically, this type of approaches learns from mistakes and then utilizes that information to suggest candidates. Neural networks or probabilistic methods are usually used for learning.
Tong and Evans [5] proposed an approach using n-grams and character confusion matrix. Only candidates which have enough common n-grams with the original token are retrieved. These candidates are then ranked using character confusion probabilities. After that, the Viterbi algorithm is used to get the best word sequence for the strings in the sentence [6]. In addition, the correct candidates are used to update the confusion matrix.
The data sparsity is always a big problem with n-gram approaches on word level like the approach above. In order to overcome that problem, Islam and Inkpen [7] proposed to use the n-grams contained in the Google Web 1T data 3-grams. This approach tries to find relevant candidates by frequency information based on Googles trigrams and chooses the best candidate by distance metric.
Some other approaches suppose a combination of trigram model and POS (part of speech) tagger [8], [9]. Each word of a sentence is examined to determine the grammatical order of the sentence based on mixed trigrams. The candidate which matches best in the sentence is chosen as correction. However, using a POS tagger on damaged historic texts is not practical.

In general, in order to deal with OCR errors from multilingual document, n-gram model is one of effective techniques. In fact, character n-gram is useful for detecting errors or learning character confusions. Additionally, word n-gram is a good choice for correct real-word error because it takes the context into consideration. However, the data sparsity of this model is still a big challenge. Other problem is that n-gram model is not powerful enough to capture long context [10]. Therefore, some other language models should be considered.

II. Methodology

This section will describe a general approach to deal with the OCR errors detection and correction. Our approach consists of three main steps, the details are described in Fig.1

Firstly, the errors are detected by using different methods, for instance character n-gram model, dictionary (lexicon) techniques.

Secondly, frequent error patterns learned from the ground truth are combined with lexicon for generating correction candidates, and the best N candidates will be chosen based on distance metrics (Levenshtein-Distance [3], Damerau-Levenshtein-Distance[11], ...), and be used as the input for the next step.

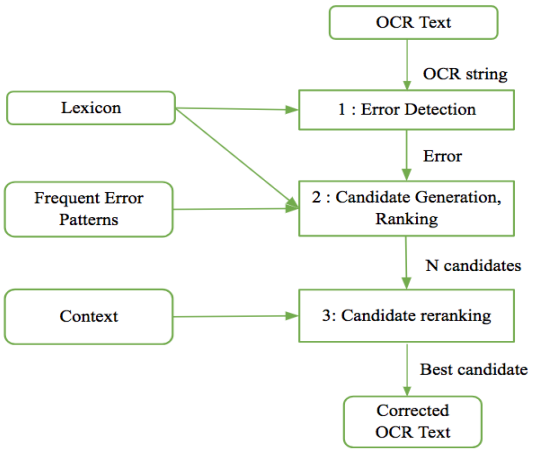

Fig. 1. The Post-OCR correction system architect.

The context is utilized to correct real-word errors by using different language models in following step. The traditional word n-gram model is a popular choice of many approach. To deal with the data sparsity of n-gram model, several smoothing technique have been suggested, such as Good-Turing smoothing [12], modified Kneser-Ney smoothing [13]. The modified Kneser-Ney smoothing (KN) is regarded as the best among smoothing techniques [14]. Despite of the popularity of traditional n-gram models, it was obvious that these models are not effectively enough to capture longer context patterns [10]. As a result, there have been some other advanced language models.

In practice, many words often occur again if they did appear in the recent history. Cache models [15] are supposed to deal with this regularity. This kind of model is often viewed as another n-gram model, which combines the basic n-gram model and the recent history.

Another way to deal with the data sparsity is to apply equivalence classes [14] obtained by clustering techniques. Firstly, each word is classified into a class. Then, n-gram model is trained on these classes. With patterns which were not seen in the training data, similar word in the same class could be used. The combination between the class based model and the n-gram model could improve the performance result.

Artificial neural networks can be successfully used for clustering. In addition, neural network are promising to capture longer context. In other words, neural network based Language Models can combine advantages of cached model and class based model. The main problem of this model is very large computational complexity [10].

III. Future Work

In first steps of my research, I have surveyed some related work to get an overview of my PhD topic. The n-gram model on character level and the frequent error patterns have been examined on some datasets with different similarity metric.

Follow the general approach, I will pay more attention on language models, and try to choose the suitable models for my research problem. Moreover, there are several OCR datasets with poor quality of ground truth, therefore it is worth taking unsupervised approach into consideration.

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] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[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] 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.

[6] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[7] A. Islam and D. Inkpen, "Real-word spelling correction using google web it 3-grams," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1241–1249.

[8] D. Fossati and B. Di Eugenio, "A mixed trigrams approach for context sensitive spell checking," *Computational Linguistics and Intelligent Text Processing*, pp. 623–633, 2007.

[9] ——, "I saw tree trees in the park: How to correct real-word spelling mistakes." in *LREC*, 2008.

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

[11] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[12] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE transactions on acoustics, speech, and signal processing*, vol. 35, no. 3, pp. 400–401, 1987.

[13] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 181–184.

[14] T. Joshua and J. Goodman, "A bit of progress in language modeling extended version," *Machine Learning and Applied Statistics Group Microsoft Research. Technical Report, MSR-TR-2001-72*, 2001.

[15] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic language model for speech recognition." in *HLT*, vol. 91, 1991, pp. 293–295.

Document Image Analysis of Balinese Palm Leaf Manuscripts

Student's name: Made Windu Antara Kesiman
Supervisor/s of the thesis: Jean-Christophe Burie, Jean-Marc Ogier and Philippe Grangé
University: Université de La Rochelle France
Starting date of the PhD: October 2014
Expected finalization date of the PhD: March 2018
Email: made_windu_antara.kesiman@univ-lr.fr

***Abstract*—In Southeast Asia, the collection of palm leaf manuscripts stores various forms of knowledge and historical record of the social life of southeast asian cultures long ago. This research collects the palm leaf manuscripts, and digitize it into digital image. But the main objective of this research is to bring added value to digitized palm leaf manuscripts by developing tools to analyze, index and access quickly and efficiently to the content of ancient documents. The document analysis tasks for palm leaf manuscript images are very difficult. The basic document analysis methods do not suffice for palm leaf manuscripts. This research studies and develops a complete document image analysis system and method, specifically for document images of Balinese palm leaf manuscripts, that includes several image processing tasks, beginning with digitization of the document, ground truth construction, text line and glyph segmentation and ending with glyph recognition. An appropriate system to transliterate the Balinese script to the Roman script is also developed. This transliteration system is needed to open a wider access to the precious content of historical Balinese palm leaf manuscripts.**

I. Short research plan

A. Introduction

The physical condition of natural materials from palm leaves certainly can not last long. As a very valuable cultural heritage which contains a wide variety of Southeast Asian social cultural life aspects, many palm leaf manuscripts discovered, for example in Bali, is a collection of the museum that has been in a state of disrepair due to age and due to inadequate storage conditions. Usually, palm leaf manuscripts are of poor quality since the documents have degraded over time. Natural materials from palm leaves and old paper manuscript certainly can not fight time. Equipment that can be used to protect the palm leaves to prevent rapid deterioration still relatively few in number, and therefore there is a need for the process of digitizing and indexing the palm leaf manuscripts. Palm leaf manuscripts offer a new challenge in document image analysis system due to the physical characteristics and conditions of the manuscripts. Document images of palm leaf manuscript have some typical characteristics. For Balinese manuscripts, there are 2 challenges: 1) the poor quality of palm leaf manuscripts, and 2) the complexity of Balinese script. Written on a dried palm leaf by using a sharp pen (which looks like a small knife) and colored with natural dyes, palm leaf manuscript images provide a real challenge in the binarization process to separate text from the background, as the early important step in document image analysis. The text line segmentation, the glyph segmentation, and the glyph recognition process are not trivial for palm leaf manuscript images. The palm leaf manuscripts contain discoloured parts and artefacts due to aging and low intensity variations or poor contrast, random noises, and fading. Severals deformations in the character shapes are visibles due to the merges and fractures of the use of nonstandard fonts, varying space between letters, and varying space between lines. The Balinese script is also considered as one of the most complex script in Southeast Asia.

The collection of palm leaf manuscripts in Southeast Asia attracted the attention of researchers in document image analysis. A new specific scheme is needed for the document images analysis of palm leaf manuscript images. This research studies and develops a complete document image analysis system and method, specifically for document images of Balinese palm leaf manuscripts, that includes several image processing tasks, beginning with digitization of the document, ground truth construction, text line and glyph segmentation and ending with glyph recognition. An appropriate system to transliterate the Balinese script to the Roman script is also developed. In the first year of this research, we focused on the digitization process of palm leaf manuscript images, on the literature study about the binarization process, and on the construction of ground truth image dataset for the collection of palm leaf manuscripts. In the second year, we studied and implemented the text line and glyph segmentation methods, the isolated character glyph recognition methods and the word spotting method for palm leaf manuscripts. And in the third year, we developed and proposed a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts.

Fig. 1. Palm leaf manuscript images

B. Methodology

To achieve this objective, this research project will be divided in three parts.

Part 1 : Constitution of the database : standardized manuscript digitization and the construction of ground truth image dataset

This research is under the scheme of the AMADI (Ancient Manuscripts Digitization and Indexation) Project. The corpus for this research includes the palm leaf manuscripts in Balinese script from Bali, Indonesia. In order to develop and to evaluate the performance of the document analysis methods, the dataset and the corresponding ground truth data for palm leaf manuscripts is necessary. The digitization and construction of palm leaf manuscript images dataset (the transcription and data annotation process) were done in Bali. The collaboration with the local partners, in ICT as well as in philology; is essential to take the analysis of these ancient documents up. A safe and efficient standard procedure for digitizing was designed, in order to preserve the physical integrity of the manuscripts and to ensure a necessary quality to analyse the document.

In this first part of the research, we proposed and analyzed the need for a specific scheme for the construction of the ground truth of binarized images [1]. The purpose of our work is to achieve a better ground truth binarized images for low quality palm leaf manuscripts. We also did an experiment in a real condition to analyse the human intervention subjectivity on the construction of ground truth binarized image and to measure quantitatively the ground truth variability of palm leaf manuscript images with different binarization evaluation metrics [2]. We finally presented the AMADI_LontarSet, the first handwritten Balinese palm leaf manuscript dataset [3][4]. It includes three components of dataset as follows: binarized images ground truth dataset, word annotated images dataset, and isolated character annotated images dataset. The dataset is publicly available for scientific use.

Part 2 : Development of content analysis algorithms : the text line and glyph segmentation methods, the isolated character glyph recognition methods, the word spotting method.

Different approaches are possible to analyse the document and extract their content. The first one is to develop specific glyph recognition. In Balinese palm leaf manuscripts with the Balinese script, the alphabets originate all from traditional Indian Pallava script. Most of the text recognition methods which naturally proposed a sequential process to recognize the words as entity/unit will face this characteristic as a very challenging task. Based on our knowledge, no OCR is indeed available to process the languages used in the studied documents. It will be necessary to develop a specific glyph recognition for specific language.

In this second part of the research, we investigated the performances of six text line segmentation methods by conducting comparative experimental studies on the collection of Southeast Asian palm leaf manuscript images [5][6]. We also implemented the supporting glyph recognition for the transliteration of Balinese script [7]. The second approach is to use a word spotting approach [8]. It consists to recognize a word (group of glyphs) in its entirety without having to recognizing each glyph separately. To optimize the chance of success of the project, the two approaches will be studied simultaneously. A combination of both approaches will be probably used to improve the recognition rate. To develop such a prototype, the research work in common will be unavoidable.

Part 3 : Design of a complete scheme for transliteration engine, indexing and retrieval system

The alphabet and numeral of Balinese script is composed of 100 character classes including consonants, vowels, diacritics, and some other special compound characters. Therefore, using a character recognition system will help to transcript these ancient documents and translate them to the current language, to give an access to the important information and knowledge in palm leaf manuscript. Transliteration system is one of the most demanding systems which has to be developed for the collection of palm leaf manuscript images.

In this third part of the research, we developed a complete scheme of spatially categorized glyph recognition for the transliteration of Balinese palm leaf manuscripts. To complete our scheme, we proposed an implementation of knowledge representation and phonological rules for the automatic transliteration of Balinese script on palm leaf manuscript. The phonological rules are finally built and are formally defined based on that glyph recognition results. A rule-based engine for performing transliterations is proposed. Our model is based on phonetics which is based on traditional linguistic study of Balinese transliteration.

C. Future Work

We will finally combine and implement the best method of indexing and retrieval for document images of balinese palm leaf manuscript. The final step of this research is the development of a transliteration system and word spotting methods, for the indexing and image retrieval of palm leaf manuscript images. By developing tools to analyze, to index and to access quickly and efficiently to the content of the manuscripts, this research tries to make the manuscripts more accessible, readable and understandable to a wider audience and to scholars all over the world.

REFERENCES

[1] M. Kesiman, S. Prum, J.-C. Burie, and J.-M. Ogier, "An initial study on the construction of ground truth binarized images of ancient palm leaf manuscripts," in *13th Int. Conf. Doc. Anal. Recognit. ICDAR*, 2015.

[2] M. Kesiman, S. Prum, I. Sunarya, J.-C. Burie, and J.-M. Ogier, "An analysis of ground truth binarized image variability of palm leaf manuscripts," in *5th Int. Conf. Image Process. Theory Tools Appl. IPTA*, 2015.

[3] M. Kesiman, J.-C. Burie, J.-M. Ogier, G. Wibawantara, and I. Sunarya, "Amadi_lontarset: The first handwritten balinese palm leaf manuscripts dataset," in *15th Int. Conf. Front. Handwrit. Recognit.*, 2016.

[4] J.-C. Burie, M. Coustaty, S. Hadi, M. Kesiman, J.-M. Ogier, E. Paulus, K. Sok, I. Sunarya, and D. Valy, "Icfhr 2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts," in *15th Int. Conf. Front. Handwrit. Recognit.*, 2016.

[5] M. Kesiman, J.-C. Burie, and J.-M. Ogier, "A new scheme for text line and character segmentation from gray scale images of palm leaf manuscript," in *15th Int. Conf. Front. Handwrit. Recognit.*, 2016.

[6] M. Kesiman, D. Valy, J.-C. Burie, E. Paulus, I. Sunarya, S. Hadi, K. Sok, and J.-M. Ogier, "Southeast asian palm leaf manuscript images: a review of handwritten text line segmentation methods and new challenges," *J. Electron. Imaging.*, vol. 26, no. 1, pp. 011 011–1–011 011–15, January 2017.

[7] M. Kesiman, S. Prum, J.-C. Burie, and J.-M. Ogier, "Study on feature extraction methods for character recognition of balinese script on palm leaf manuscript images," in *23rd Int. Conf. Pattern Recognit.*, 2016.

[8] M. Kesiman, J.-C. Burie, J.-M. Ogier, G. Wibawantara, and I. Sunarya, "Historical handwritten document analysis of southeast asian palm leaf manuscripts," in *Handwriting: recognition, development and analysis*, B. L. D. Bezerra, C. Zanchettin, A. H. Toselli, and G. Pirlo, Eds. Hauppauge, New York: Nova Science Publishers, Inc., 2017, ch. 9, pp. 227–267.

Deep Learning Based Approach to Handwritten Mathematical Expression Recognition

Student's name: Jianshu Zhang
Supervisor/s of the thesis: Jun Du and Lirong Dai
University: University of Science and Technology of China
Starting date of the PhD: 1 Sep 2017
Expected finalization date of the PhD: 30 June 2020
Email: xysszjs@mail.ustc.edu.cn

***Abstract*—Machine recognition of handwritten mathematical expressions (HMEs) is challenging due to the ambiguities of handwritten symbols and the two-dimensional structure of mathematical expressions. Inspired by recent work in deep learning, my PhD research focus on automatically recognizing HMEs based on deep learning. So far, I have presented the Watch, Attend and Parse (WAP) model to deal with off-line handwritten mathematical expression recognition (HMER) and the GRU-based encoder-decoder model for on-line HMER. Inherently unlike traditional methods, the proposed models avoid problems that stem from symbol segmentation, and they do not require a predefined expression grammar as they are data-driven models. These models take HME images or trajectory points as input and output their corresponding LaTeX strings automatically. The HME recognizer consists of convolutional neural networks (CNN) and recurrent neural networks (RNN), which have been proven to be powerful tools in processing images and sequential signals, respectively. Also, an attention mechanism is incorporated, it has been found to be one of the most distinct aspects of the human visual system. Validated on a benchmark published by the CROHME international competition, the deep learning based approaches significantly outperform the state-of-the-art methods on CROHME 2014 and CROHME 2016 test set, using only the official training dataset.**

I. Short research plan

A. Introduction

Mathematical notations play an essential role in scientific documents and are indispensable for describing problems and theories in math, physics and many other fields. Recently, people have begun to use handwritten mathematical notations as input due to the rapid emergence of new technologies such as digital pens, tablets, smart phones, etc. While this natural input method is convenient, it also requires the development of systems that are able to recognize HMEs. However, the automatic recognition of these HMEs is quite different from the traditional character recognition problems with more challenges, e.g., the complicated two-dimensional structures, enormous ambiguities in handwritten input and the strong dependency on contextual information.

HMER comprises two major problems: symbol recognition and structural analysis. These two problems can be solved sequentially or globally. Sequential solutions first segment the input expression into math symbols and recognize them. The analysis of two-dimensional structures is then carried out based on the best symbol segmentation and symbol recognition results. In contrast, the goal of global solutions is to recognize symbols and analyse two-dimensional structures simultaneously. The segmentation is achieved as a byproduct of the global optimization. In sequential solutions, the errors of symbol recognition and segmentation are subsequently inherited by the structural analysis. Conversely, global solutions can well address this problem, but they are computationally expensive as the probabilities for segmentation composed of strokes are exponentially expanded. Many approaches for performing structural analysis of an ME language have been investigated, among them, the grammar-based methods seem to be more dependable and have performed well in several systems. These grammars are constructed using extensive prior knowledge and the corresponding parsing algorithms are also computed.

Overall, both conventional sequential and global approaches have common limitations that my PhD research aims to address: 1) symbol segmentation during symbol recognition is inevitable, which introduces many difficulties; 2) structural analysis requires a priori knowledge that defines an ME grammar; 3) the complexity of parsing algorithms increases exponentially with the size of the predefined grammar.

B. Methodology

In my proposed models, I utilize the public LaTeX markup for HMER. Unlike common traditional ways, which recognize HMEs as expression trees, I recognize HMEs as their corresponding LaTeX markups. Besides, note that, mathematical expression (ME) language is a typical instance of two-dimensional languages. Achieve success in this domain could, in turn, accelerate progress in machine recognition of other two-dimensional languages.

In [1], We introduce Watch, Attend and Parse (WAP), a novel neural network model for off-line HMER. As shown in Fig. 1, this model learns to "watch" an HME image and parse it into a LaTeX string. So WAP has two components: a watcher and a parser, the parser is equipped with the attention mechanism. In WAP, the watcher is a fully convolutional network (FCN) encoder that maps ME images to high-level features. The parser is a Gated Recurrent Units (GRU) decoder that converts these high-level features into output strings, word by word. For each predicted word, the attention mechanism built into the parser scans the entire input ME image and chooses the most relevant region to describe a segmented symbol or implicit spatial operator. Unlike traditional approaches, WAP optimises symbol segmentation automatically through

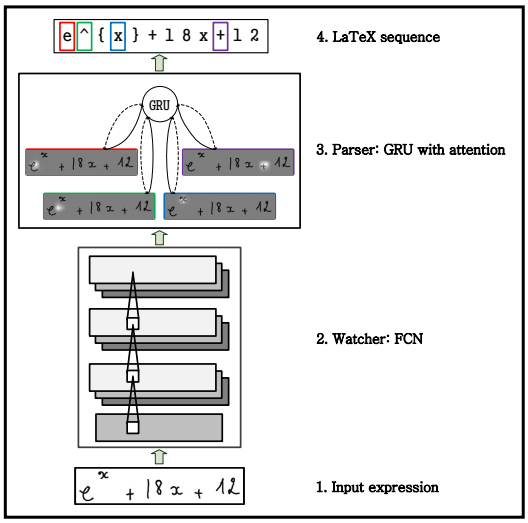

Fig. 1. Architectures of Watch, Attend, Parse for handwritten mathematical expression recognition.

its attention mechanism, and structural analysis does not rely on a predefined ME grammar. Moreover, the watcher and the parser are jointly trained. By doing this, not only the watcher extracts good features for the parser to decode but the parser also provides contextual information to tune the watcher and guide the attention. Although WAP is designed to deal with off-line HMER, it still outperforms the state-of-the-art method with an expression recognition accuracy of 46.55% on CROHME 2014 (Competition on Recognition of Online Handwritten Mathematical Expressions). In our experiments, we first transfer the generated LaTeX strings into MathML representations and then evaluate them by using the official tools provided by the organizer of CROHME. We use an old way for evaluation which compares the mathML trees rather than compares stroke based LG graph like other participants in CROHME 2014 and CROHME 2016, we do so as we do not have symbol segmentation results.

In [2], We propose to use a GRU-based encoder-decoder model for on-line HMER. We replace the FCN watcher with the GRU watcher so that we can make full use of trajectory information. The attention mechanism equipped in parser is converted from two-dimensional to one-dimensional. Validated on CROHME 2014 test set, the proposed model outperforms the state-of-the-art method with an expression recognition accuracy of 52.43% (shown in Table I).

In [3], We believe it is clearly superior to combine the on-line and off-line HMER model than a singular one. For example, in off-line HMER model, it is difficult to differentiate between the math symbol 'a' and 'α' but on-line model doesn't. While in on-line HMER model, we find the insert-strokes and inv-strokes have serious effect in the recognition accuracy but off-line model doesn't. In this work, we also blend a language model in the ensemble of on-line and off-line HMER model. Besides, a hybrid attention mechanism is also proposed which significantly improves the performance

TABLE I. CORRECT EXPRESSION RECOGNITION RATES (IN %) OF DIFFERENT SYSTEMS ON CROHME 2014 TEST SET.

System	Correct(%)	$\leq$ 1(%)	$\leq$ 2(%)	$\leq$ 3(%)
I	37.22	44.22	47.26	50.20
II	15.01	22.31	26.57	27.69
IV	18.97	28.19	32.35	33.37
V	18.97	26.37	30.83	32.96
VI	25.66	33.16	35.90	37.32
VII	26.06	33.87	38.54	39.96
P1	**42.49**	**57.91**	**60.45**	**61.56**
P2	**46.86**	**61.87**	**65.82**	**66.63**
P3	**52.43**	**68.05**	**71.50**	**72.31**

of HMER model. On CROHME 2014 test set, we achieve an expression accuracy of 62.37% by using only official training dataset. The performance even approaches MyScript (62.68%), who also uses private data for training their best system. Validated on CROHME 2016 test set, the model proposed in [3] also significantly outperform the current state-of-the-art method by using official training dataset.

C. Future Work

The current deep learning based approach to HMER still have the following limitations:

- Insufficient training data, the proposed neural network models reveal the over-fitting problem.
- Mathematical expressions whose LaTeX strings are quite long are difficult to parse correctly, which is due to the mismatch between training procedure and testing procedure.

My future research work will try to deal with the above limitations and focus on the following aspects:

- The innovation of the proposed Watch/Track, Attend and Parse (WAP/TAP) model. The watcher and attention mechanism could be improved further.
- It is inevitable to investigate the expression detection if we want to recognize handwritten mathematical expression in the natural scenes.
- Apply the proposed neural network in other document analysis tasks.
- Concerning that math expressions are not sequence, but a 2D language, generally represented by a tree, we will further re-introduce math grammar into our previous grammar-independent models.

REFERENCES

[1] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, 2017.

[2] J. Zhang, J. Du, and L. Dai, "A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition," in *ICDAR 2017*, 2017, in press.

[3] ——, "Track, attend and parse," unpublished.

Phd Research Work of Scene Text Detection

Student's name: Minghui Liao
Supervisor/s of the thesis: Xiang Bai
University: Huazhong University of Science and Technology, China
Starting date of the PhD: September 1st, 2017
Expected finalization date of the PhD: June 30, 2021
Email: mhliao@hust.edu.cn

***Abstract*—Scene text is one of the most frequent general objects in natural scene. Reading text in the wild usually consists of two steps including scene text detection and scene text recognition. My topic is mainly on scene text detection, which is aimed to localize texts in the natural images. Different from general object detection, the main challenges of scene text detection lie on arbitrary orientations, small sizes, and significantly variant aspect ratios of texts in natural images.**

I. Short Research Plan

A. Introduction

Scene text is one of the most general visual objects in natural scenes, which frequently appears on road signs, license plates, product packages, etc. Reading scene text facilitates a lot of useful applications, such as image-based geolocation. Despite the similarity to traditional OCR, scene text reading is much more challenging, due to the large variations in both foreground text and background objects, as well as uncontrollable lighting conditions, *etc*.
Generally, there are two steps for reading text in the wild: localizing and recognizing. My major topic is scene text detection, which is aimed to localize texts in the scene images. Different from general object detection, the main challenges of scene text detection lie on arbitrary orientations, small sizes, and significantly variant aspect ratios of texts in natural images. Owing to the inevitable challenges and complexities, traditional text detection methods tend to involve multiple processing steps, *e.g.* character/word candidate generation [1], [2], candidate filtering, and grouping.

B. Methodology

We proposed an end-to-end trainable fast scene text detector, named TextBoxes [3], which detects scene text with both high accuracy and efficiency in a single network forward pass, involving no post-process except for a standard non-maximum suppression.
TextBoxes [3] is inspired by an object detection method named SSD [4]. SSD aims to detect general objects in images, but fails on words that have extreme aspect ratios. Thus, we applied longer convolutional kernels and special default boxes to ease this problem.
Furthermore, we argue that word recognition is helpful to distinguish texts from backgrounds, especially when words are confined to a given set, *i.e.* a lexicon. We adopt a successful text recognition algorithm, CRNN [5], in conjunction with TextBoxes. The recognizer not only provides extra recognition outputs, but also regularizes text detection with its semantic-level awareness, thus further boosting the accuracy of word spotting considerably. The combination of TextBoxes and CRNN yields the state-of-the-art performance on word spotting and end-to-end text recognition tasks, which appears to be a simple yet effective solution to robust text reading in the wild. The code[1] of TextBoxes has been released for about half of a year and earns more than 200 stars on Github.
After that, we also adapt TextBoxes into a more robust text detector which can detect text with arbitrary orientations.

C. Future Work

I will focus on two important characteristics of text, including oriented texts and extremely long texts. Scene texts are usually in arbitrary orientations,
The extremely long texts are usually in non-Latin languages, such as Chinese, Japanese and Korean, where words are not split by blanks. There are some examples of RCTW dataset[2] are shown in Fig. 1
Firstly, it is hard to solve the challenge of the extremely aspect ratios with detecting the whole text directly. The main idea for solving these problems is to focus on the relations between parts of texts. Secondly, a good way to express the orientations of texts is worth to discuss about.

Fig. 1. Some challenging images in RCTW dataset.

[1]https://github.com/MhLiao/TextBoxes
[2]http://mclab.eic.hust.edu.cn/icdar2017chinese

II. Curriculum Vitae

Fig. 2. Student's picture.

A. Education

I got my B.E. degree in Electronics and Information Engineering of Huazhong University of Science and Technology in 2016. Then, I started to be a master student of Prof. Xiang Bai in 2016.
I am a PhD student of Prof. Xiang Bai in Huazhong University of Science and Technology at present.

B. Experience

1) I was a volunteer of VALSE 2016 conference[3].
2) I have given an oral presentation in AAAI-17 for [3].

C. Publications

I published a oral paper [3] in AAAI-17, which is a top conference in the field of artificial intelligence.

References

[1] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. CVPR*, 2012, pp. 3538–3545.

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *IJCV*, vol. 116, no. 1, pp. 1–20, 2016.

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

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *ECCV 2016*, 2016, pp. 21–37.

[5] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *CoRR*, vol. abs/1507.05717, 2015.

[3]http://valser.org

Graph-based representations for Document Analysis

Student's name: Pau Riba
Supervisor/s of the thesis: Josep Lladós and Alicia Fornés
University: Universitat Autnoma de Barcelona
Starting date of the PhD: October 2016
Expected finalization date of the PhD: 2019
Email: priba@cvc.uab.cat

***Abstract*—The research has been focused on proposing a graph-based technique for word spotting as an alternative to traditional statistic methods. The motivation is to describe deformations present in handwritten words using structural information. Moreover, graph indexation has been studied in order to speed up the previous process. This indexation is used to find candidates where graph matching must be applied. The indexation will discard several comparisons. Furthermore, a hierarchical graph representation is used to get information at different abstract levels in order to deal with noise of the original graph and also to get a smaller representation in terms of the number of nodes. Then, a stochastic graph embedding is proposed to codify the whole hierarchy. Finally, we explore the the utility of graph of terms to improve the retrieval performance of a given framework.**

I. Introduction

During the last years, graph-based representations have experienced a huge increase in their usage. Not only in social media analysis but also chemistry, visual recognition and image retrieval. In computer vision, these representations allow to enhance the information provided by classical statistical approaches with structure and relationship between elements, *i.e.* scene graphs [1]. Although there is a collection of techniques using these representations, few methodologies deal with large scale retrieval due to their complexity in terms of time. In the literature, graph embeddings are classical approaches to deal with large scale graph problems such as retrieval and classification. However, these embeddings have been hand-crafted by researchers leading to a representation that it is not able to generalise. A new trend in the community tries to extend deep learning methodologies to non-euclidean data such as graphs and manifolds [2], [3].

The research proposed by this thesis has focused on the development of graph-based techniques for document analysis. Firstly, a graph-based approach has been proposed in order to deal with the 2 dimensional structural information of handwritten words. Then, in order to deal with large-scale scenarios, we proposed an indexation framework able to encode local node information. This information is used to detect partitions on the graph where it is promising to perform a more accurate matching. However, the local information is highly influenced by the noise present in the data. Therefore, a hierarchical graph approach able to deal with different abstract levels of the graph have been proposed. This representation uses global graph information to reduce the size of these graphs allowing a faster matching between them. Despite, this pyramidal representation has been developed in order to speed up the matching between graphs, we noticed that it was able to obtain useful information.

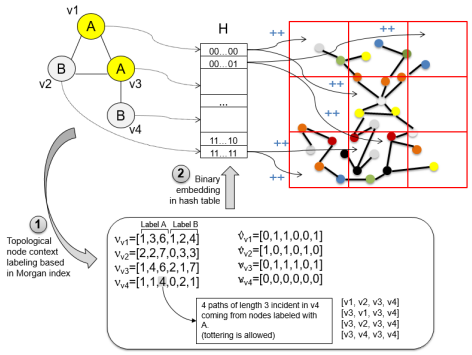

Fig. 1. Overview of the proposed node indexation framework. A topological node embedding based in the Morgan Index is computed. Then, this node representations are binarized to be used in a hash table that points to the candidate partitions in the graph where we are performing the search.

Hence, we proposed a pyramidal graph embedding combining information at different levels. Finally, graph of terms, where nodes are images and edges are similarity measures between the nodes, has been explored to compute a re-ranking for retrieval purposes. From now on, the idea is to work on deep learning structural approaches where the embeddings and classification of graphs are no more hand-crafted.

II. Methodology

Graph-based word spotting has been addressed in different works [4], [5]. However, we consider that primitives should be defined in order to obtain a robust representation of the handwritten word. In our previous work [6] we proposed a graph representation based on the convexities of the handwritten words. This work, faces a very specific problem, *segmentation based word spotting*. Therefore, a previous word segmentation is needed. Our next work focuses on two problems, speed-up the matching between word graphs pruning unnecessary comparisons and move to a *segmentation free* scenario. We propose an indexation scheme extending the attributes associated to graph nodes by a binary embedding function describing the local context of the corresponding node. Therefore, graph retrieval is formulated in terms of finding target (sub)graphs in the database whose nodes have a small Hamming distance from the query nodes. Figure 1 shows an overview of the proposed indexation framework. The before mentioned indexation framework relies on the local context of nodes. Thus, the node embedding becomes very sensitive to noise. In order to deal

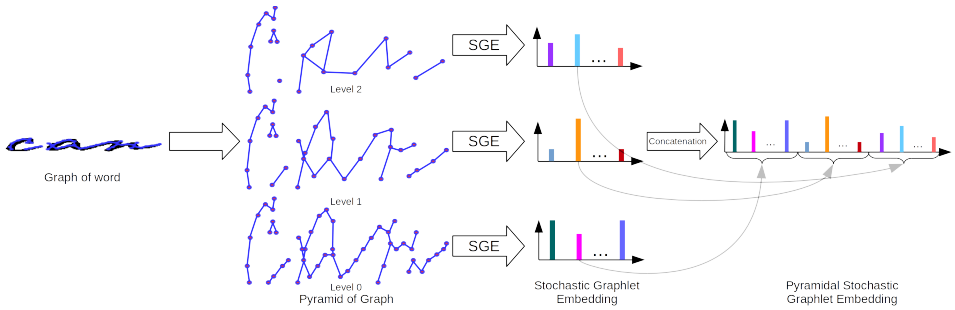

Fig. 2. Overview of our Pyramidal Stochastic Graphlet Embedding framework. Firstly, a pyramidal representation is computed. Secondly, a SGE is calculated for each level. Finally the different embeddings are concatenated in order to obtain a descriptor of the whole hierarchy.

with noisy and large graphs, we developed a hierarchical graph embedding [7] able to represent the input graph at different sizes, abstraction levels or resolutions. It has been proved, that using this representation, we are able to perform a fast graph comparison through a coarse-to-fine matching approach. This means, that we can perform most of the comparisons at coarse levels very fast, avoiding large graphs with lots of nodes. Table I shows a comparison in the BH2M dataset [8].

TABLE I. ORIGINAL GRAPHS [6], HIERARCHICAL [7] AND INDEXATION [9] FRAMEWORKS. MAP EVALUATES THE WORD SPOTTING PROBLEM; R AND SPC ARE COMPUTED ON THE SELECTED GRAPHS.

	mAP (%)	R (%)	SPC (%)	T/Q (s)
Original[6]	69.45	100.00	0.00	19.58
+ind.[9] (t=0.20)	66.13	92.54	46.13	16.34
+ind.[9] (t=0.30)	61.15	83.55	63.04	12.74
+abst.[7] (t=0.30)	**68.27**	90.91	69.98	**12.46**
+abst.[7] (t=0.25)	**61.71**	67.93	97.91	**3.94**

Seeing graphs at different detail levels can provide useful information to classify them. Hence, we propose to use these different abstraction levels together to construct a *pyramidal stochastic graphlet embedding* (PSGE). SGE randomly samples graphlets (subgraphs of different sizes) to construct a histogram. Thanks to a hashing function, identifying these graphlets is performed in real time. Graphlets of coarse graph levels, provide global information whereas fine levels provide information of the local structures. Figure 2 shows the pipeline used to construct the previously mentioned PSGE.

Lately, we have also focused on improving the information retrieval provided by other methodologies through graph-based techniques. Given a framework able to compare two images of a dataset, construct the k-nearest neighbour graph is a straightforward task. Using this graph, diffusion techniques [10] can be applied in order to obtain a more accurate similarity measure between samples. Moreover, the constructed graph follows the feature space topology avoiding wrong matchings.

III. FUTURE WORK

During this year, we plan to explore geometric deep learning[1] approaches dealing with structural data. These frameworks are able to perform a set of different tasks in the computer vision field. We can use it either to obtain information about whole graph or local information to the nodes. Moreover, these graphs can represent directly the information extracted from the images or graph of knowledge to improve the information extracted by non-structural methodologies. Currently, we have work implementing the framework proposed by Gilmer *et al.* [2][2]. Some experiments have been performed on classical graph datasets in order to test these approaches in a controlled scenario. The preliminar tests have proved that they are powerful techniques that can learn important information about the structure. However, this techniques needs a great amount of data to be able to generalise. Hence, classical graph datasets such as the one proposed by Riesen and Bunke [11] leads the model to overfitting to the training data.

Moreover, the previous work on PSGE will be extended. There, hierarchical edges will be included to the embedding. Graphlets will provide information about the contraction between levels combining resolutions of the graph.

REFERENCES

[1] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," *CVPR*, 2017.

[2] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *ICML*, 2017.

[3] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *arXiv preprint arXiv:1611.08097*, 2016.

[4] P. Wang, V. Eglin, C. Garcia, C. Largeron, J. Lladós, and A. Fornés, "A novel learning-free word spotting approach based on graph representation," in *DAS*, 2014, pp. 207–211.

[5] M. Stauffer, A. Fischer, and K. Riesen, "Graph-based keyword spotting in historical handwritten documents," in *S+SSPR)*, 2016, pp. 564–573.

[6] P. Riba, J. Lladós, and A. Fornés, "Handwritten word spotting by inexact matching of grapheme graphs," in *ICDAR*, 2015, pp. 781–785.

[7] P. Riba, J. Lladós, and A. Fornés, "Error-tolerant coarse-to-fine matching model for hierarchical graphs," in *GbRPR*, 2017, pp. 107–117.

[8] D. Fernández-Mota, J. Almazán, N. Cirera, A. Fornés, and J. Lladós, "Bh2m: The barcelona historical, handwritten marriages database," in *ICPR*. IEEE, 2014, pp. 256–261.

[9] P. Riba, J. Lladós, A. Fornés, and A. Dutta, "Large-scale graph indexing using binary embeddings of node contexts for information spotting in document image databases," *PRL*, vol. 87, pp. 203 – 211, 2017.

[10] X. Yang, L. Prasad, and L. Latecki, "Affinity learning with diffusion on tensor product graph," *PAMI*, vol. 35, no. 1, pp. 28–38, 2013.

[11] K. Riesen and H. Bunke, "Iam graph database repository for graph based pattern recognition and machine learning," *S+SSPR*, pp. 287–297, 2008.

[1] http://geometricdeeplearning.com/

[2] https://github.com/priba/nmp_qc

Computational Analysis of Writing Style in Digital Manuscripts

Student's name: Hussein Mohammed
Supervisors of the thesis: Prof. Dr.-Ing. H. Siegfried Stiehl* † and Dr. -Ing. Volker Märgner* ‡

* SFB 950 / Centre for the Study of Manuscript Cultures (CSMC), Universität Hamburg, Hamburg, Germany
† Department of Informatics, Universität Hamburg, Hamburg, Germany
‡ Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany

Starting date of the PhD: Nov 2015
Expected finalization date of the PhD: Feb 2019
Email: hussein.adnan.mohammed@uni-hamburg.de

***Abstract*—The goal of this PhD is to develop and implement a novel computer vision method to tackle the problem of writing style identification in digital manuscripts and to fulfil the requirements of the sub-projects of the Collaborative Research Centre (SFB 950) at the Centre for the Study of Manuscript Cultures (CSMC) such as the lack of sufficient training data and the prevalence of unbalanced data. A classifier has been developed for offline, text-independent, and segmentation-free writer identification based on the Local Naïve Bayes Nearest-Neighbour (Local NBNN) classification. The proposed method takes into consideration the particularity of handwriting patterns by adding a constraint to prevent the matching of irrelevant keypoints. Furthermore, a normalisation factor is proposed to cope with the prevalent problem of unbalanced data. The method has been evaluated on several public datasets of different writing systems and state-of-the-art results are shown to be improved. A thorough investigation is to be carried out to analyse the impact of degradation factors, e.g. resolution, and contrast.**

I. Introduction

A. Background

The first and the ongoing second phase of the Collaborative Research Centre (SFB 950) - "Manuscript Cultures in Asia, Africa and Europe" (2012-2019) at the Centre for the Studies of Manuscript Cultures (CSMC) in Hamburg is engaged in a fundamental research, investigating from both a historical and comparative perspective, the empirical diversity of manuscript cultures.

As a part of the scientific service projects, specifically Z03 [1] "Methods of image processing for the determination of visual manuscript and character features" , image processing methods are to be developed for determining the visual features in historical manuscripts. In this PhD project, a research on computational analysis of writing styles in digital manuscripts has to be carried out.

Identifying the style of handwriting is still a challenging task for law agencies and forensic documents analysis. Addressing this problem in historical manuscripts poses additional challenges due to the nature of these documents and different kinds of degradation factors. Therefore, a thorough investigation needs to be carried out to analyse the impact of these degradation factors.

B. Methodology

The requirements from the sub-projects at SFB 950 is collected and analysed, then a method has been developed and implemented to fulfil these requirements. Finally, a thorough analysis of the method is to be done with regards to degradation factors.

II. Short research plan

A. Analysing requirements from sub-projects in SFB 950

The requirements of the sub-projects can be divided into general requirements and project-specific requirements, and they can be summarised as follows:

The *general requirements* shared by all sub-projects is a method that can:

- Perform writer identification task.
- Work with different writing systems.
- Work with limited amount of training data.
- Work with unbalanced data.

The *Project-specific requirements* is a method that can:

- Sort samples by similarity to a given query.
- Provide an intuitive similarity measure.

B. Developing and implementing a method to tackle the mentioned requirements above

In this method, we propose an offline, text-independent, and segmentation-free writer identification method based on Local NBNN classification; See reference [2]. First, both query and labelled images of handwritten pages are converted to grey scale, then keypoints are detected and descriptors

are calculated from all images. Dense keypoints detection algorithms such as SIFT [3] or Features from Accelerated Segment Test (FAST) [4] are used for our proposed method in order to provide a sufficient number of keypoints for reliable nearest-neighbour search.

In order to match the calculated descriptors, a learning-free matching algorithm is used due to the fact that in many practical cases (as well as in many public datasets) the number of samples per writer is very small. A non-parametric learning-free classifier is proposed by Boiman et al. [5] and they demonstrated state-of-the-art results for image classification tasks. They showed that conditional class probabilities can be well approximated by the squared Euclidean distance to the nearest feature vector belonging to the correct class:

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \parallel d_i - \mathrm{NN}_c(d_i) \parallel^2 \right], \tag{1}$$

where $\hat{C}$ is the predicted class of the query image, C is the set of all classes, n is the number of query descriptors, d_i is a descriptor in the query image and $\mathrm{NN}_c(d_i)$ is the nearest-neighbour of d_i in class c.

The two main limitations of this approach are: The need to search for a neighbour in each class, and the bias toward classes represented by more descriptors. The first problem is tackled by McCann et al. [6] by searching for a neighbour only in the nearby classes, we can reformulate the Local NBNN Algorithm (2) [6] in equations as follows:

$$\begin{aligned} Dist^c_{local} = \sum_{i=1}^{n} \Big[\big(\parallel d_i - \phi(\mathrm{NN}_c(d_i)) \parallel^2 \\ - \parallel d_i - \mathrm{N}_{k+1}(d_i) \parallel^2 \big) \Big], \end{aligned} \tag{2}$$

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left(Dist^c_{local} \right), \tag{3}$$

where

$$\phi(\mathrm{NN}_c(d_i)) = \begin{cases} \mathrm{NN}_c(d_i) & \text{if } \mathrm{NN}_c(d_i) \leq \mathrm{N}_{k+1}(d_i) \\ \mathrm{N}_{k+1}(d_i) & \text{if } \mathrm{NN}_c(d_i) > \mathrm{N}_{k+1}(d_i), \end{cases}$$

and $\mathrm{N}_{k+1}(d_i)$ is the neighbour $(k+1)$ of d_i.

Furthermore, we proposed a normalisation step in order to tackle the second problem of unbalanced data, and a constraint to prevent the matching of irrelevant keypoints. The method has been evaluated on several public datasets of different writing systems and state-of-the-art results are shown to be improved.

C. Participating in international competitions

ICFHR-2016 Competition on Multi-script Writer Demographics Classification (Tasks 1A and 1B) [7]:

Tasks on writer identification (1A and 1B) has been carried out on writing samples (Arabic and English) of 800 writers from the database, 400 writers in each task. 200 test samples per task are used to evaluate the system performance. In this competition, we used SIFT keypoints and the normalisation factor has not been developed yet; nevertheless, the proposed method achieved the first rank in both tasks.

ICDAR-2017 Competition on Historical Document Writer Identification [8]:

This competition deals with the identification of writers in historical handwritten documents. The task is to generate a ranking of the images stored in the database according to the similarity of the handwriting. The dataset used for this competition consists of 3600 document images, which have been written by 720 different writers. Each writer has contributed 5 documents. In this competition, NBNN matching is used with SIFT keypoints instead of Local NBNN matching due to the size of dataset.

D. Using public datasets to evaluate the performance of the developed method

The proposed approach has been evaluated on several public datasets with different character sets, languages, and even musical scores to evaluate its generality. The state-of-the-art results are shown to be improved; see the publication in [2].

III. Future Work

A. Preparing datasets from the sub-projects at SFB 950

preparing a validation dataset with *confirmed* ground truth from scholars is a challenging task in most of the cases; nevertheless, it is very important for the detailed analysis of the method.

B. Analysing the method with respect to degradation factors

The degradation factors need to be systematically produced and controlled, and they need to be relevant to the possible degradations resulting from the digitisation process or found in manuscripts (e.g. resolution, rotation and contrast).

C. Compiling a draft of the thesis

References

[1] CSMC. (2015) Methods of image processing for the determination of visual manuscript and character features. [Online]. Available: http://www.manuscript-cultures.uni-hamburg.de/Projekte_p2.html#Z03

[2] H. Mohammed, V. Märgner, T. Konidaris, and H. S. Stiehl, "Normalised local naïve bayes nearest-neighbour classifier for offline writer identification," in *Accepted ICDAR paper*, 2017.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 105–119, 2010.

[5] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," *2008 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[6] S. McCann and D. G. Lowe, "Local Naïve Bayes Nearest Neighbor for image classification," *2012 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3650–3656, Jun. 2012.

[7] C. Djeddi, S. Al-Maadeed, A. Gattal, I. Siddiqi, A. Ennaji, and H. El Abed, "ICFHR2016 competition on multi-script writer demographics classification using" quwi" database."

[8] B. Gatos, M. Diem, F. Kleber, S. Fiel, G. Louloudis, V. Christlein, and N. Stamatopoulos. (2017) ICDAR 2017 competition on historical document writer identification (historical-wi). [Online]. Available: https://scriptnet.iit.demokritos.gr/competitions/6/

Document counterfeit detection through background texture printing analysis

Student's name: Albert Berenguel Centeno
Supervisor/s of the thesis: Josep Lladós, Oriol Ramos Terrades, Cristina Cañero
University: Computer Vision Center (CVC), Universitat Autonoma of Barcelona (UAB)
Starting date of the PhD: 01-02-2015
Expected finalization date of the PhD: 01-02-2018
Email: aberenguel@cvc.uab.es

***Abstract*—The objective of this Ph.D. is to validate the security background texture patterns present in documents, e.g., passports or banknotes. The main constraint is that we acquire the images using a smartphone in a non-controlled environment using visible light. We have used a classical computer vision classification pipeline to get a baseline results for the created dataset. A comparison study with the state-of-art texture descriptors has been done achieving good results. However we want to go one step further, proposing a new pipeline approach where we have a lack of counterfeit samples (real world scenario) and create a model which is able to learn from few samples. Furthermore we would also like that this model will be able to generalize or have a rapid learning for new documents. We think that going towards one-shot or few-shot learning schemas are the approaches that will accomplish this objective.**

I. Short research plan

A. Introduction

The use of counterfeit identity documents has been increasing steadily over the last years. Today, technology makes possible for anyone with a simple scanner, a high-end printer and some basic knowledge in image edition software to jump into the world of counterfeit. One security counterfeit measure is the background/security printing. The main difficulty is to detect this background counterfeit textures on a real industrial scenario without any constraint about the acquisition device and with a single image. The non-controlled acquisition environment raises problems such as sensor noise and luminance camera conditions that will affect the acquisition quality. This PhD thesis focuses in the detection of counterfeit identification documents using background texture printing analysis. Few research literature about document counterfeit is available when single images are acquired from a mobile device within a non-controlled environment. The novelty of this thesis is to propose a baseline and state-of-the-art results using the previous constraints.

B. Dataset

The copyrights laws and government data protection rules make counterfeit detection datasets very difficult to obtain. Several banknotes datasets have been built by other researchers, but are not public and the images are taken under controlled and constraint environment. We create our own highly secured textured documents datasets which are acquired as a normal user could do with a smartphone. We acquire the full document at close undetermined distance (allowing background) within a non-controlled environment. We join Euro banknotes and the Spanish identity card, due both have anti-counterfeit background security patterns, see Table I.

TABLE I. Created datasets. Obverse (A) or reverse(R) of the document.

Banknote	nImages	nTextures	Train/Test	%Counterfeit
€5 A	391	10	273/118	71.0/72.8
€5 B	331	10	231/100	66.6/72.0
€10 A	624	6	436/188	39.4/44.6
€10 B	614	6	429/185	40.0/37.8
€20 A	639	8	447/192	34.6/29.6
€20 B	622	5	435/187	34.0/37.9
ID	**nImages**	**nTextures**	**Train/Test**	**%Counterfeit**
ESPA	1865	10	1305/560	24.5/24.6
ESPB	1268	7	887/381	13.5/17.5

We center our efforts in detecting low level counterfeiting. Most of this type of counterfeit just follow the procedure of scan a real document, alter data and then print the document with a common commercial printer, which we call scan-printing procedure, see Fig. 1. Following this procedure, we expect that the texture background print design will loose detail, and hence it will be possible to classify it as counterfeit, see Fig. 2. The minimum resolution threshold of the cropped banknotes that forms the dataset is 400 dpi. All the images have been resized to 600 dpi as working resolution for stable intrinsic feature detection.

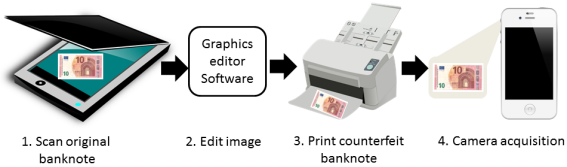

Fig. 1. Counterfeit generation images. The genuine scanned banknote is modified with a image software. Afterwards is printed with a high resolution printer to finally acquire the counterfeit printed banknote with a camera.

Two different configurations of the dataset are considered. The first configuration contemplates the full banknote image, where any region contained within the banknote is susceptible of being analysed. In the second, we select manually several

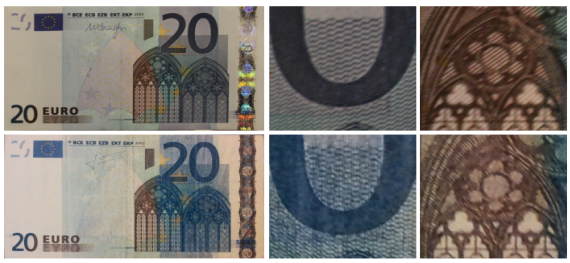

Fig. 2. Example of resolution loose in fine-line patterns. First row €20 genuine banknote and ROIs extracted, second row scanned and printed €20 counterfeit banknote. Acquired with BQ Aquaris M5 smartphone at 600 dpi.

background Regions of Interest (ROIs) containing fine-line patterns for its posterior classification.

C. Methodology

Our initial approximation was to create sparse coding dictionaries with existing algorithms to represent all the possible variations of each texture ROI [1], where algorithms such as K-SVD or SCSPM where tested. We have also compared our results with image quality assessment (IQA) metrics and a state of the art use of statistical moments with wavelet coefficients for banknote counterfeit detection. The results obtained with the sparse coding approaches outperformed other methods. Afterwards, we have built a end-to-end mobile-server architecture, which provides a service for non-expert users and therefore can be used in several scenarios with industrial purposes [3]. Inside this service-oriented architecture (SOA), we create a Counterfeit module which applies the previous best algorithms (K-SVD and SCPSPM) along with sparse texture feature representation. We apply BoVW, VLAD and Fisher Vectors (FV) to encode the dense SIFT descriptors. From this work, we have obtained the conclusion that dense SIFT with FV encoding can be used for counterfeit texture classification surpassing other algorithms. Following this line of work, instead of using uniquely dense SIFT as texture descriptor we evaluate 30 different state-of-art texture descriptors in terms of texture counterfeit classification [2]. In this comparison we can observe how HoG and CNN based descriptors stands out statistically over the rest if we include the F1-score/time ratio performance. Although hand-crafted features are outperformed in most cases by the learned features of the latests generations of deep CNNs, when the dataset is small sized and unbalanced, classical CNNs approaches for feature extraction may not be the best solution. Hence the ideal solution should be towards the creation of a single model, which could differentiate the introduction of artifacts between genuine and counterfeit textures, regardless of which type of document or texture we are dealing with.

D. Future Work

Following the previous reasoning, a future plan to distinguish the counterfeits is to use autoencoders which could learn the main characteristics of the background textures and combine it with a K-NN for posterior classification. However we think that we should continue our research towards CNN models which are able to learn progressively with small quantities of data. Just like how we teach our kids to differentiate between objects, we show them a few categories of an object and has to classify a new object which belongs to a training category. To imitate this training procedure we need to train a network to do rapid learning with attention and memory modules. One-Shot learning, Few-shot learning and Zero-shot learning are the topics which covers those ideas, being currently a field of interest for many researchers. We want to apply all this recent advances for rapid learning into the counterfeit detection area. We would like to create a fully differentiable network which is capable of learning the differences between a genuine and a counterfeit document using neural attention mechanisms.

A collateral effect of learning a fully differentiable network with sets of datasets with few annotated examples per class (meta-datasets), is that the actual variants of gradient-based optimization algorithms are designed to converge to a good solution after what could be many millions of iterations. Secondly, the network would have to start from a random initialization of its parameters, for each small dataset, which also hurts the ability to converge. For this reason, we believe it is mandatory to provide a systematic way to learn a beneficial common initialization that would serve as a good starting point to start training for the sets of datasets being considered. A solution for this learning problem is meta-learning, where two levels of learning are proposed. The first level consists of quick acquisition of knowledge within each meta-dataset. The second level acts as a higher level of knowledge extraction, which learns slower than the first level the information across meta-datasets. Zero-shot is the ideal solution, discarding the need of using counterfeit samples. Then we would only learn with the genuine samples and likewise a K-NN the network would be able to classify the genuine samples in its category by the proximity to their center clusters. However we want to start with simpler approaches like few-shot learning. Although having few counterfeit samples in our database, this samples can guide the network to learn a better representation. Therefore we have two possible ways of representing a document:

1) Individual textures ROIs: This schema treats each texture ROI independently from the others ROIs within the document. Afterwards the ROIs are joined to form a final classification decision of document authenticity. This schema is applied in [1], [3], [2]. Having independent ROIs allows higher combinations of meta-datasets and it can be seen as one way to perform data augmentation. However we have also to add an extra classifier for the final decision. Furthermore, we are loosing a global view of the document, which could lead the network to classify it incorrectly.

2) Sequence of texture ROIs: This schema creates a sequence of texture ROIs within the document, resembling human behaviour. First we would look at one texture ROI and depending on the details we expect to find a similar behaviour in another texture ROI of the background. This way of learning ensures that information across different textures regions are shared, but arises the question if this dependency will be beneficial or if we are imposing an unnecessary correlation between the ROIs which it might cause losing generality on the learned model. Using this approach we substantially restrict the number of learning meta-datasets combinations, but we expect to overcome this difficulty using the one-shot/few-shot learning approaches.

Detection and localization of text lines in heterogeneous document images with deep neural networks

Student's name: Bastien Moysset
Supervisor/s of the thesis: Christian Wolf and Christopher Kermorvant
University: INSA Lyon
Starting date of the PhD: November 2014
Expected finalization date of the PhD: December 2017
Email: bastien.moysset@a2ia.com

***Abstract*—This Phd. thesis is about detecting and localizing text lines in highly heterogeneous document databases with complex layouts. For this, we use Machine Learning techniques and, in particular, neural networks. The two main contributions, up to date, are the use of a Recurrent Neural Network within a CTC framework for the segmentation of a paragraph into lines and the adaptation of a convolutional neural networks to directly and efficiently predict line bounding boxes coordinates within a full document image by making its predictions local, by sharing the network parameters among the localizations, and by taking the context into account with Multi-Dimensional Long Short-Term Memory cells (MD-LSTM).**

I. Short research plan

A. Introduction

Text recognition and information extraction systems for document images, which have been developped during the late twenty years, need preliminary steps of document layout analysis. The aim of these steps is to automatically extract the text which is in the images. Because the text recognition takes line or word images as input, a good detection and localization of the text is crucial for good performances in recognition or keyword spotting.

Current techniques for text line localization are usually based on image processing heuristics and give decent results when applied on homogeneous datasets with simple layouts and digitized in good conditions. But these techniques show a lack of robustness when they have to deal with heterogeneous databases, complex layouts with several line orientations, presence of graphical elements or both handwritten and printed texts and different alphabets (latin and arabic for example) as was shown in recent evaluations [11].

The existence of consequent databases like the Maurdor dataset [11] enables the use of automatic learning techniques. In this field, the deep learning and especially the Convolutional Neural Networks (CNN) and the Recurrent Neural Networks (RNN) have given state of the art results on several challenging tasks in computer vision and speech processing as shown by last results in object detection competitions like PASCAL VOC or ImageNet [12] [13], text recognition competitions [9] and motion [14][15] or speech [16] recognition.

The aim of the thesis is to set up new methods for text line localization using this machine learning approach.

B. Contributions

1) Paragraph segmentation: First [4], we used recurrent neural networks to segment paragraphs in lines with the use of recurrent neural networks with an architecture and a cost function similar to those used for handwritten text recognition. It means that we used the Connectionist Temporal Classification (CTC) to align the sequences. This technique has the advantage of not needing the annotation of line positions because only the number of lines in the paragraph is needed. The limit of this method is that the segmentation is uniquely vertical and, therefore, cannot be applied to documents with complex layouts or to documents with skewed text.

2) Full page segmentation: In order to detect boxes in full pages, we have to tackle two main issues:

- We need to detect a variable number of objects.
- We need to be able to detect bounding boxes that overlap.

Several algorithms that use Machine Learning to localize objects use sliding windows and classify parts of the image as belonging or not to a given class. Successive lines within a paragraph can touch each other, especially for handwritten documents. For this reason, the sliding windows techniques need heavy post-processings which are often based on task-related heuristics. The algorithm presented in Erhan et al. [17] for natural scene object detection uses a neural network to directly predict the box coordinates and meet the two conditions mentioned previously. We implemented this method and tested it on the text line detection task.

But, due to the specificity of the full-page line detection task and of the datasets we work on, mainly the fact that the dataset are really smaller and that the number of objects to detect per image is higher, the Erhan et al. [17] algorithm had difficulties to generalize, even with the use of data augmentation or regularisation (namely dropout) during the network training.

We think that this may be related to the high number of parameters on the last layer of the network. For this reason, we

brought a new network layer that we call Space Displacement Localization (SDL) [5] that predicts the position of the objects locally and that share the parameters between the different zones of the image. This led to a drop in the number of parameters of our network and to a significant improvement of the performances on the Maurdor dataset [1].

In order to recover the global context information, lost due to the local behaviour of our network, we interleaved two dimensional Long short-term memory (2D-LSTM) layers between the convolutional layers. These contributions to the changes in the network architecture are illustrated in Figure 1.

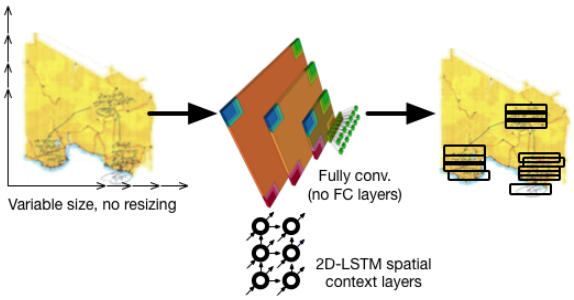

Fig. 1. A fully convolutional model with high spatial parameter sharing and fully trainable 2D-LSTM context layers learns to detect potentially many objects from few examples and inputs of variable sizes.

We also observed that the position of the predicted objects were more correct when the number of predicted coordinates was smaller and we proposed two techniques [3] [2] that use this characteristic to predict the objects and then recognize them. The first one uses two neural networks to detect respectively the points left and right from the text lines. These points are then paired to form the text line bounding boxes. The second predicts only the left side of the bounding boxes and trust the text recognizer to find the end of the line.

II. Curriculum Vitae

A. Education

- Phd. student at Liris, INSA Lyon: Detection, localization and typing of text in heterogeneous document images with deep neural networks (Since November 2014).
- Ecole Centrale Marseille: Graduate engineering school (2009 - 2012).
- MSc. in signal and image processing Université de Provence, Marseille (2011 - 2012).

B. Experience

- Research engineer in Computer Vision and Machine Learning applied to document image understanding at A2iA (since October 2012).

C. Publications

[1] B. Moysset, C. Kermorvant, and C. Wolf, Full-Page Text Recognition: Learning Where to Start and When to Stop. In International Conference on Document Analysis and Recognition 2017.

[2] B. Moysset, C. Kermorvant, and C. Wolf, Learning to detect and localize many objects from few examples, arXiv preprint arXiv:1611.05664

[3] B. Moysset, J. Louradour, C. Kermorvant, and C. Wolf, Learning text-line localization with shared and local regression neural networks. In International Conference on Frontiers in Handwriting Recognition 2016.

[4] B. Moysset, C. Kermorvant, C. Wolf and J. Louradour, Paragraph text segmentation into lines with Recurrent Neural Networks. In International Conference on Document Analysis and Recognition 2015.

[5] B. Moysset, P. Adam, C. Wolf and J. Louradour, Space Displacement Localization Neural Networks to locate origin points of handwritten text lines in historical documents. In Historical Image Processing Workshop 2015.

[6] B. Moysset, T. Bluche, M. Knibbe, F. Benzeghiba, R. Messina, J. Louradour and C. Kermorvant, The A2iA multi-lingual text recognition system at the second Maurdor evaluation. In International Conference on Frontiers in Handwriting Recognition 2014.

[7] B. Moysset, R. Messina, and C. Kermorvant, A Comparison of Recognition Strategies for Printed / Handwritten Composite Documents. In International Conference on Frontiers in Handwriting Recognition 2014.

[8] T. Bluche, B. Moysset, and C. Kermorvant, Automatic Line Segmentation and Ground-Truth Alignment of Handwritten Documents. In International Conference on Frontiers in Handwriting Recognition 2014.

[9] T. Bluche, J. Louradour, M. Knibbe, B. Moysset, F. Benzeghiba, and C. Kermorvant, The A2iA Arabic Handwritten Text Recognition System at the OpenHaRT2013 Evaluation. In International Workshop on Document Analysis Systems 2014.

[10] B. Moysset, and C. Kermorvant, On the evaluation of handwritten text line detection algorithms. In International Conference on Document Analysis and Recognition 2013.

References

[11] I. Oparin, J. Kahn, and O. Galibert, First maurdor 2013 evaluation campaign in scanned document image processing, in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 50905094.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in Advances in Neural Information Processing System, 2012.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, Going deeper with convolutions, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 19.

[14] S. Escalera, X. Bar o, J. Gonzalez, M. A. V. Ponce-L opez, H. J. Escalante, J. Shotton, and I. Guyon, Chalearn looking at people challenge 2014: Dataset and results. in ECCV Workshops (1), 2014.

[15] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, Multi-scale deep learning for gesture detection and localization, in Workshop at the European conference on computer vision. Springer, 2014.

[16] A. Graves, A.-r. Mohamed, and G. Hinton, Speech recognition with deep recurrent neural networks, in Acoustics, speech and signal pro- cessing (icassp), 2013 ieee international conference on. IEEE, 2013, pp. 66456649.

[17] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, Scalable object detection using deep neural networks, in IEEE Conf. on Computer Vision and Pattern Recognition, 2014.

Grouping and Recognition of Digital Ink Diagrams

Student's name: Amir Ghodrati
Supervisors of the thesis: Dr. Rachel Blagojevic, Prof. Hans Guesgen, Prof. Stephen Marsland
University: Massey University
Starting date of the PhD: 1/11/2015
Expected finalization date of the PhD: 1/2/2019
Email: a.ghodrati@massey.ac.nz

***Abstract*—Pen-and-paper sketches are used by designers and engineers (amongst many other groups) to convey ideas and concepts. Sketching on computers is easier and more natural following developments in stylus. There is a consequent interest in automatic recognition of sketched diagrams by computer. Given the variation in how users draw this is a non-trivial task. The first step in automatic diagram recognition is often considered to be grouping, whereby the individual lines(strokes) drawn by the user are combined into candidate shapes for consideration by the recognizer. In order to recognize an object we need to group the strokes, while the strokes cannot be grouped until the shape is known. Therefore, in this PhD study we treat the grouping and recognition as a simultaneous task that needs to be done at the same time. A method of hypothesizing candidate shapes based on spatial proximity has been developed, which requires a recognizer with rejection capability. In the following we will further discuss the details of the grouping algorithm followed with a plan on how a recognizer with rejection capabilities can be achieved.**

I. Short Research Plan

A. Introduction

Hand-drawn sketches are frequently used for creating conceptual designs. It is usually easier and quicker for people to draw sketches on papers before preparing a formal computerized format. It is desirable to allow users to produce digital versions of sketches with stylus and have them analyzed by the system. This has raised an interest in developing sketch recognition systems.

One of the challenges in diagram recognition systems is to determine which strokes together form a distinct object which is referred to as the "grouping" process. There is a potential chicken-and-egg problem with grouping and recognition, since identifying a group of primitives that form an object requires recognition of the object, while recognizing an object requires that the correct group of primitives (or features derived from them) are presented to the recognizer. In this PhD we are investigating an approach to the task of sketch recognition of diagrams based on simultaneous grouping and recognition of sets of strokes drawn by a user.

There has been some prior work in simultaneous grouping and recognition through rejecting invalid shape candidates. These approaches either use hard-coded rules and constraints [1], [2], [3], which obviously does not generalize to new situations well, or use negative examples [4], [5], [6]. However, the number of possible negative examples is larger than the number of positive examples, which means that a huge training set is required. We plan to develop a method based on novelty detection, which should avoid both of these problems, and to the best of our knowledge has not been attempted in the diagram recognition literature.

We have developed a simultaneous grouping and recognition system in which the grouper hypothesizes shape candidates and the recognizer accepts or rejects them. The proposed grouping algorithm requires a recognizer capable of rejecting candidates, which needs to be added to the existing recognizers since they do not perform this task. The grouping algorithm has been demonstrated using a mock recognizer. Our experiments showed that the algorithm works efficiently in terms of computation time and accuracy.

B. Methodology

The use phase of our sketch recognition system as shown in Fig 1, is comprised of two main steps: grouping/recognition of shapes, and connector recognition. In the grouping process, the grouper and the shape recognizer work simultaneously to find and recognize shapes in a given diagram. The connector recognizer gets a list of recognized shapes and unrecognized strokes from the grouper and identifies the connectors in the unrecognized strokes list. In fact, connectors are also shapes. However, their appearance can vary markedly between examples, and they have requirements that are not true of general shapes, such as the need to connect shapes. We therefore treat them separately. It has the additional benefit that parts of shapes are less likely to be misclassified as connectors.

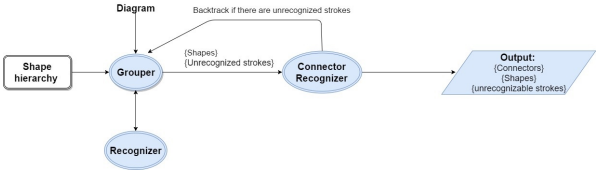

Fig. 1. Our sketch recognition scheme (use phase)

At this stage in our sketch recognition system we assume that all the input strokes are recognizable. Hence, if the connector recognizer's output contains unrecognized strokes, it implies that there might be a mistake from the previous steps. There is a backtracking process that can help solve some of the mistakes made in the previous steps. In the end, if the backtracking process still cannot find a solution where all strokes are recognized, then the solution with the highest number of recognized strokes will be reported.

The training process as shown in Figure 2 takes a set of labelled diagrams as input and produces the shape recognizer,

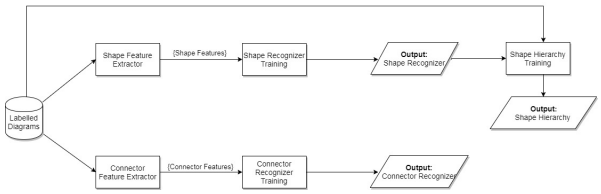

Fig. 2. Our sketch recognition scheme (training phase)

shape hierarchy and connector recognizer. The input diagrams are turned into feature sets to be used for the training.

The shape recognizer is a classifier trained to recognize an input as one of the shapes it has seen in the training set, and being able to reject the rest.

Some shapes are built from other shapes (i.e. complex shapes). We believe it will be useful to form such shapes into a hierarchy because it can help reducing the search space in the process of grouping. This hierarchy can be learnt by checking if any of the subsets of a shape represent one of the shapes in the training set. An example of a part of a shape hierarchy in digital circuit diagrams is shown in Figure 3.

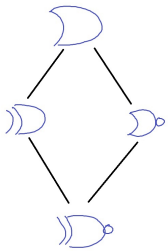

Fig. 3. A part of the shape hierarchy for digital circuit diagram

Up to this stage, we have worked on the use phase, which includes the neighbourhood search-based grouping algorithm and preliminary work towards connector recognition. As for the training phase, we have used the mock recognizer for learning the shape hierarchy as described above.

C. Future Work

In the remainder of this PhD we are planning to add the rejection capability to recognizers. There are several approaches that enables a recognizer to reject invalid candidates. The idea of these approaches is to measure the distance between the prototype shape and test shapes. One method to do that is a probabilistic approach, which puts a probability distribution over all shapes in the dataset. The classifier gives the probability for each class which then can be used to decide whether the given input is a shape or should be rejected as an invalid candidate. Different classifiers can be used to get the probabilities such as Bayesian Networks. Another approach for rejection is to measure the distance between the prototype shape and test shapes in the feature space. Linear Discriminant Analysis (LDA) searches for a combination of features that best separates classes. The aim of LDA is to minimize the distance within classes while maximizing the distance between classes. Based on the calculated distance for each class the decision can be made whether to accept or reject the input data. Both these approaches need a threshold value to make the decision whether the shape candidate should be rejected. We need to develop an algorithm that produces the threshold value from a training dataset. In this PhD study, we are planning to implement both approaches for rejection and compare their results.

To the best of our knowledge, there is only one research [7] that performs rejection which is used for shape auto-completion purposes. In this PhD study, we are aiming to add the rejection ability to recognizer introduced in [8] and compare our results with that of [7]. The choice of [8] as our recognizer is due to its high recognition accuracy, low computation cost and more importantly its support for complex shapes.

Once the rejection ability is added to the recognizer, we need to evaluate its ability in rejection. In the process of grouping, three types of candidates will be hypothesized: valid shapes, incomplete shapes and invalid candidates. For this research, after adding the rejection ability to a recognizer, we will evaluate the recognizer's ability in rejecting both incomplete and invalid candidates. The invalid and incomplete candidates can be hypothesized with the same grouping algorithm we have developed. It is also important to check if the rejection ability has affected the recognizer's accuracy in recognizing valid shapes. Once we achieve a full sketch recognition system, we will compare our results with other sketch recognition systems such as [9].

REFERENCES

[1] C. Alvarado and R. Davis, "Sketchread: a multi-domain sketch recognition engine," in *Proceedings of the 17th annual ACM symposium on User interface software and technology*. ACM, 2004, Conference Proceedings, pp. 23–32.

[2] ——, "Dynamically constructed bayes nets for multi-domain sketch understanding," in *ACM SIGGRAPH 2006 Courses*. ACM, 2006, Conference Proceedings, p. 32.

[3] T. A. Hammond and R. Davis, "Recognizing interspersed sketches quickly," in *Proceedings of Graphics Interface*. Canadian Information Processing Society, 2009, Conference Proceedings, pp. 157–166.

[4] M. Bresler, D. Prua, and V. Hlavc, "Simultaneous segmentation and recognition of graphical symbols using a composite descriptor," in *Computer Vision Winter Workshop*, vol. 13, 2013, pp. 16–23.

[5] M. Bresler, D. Prua, and V. Hlavc, "Modeling flowchart structure recognition as a max-sum problem," in *12th International Conference on Document Analysis and Recognition*. IEEE, 2013, Conference Proceedings, pp. 1215–1219.

[6] B. Kang, H. Hu, and J. J. LaViola Jr, "Mixed heuristic search for sketch prediction on chemical structure drawing," in *Proceedings of the 4th Joint Symposium on Computational Aesthetics, Non-Photorealistic Animation and Rendering, and Sketch-Based Interfaces and Modeling*. ACM, 2014, Conference Proceedings, pp. 27–34.

[7] C. Tirkaz, B. Yanikoglu, and T. M. Sezgin, "Sketched symbol recognition with auto-completion," *Pattern Recognition*, vol. 45, no. 11, pp. 3926 – 3937, 2012.

[8] T. Y. Ouyang and R. Davis, "A visual approach to sketched symbol recognition," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, ser. IJCAI'09. Morgan Kaufmann Publishers Inc., 2009, pp. 1463–1468. [Online]. Available: http://dl.acm.org/citation.cfm?id=1661445.1661680

[9] V. Deufemia, M. Risi, and G. Tortora, "Sketched symbol recognition using latent-dynamic conditional random fields and distance-based clustering," *Pattern Recognition*, vol. 47, no. 3, pp. 1159–1171, 2014.

Optical Music Recognition with Deep Learning

Student's name: Alexander Pacha
Supervisor/s of the thesis: Horst Eidenberger
University: TU Wien
Starting date of the PhD: March 2017
Expected finalization date of the PhD: March 2019
Email: alexander.pacha@tuwien.ac.at

***Abstract*—Optical Music Recognition aims to recognize and understand written music scores. My research goal is to improve Optical Music Recognition by applying Deep Learning techniques and train a system to learn and understand images of printed as well as handwritten music scores. The system should be capable of handling realistic images that might be imperfect and miss some information. Such a system can be used by musicians to robustly digitize music scores.**

I. Short research plan

A. Introduction

Music plays a central role in our cultural heritage with written music scores being an essential way of communicating the composer's intention to musicians that perform a piece of music. The music notation encodes the information into a graphical form that follows certain syntactic and semantic rules to encode pitch, rhythm, tempo and articulation. Optical Music Recognition (OMR) tries to recognize and understand the notation and the contents of an image for a machine to be able to comprehend the music. Given a system that is able to translate an image into a machine-readable format, the applications are manifold, including preservation and digitization of hand-written manuscripts, supporting music education or accompanying musicians that practice their performance, just to name a few.

Although considerable research has been conducted and many systems have been developed, that reportedly perform well on the specific set of music scores for which they have been designed for, the robustness and extensibility of these systems is limited due to the underlying architecture and used algorithms that discard information and propagate errors from one step to the next, e.g. an error in the binarization which is often the first step of an OMR system might cause the symbol detection to detect notes where there are none. Many algorithms have been proposed to improve individual steps of this linear process, but to the best of our knowledge, there exists no system that is capable of automatically recognizing a large set of real world data with satisfactory precision, good usability, and reasonably low editing costs. Such a system could be used to automatically digitize online sources like the IMSLP[1] dataset that contains high-quality scans of almost 400.000 public domain music scores. Many people could benefit from digitizing a large body of music scores that is accessible and searchable. As a result, there are ongoing projects to do. To support such projects, I propose a new approach that breaks with existing solutions: rather than designing features and defining rules by hand, the system should learn to extract features and appropriate rules by itself (given a certain amount of supervision). Ideally, such a system is capable of transcribing music scores as accurately as humans.

[1] http://imslp.org/

B. Methodology

For building a self-learning Optical Music Recognition system, I will break the problem down into pieces of increasing challenge that ultimately converge towards a complete system. I came up with following five questions that define my research program for the next two years. For each question, I will perform a data-driven investigation and attempt to answer it.

Can a machine mimic human behavior in ...

- Q-I distinguishing between music scores and arbitrary content?
- Q-II understanding the structure of music scores (staves, systems) and distinguish basic music symbols from each other and from the background?
- Q-III detecting and locating the music symbols (notes, rests, ornaments, accidentals, bar-lines, articulations, ...) in the scores?
- Q-IV understanding the relation of objects to each other in music scores (the relation between a note and the staff-lines, an accidental to the left of a note which relates to that note, etc.)?
- Q-V fully understanding the syntax and semantics of music scores (inferring the actual note from relative position, shape and preceding symbols such as key signatures or accidentals)?

In my opinion, each question can be solved using an appropriate model and sufficient evaluation data. Note that Q-V represents a complete system that is capable of reading scores and fully understanding their content like humans.

So far I've been able to answer Q-I and partially Q-II with yes. The respective conference paper has been submitted to an upcoming conference and is under review at the moment.

C. Future Work

Another paper that describes my work on Q-II has been submitted to the GREC workshop [1] and will be presented there. The extension of this, is what I am currently working on, with the goal of building a music symbol (object) detector that is capable of answering Q-III.

Once I have been able to answer Q-III, I will approach Q-IV and Q-V, by attaching recurrent neural networks (RNN) to my previous steps, as they can learn relationships in sequential data and already achieved remarkable results in Optical Character Recognition, a task that is comparable to OMR but in many regards simpler.

D. Novelty

The major novelty of my work is that I strive towards building a system that is capable of truly learning to read music scores with the quality of a skilled musician. Everything that a trained human can read, the computer should be able to read as well. It is not just applying a new technique to an existing field of research, but finally I want to be able to come up with an architecture that allows end-to-end training of the entire system.

II. Curriculum Vitae

Fig. 1. Alexander Pacha

A. Education

Doctoral studies
TU Wien, Austria
since 2017

Visiting researcher
Human Interface Technology Laboratory, New Zealand
2013

Elite graduate program Software Engineering
TU Munich, University of Augsburg and LMU Munich, Germany
2011 - 2013

Bachelor of Software and Information Engineering
TU Wien, Austria
2009 - 2011

B. Experience

Since 2014, I am a Professional Software Engineer at Zuehlke Engineering, Austria who is also working as a trainer for *Clean Code*.

Before that, I worked at several companies as an intern during the summer holidays, including Zuehlke Engineering (Munich), Andritz Hydro (Vienna), bwin (Vienna), IT-eXperience (Vienna), TU Wien (Vienna) and Siemens.

Besides that, I have worked three semesters as a tutor for *Algorithms and data structures* during my bachelor studies, where I was responsible for the exercises of the course and grading the exams.

C. Publications

Besides a few blog-posts, I have published the following papers and theses so far:

- My Bachlor thesis [2] along with a conference paper [3]
- My Master thesis [4]
- A conference paper on a talk I gave at a Software Quality conference in Vienna this year [5]

For a more extensive list of other works, including projects and blog-posts, please visit https://alexanderpacha.com as well as https://github.com/apacha and https://bitbucket.org/apacha for some of my open-source projects.

References

[1] A. Pacha and H. Eidenberger, "Towards a universal music symbol classifier," in *Proceedings of the GREC workshop at ICDAR*, 2017.

[2] A. Pacha, "Revision-control for excel," Master's thesis, Institute of Software Technology and Interactive Systems, TU Wien, 2011.

[3] R. Mordinyi, A. Pacha, and S. Biffl, "Quality assurance for data from low-tech participants in distributed automation engineering environments," in *Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on*. IEEE, 2011, pp. 1–4.

[4] A. Pacha, "Sensor fusion for robust outdoor augmented reality tracking on mobile devices," Master's thesis, University of Augsburg, Germany, 2013.

[5] ——, "1, 2, 3 – Build!" in *Software Quality Days 2017*, no. Conference Journal 2017. Software Quality Lab GmbH, Linz, Austria, 01 2017, pp. 21–25.

Analysis of heterogeneous documents and user behavior to improve accessibility

Student's name: Axel Jean-Caurant
Supervisors of the thesis: V. Courboulay & J.-C. Burie
University: La Rochelle, France
Starting date of the PhD: October 2014
Expected finalization date of the PhD: end of 2017
Email: axel.jean-caurant@univ-lr.fr

***Abstract*—Today, there are more and more ways to access information online. However, users are not necessarily trained to use the tools at their disposal. Indeed, the digitization of many resources has a huge impact on the habits of researchers. This is especially true in the field of Social Sciences or Literacy where users are often unfamiliar with how computers process information. Thus, there is a need to train these users to detect and avoid search biases. At the same time, we should also improve the way information is accessed. The digitization processes are not perfect and can create bad quality data. In this thesis, we try to understand two distinct but related issues. How can we identify biases and improve the search practices of users? How can we improve the quality of data users are interested in?**

I. Short research plan

My thesis will be organized in two parts. The first one concerns the study of the behavior of users inside digital libraries. In this part, we explain how to observe users during a search task. We also talk about the need to explain the mechanincs behind search engines to the aforementioned users. The second part of this thesis focuses on the other aspect of the problem: the data. We present a way of aprehending heterogeneous data and different tools used to exploit it. We also talk about data quality and how to improve it in the context of OCR.

A. Users and Digital Libraries

There is an increasing number of platforms available to access heterogenous resources. The new challenge for researchers is to be able to find the relevant information inside this huge amount of data. The digital turning point of the last decade has tremendously change the way researchers seek information. If this is undoubtedly a big progress, this change has brought to light many biases that can impact the accessibility of documents. For example, there is a trend to overlook some results returned by a search engine. A typical user is only interested in the first few results without considering the rest. However, we rarely know how these results were generated and why a given documents has scored higher than another. Users often have the feeling to have access to the totality of information. However, the fact that you did not get any results after a search does not mean that the information you were looking for does not exist.

To limit the impact of thoses biases, we believe there is a need to train users to use search platforms. This is the reason why we developped an experimental platform similar to most Digital Librairies [1]. The first use of this platform is to observe users to try to understand their behavior [2]. We implemented several observers responsible for logging the value of features such as the query length, the items clicked in the search engine result page, the time the user spent investigating documents, *etc*. We conducted an experiment with forty master student at the beginning of their specialization in Cultural Heritage. By analysing the actions they took, we were able to identify the features able to discriminate between two main task archetypes (look-up tasks and exploratory tasks) defined in [3]. To be able to detect the task of type a user is engaged in as soon as possible, we differentiated the observation of the first query iteration (actions between the first user input and the next search) and the overall task. Results are presented in table I. Moreover, this platform can be used as a pedagogical tool to help users understand how they can avoid some important biases. By providing new insights on the data and the users' actions (*c.f.* figure 1), this tool can serve has a way to engage discussions about the different biases a user can encounter inside a Digital Library. We defined roles for a trainer and some learners. The trainer has the ability to modify the search context (different parameters of the search engine) dynamically.

	first query			overall task		
	T4			T4		
	T1	T2	T3	T1	T2	T3
			query length			
p	***	***	***	*	***	***
Z	-4.28	-3.33	-4.16	-2.56	4.05	-4.08
			duration			
p	0.1	***	***	***	***	***
Z	1.16	-3.75	-3.78	-4.33	-4.55	-4.53
			max scroll value			
p	0.06	**	**	**	0.18	*
Z	-1.87	-2.97	-3.74	-2.64	-1.34	-2.09
			clicked items position			
p	*	*	0.05	***	***	***
Z	-2.25	-2.31	-1.96	-3.39	-3.92	-3.51

TABLE I. Statistical differences between an Exploratory task (T4) and three different Look-up tasks (T1, T2 and T3). The left-hand side of the table describes the results during the first query iteration, while the right-hand side concerns the overall task. We used a Wilcoxon signed rank test. p-values with *s are statistically significant, with * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$.

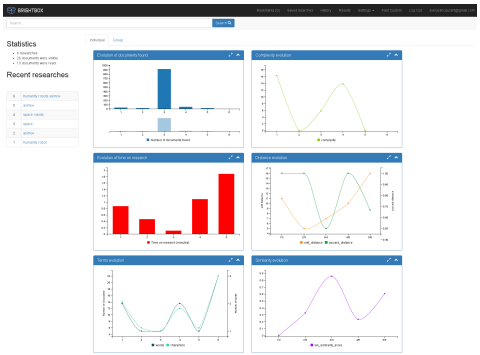

Fig. 1. Different behavioral features can be visualized using our pedagogical tool. They can highlight different biases that could affect the quality of a search.

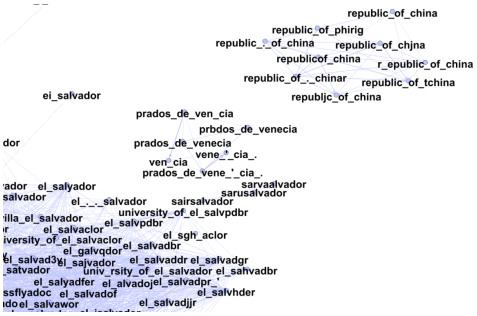

Fig. 2. Small portion of a similarity graph linking contextually similar named entities together. The weight associated with each edge is a normalize Levenshtein similarity.

B. Data and algorithms

Training users is one way to improve the accessibility of documents inside a Digital Library. At the same time, we believe it is important to develop new methods to limit the impact different algorithms can have on the quality of available data. For example, an OCR process is often used prior to indexing documents. However, these processes are not perfect and can create errors in the indexed documents by mistaking a character for another for example. If such an error occur on an important word (typically a named entity), it can limit the accessibility of a given document.

There are several works trying to correct OCR errors. Most of them make use of external knowledge such as lexicons or annotated ground truth to learn probabilistic error models . The main problem when using external knowledge is the loss of genericity. A model can perform well on a particular dataset but would have trouble to generalize. To overcome this issue, we developped a method ([4]) to automatically identify similar named entities within a corpus containing various OCR errors. It is based on two similarity measures. After identifying the named entities in a noisy corpus, we build a contextual language model using WORD2VEC [5] to infer a representative vector for each entity. By comparing the vectors of named entities and their edit distance (using Levenshtein for example), we build a graph as follows :

1) contextually similar entities are linked together
2) the weigth associated with the edge is the normalize Levenshtein similarity.

This graph is then clustered to extract similar named entities (*c.f.* figure 2). The choice of the clustering algorithm was conditionned by the following constraints :

- the number of clusters is unknown *a priori*
- a single word can belong to several clusters (the word "alite" for example, could be the result of a recognition error on the word "alive" as well as the word "alike")

This led us to use the weighted Clique Percolation Method (CPMw) [6] implementd in the software *CFinder*. However, because of the size of the graph, we had to limit the time the algorithm spent on each vertex. This resulted in approximative clusters. We evaluated our approach against a ground truth using clustering measures (MUC, B^3, $CEAF$) and a popular correction measure, the Word Error Rate. The obtained results show that we are able to correct around 50% of the detected named entities.

C. Future Work

First, we would like to experiment the benefits of using our teaching platform to support a class. We have to define learning objectives and see how this platform can help user to avoid some search biases. Considering the method to identify similar named entities, we would like to improve the clustering step. Using a more powerful (distributed) algorithm would certainly improve our results. There are several applications conceivable using these clusters of similar named entities. We can use them to suggest different possible alternative search for someone using a search engine. We can deduce some common OCR errors inside the corpus and correct them on every words using the information we gathered from named entities.

REFERENCES

[1] C. Suire, A. Jean-Caurant, V. Courboulay, and J.-C. Burie, "Ouvrir les boîtes noires : un outil pédagogique pour une approche critique de la recherche d'information en ligne." in *DH 2017: Proceedings of the Digital Humanities Conference*, 2017.

[2] C. Suire, A. Jean-Caurant, V. Courboulay, P. Estraillier, and J.-C. Burie, "User activity characterization in a cultural heritage digital library system," in *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*. IEEE, 2016, pp. 257–258.

[3] G. Marchionini, "Exploratory Search: From Finding to Understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1121949.1121979

[4] A. Jean-Caurant, N. Tamani, V. Courboulay, and J.-C. Burie, "Lexicographical-based order for post-ocr correction of named entities," in *Document Analysis and Recognition (ICDAR), 2017 International Conference on*, in press.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] G. Palla, I. Dernyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," in *Nature*, vol. 435, no. 7043, 2005, pp. 814–818.

Content Based Analysis and Retrieval of Architectural Floor Plans

Student's name: Divya Sharma
Supervisor/s of the thesis: Dr. Chiranjoy Chattopadhyay
University: Indian Institute of Technology Jodhpur
Starting date of the PhD: 27-07-2015
Expected finalization date of the PhD: 27-07-2019
Email: sharma.12@iitj.ac.in

***Abstract*—There is a recent growth in online platforms for real estate rent and sale. Moreover, architects refer to previous floor plan design history for cues while curating new plans. Architectural floor plans are essentially documents depicting 2D cross-section of a building. Text based description of a floor plan can at times be inadequate to capture the desired features a user wants in the floor plan. Also, manual look-up through the existing layouts can be a tedious and inefficient way to search for matching floor plans. Hence there is a need for automating the process of representation, matching and retrieval of similar floor plans. Thus, through this thesis we aim to propose new content to be analysed in architectural floor plans, develop new features to capture the appropriate content according to our application and finally, experiment with varying the modes of querying (images, hand-drawn sketches and hybrid representations) while retrieval.**

A. Introduction

Exponential growth of population and rapid urbanization has led to usage of online platform to search for homes by nearly 92% [1] of home buyers. To balance this demand and supply chain, while designing a new floor plan, architects often refer to existing projects. This process aids in providing insight on how similar architectural situations were solved in the past. An architect comprehends the buyers requirements, recalls a similar floor plan, and manually finds it from the previously designed projects. Moreover, with the immense progress of the digital world, several architectural projects are archived in digital form. Therefore, fast automatic techniques for retrieval of similar projects need to be adopted. This has motivated us to work on semantic matching of layouts.

A floor plan image captures the horizontal cross section of a house, as well as its interiors. In the past, work related to floor plans, has been in the area of *symbol spotting* in floor plans where given an image of a query symbol, location of the query symbol and associated documents are retrieved. For the same, usage of moment invariants such as Zernike moments for symbol spotting has been proposed in [2] and techniques like hashing the shape descriptors of graph paths (Hamiltonian paths) has been proposed by [3] which reduces time complexity. Another domain well researched is *floor plan analysis*, where existing works range from simple techniques for interpreting a hand-sketched floor plan for creating CAD models [4], [5] and automatically interpreting map drawings, to analysing differentiation between thick, medium, and thin lines for segmenting the room layouts in an efficient manner [6]. On the other hand, in sketch based retrieval of floor plans, only basic representation of floor plans [7], using primitive shapes like rectangles is analysed for correctness of relationships between rooms while sketching. Thus, upon doing a keen analysis of the existing work in the area of architectural floor plans, we came to a conclusion that retrieval in this particular domain is a challenging yet less researched area and has a lot of applications in today's digital scenario. For e.g., property buyers these days might have specific requirements while purchasing/renting a furnished house, such as : "*Required* 1 *master bed room,* 2 *smaller bedrooms, kitchen on the southern part of the house, with a garage to right side of the entry...*". In such a scenario, both external similarity, and interior designs need to be looked into. Such specific requirements can be met using a composite, automated framework that takes into account semantics and content inside a floor plan for retrieval.

A content based retrieval system can have different modalities of representing the input query floor plan, for e.g. query can be an image (I), a sketch (S) (a convenient choice) or a combination (H) of sketch, image and text to capture the buyer's intent perfectly. Further, the query and the existing layouts in the repository can be analysed for structural features as well as semantic features. Thus, such an amalgamation of various attributes while retrieval gives a great array of challenging problems to be dealt with while representation, analysis and finally matching and retrieval of floor plans.

B. Methodology

To initially approach the problem of floor plan retrieval we proposed a unified framework taking into consideration matching and retrieval based on similar layout designs as well as the room decor present in the layouts. The query mode proposed was query by example in the form of an image. Our proposed approach, not only matched and retrieved similar layouts with high accuracy through graph spectral embedding of the floor plans, it also went one level further in determining the room decor level matching between the type, number and arrangement of furnitures inside the layout for a complete semantic layout based matching [8][9]. Given a layout from the floor plan database we first identified walls in the layout which helped in segmenting the layout into semantic blocks or rooms present inside the floor plan. The rooms were numbered based on their centroid positions from the layout origin and thus, a sense of direction was incorporated during identifying room adjacencies. The result of this step was passed to the adjacent

room detection step, and relations between the rooms were determined. Considering these adjacencies we then generated a graph of the floor plan using the concept of topology graph. The adjacency matrix obtained from this layout graph was used to extract the spectral features. The layout graph was then embedded into pattern space to create a uniform feature vector. Such a spectral embedding helped in efficient matching of layout graphs through clustering similar graphs in the pattern space. For matching purpose two stage process of, (1) Room Layout Matching (RLM) and (2) Room Decor Matching (RDM) was carried. A match cost computing the similarity between each image of the layout category with the query image was used for rank ordering the results. We demonstrated our results on the SESYD dataset [10] and due to the high intra-category similarity of layouts in this dataset we achieved a perfect Precision on every Recall. While approaching this problem, we observed that there is a dearth of publicly available datasets to test floor plan retrieval algorithms. Two existing public datasets SESYD [10] and CVC-FP [11] have points like SESYD has high intra class similarity and low inter class similarity, rendering this dataset unsuitable for retrieval, and in the CVC-FP dataset, the samples are insufficient (122) in number for the task of floor plan retrieval. Therefore, we proposed a new floor plan dataset ROBIN with 510 floor plans. The unique characteristic of ROBIN is that the dataset is designed keeping in mind the needs of a potential buyer. Every prospective buyer wants to have certain amenities and functionalities in their house. Thus, in the dataset, there are three broad categories, which are different from each other in terms of the number and type of rooms present in a floor plan. Each broad category is further classied into 17 sub-categories depending upon the global layout shape of the floor plan. We have followed standard notations given in [10] for constructing the floor plans and have made ROBIN publicly available for other floor plan analysis tasks as well, apart from retrieval [12]. ROBIN helps in better visualization of the floor plans and aids in efficient capturing of various high-level features during fine-grained retrieval tasks. The growing application of deep learning in document processing tasks, intrigued us to implement and observe the results of the same in floor plan retrieval. For the said purpose, we applied the convolutional neural network on our ROBIN dataset and addressed two open issues: (i) how to learn new efficient deep learning feature representation for floor plan retrieval task?; and (ii) how the individual deep feature layers will affect the performance of the retrieval system?. The deep learning approach proved to be very efficient as compared to the state of the art matching approaches. It increased the mean average precion value from 0.23, as obtained using our earlier handcrafted feature approach [8] to 0.56 using deep learning[13]. Both the dataset and the deep learning approach is due to be published during the proceedings of ICDAR 2017.

C. Future Work

As discussed, we have been successfully able to implement query mode as image for identifying structural features while retrieval. In future we would study the role of sketch as modality while querying. Sketches come with their own set of challenges in terms of variability in representation and accurate recognition due to noise. In addition, sometimes a sketch produced by a user looks rough because of poor drawing skills or limited time for drawing. So it is possible that the sketches are different from natural scene images in some aspects. Thus, an effective sketch based image retrieval system must be able to diminish the ambiguity existing between sketch and natural images, which we seek to achieve through our system by proposing robust descriptors. Also while sketch representation, it would be interesting to note the role of temporal information, so as to gauge user's priorities as well as, to aid partial matching in our system. We would also like to explore multi-modal approach while querying layouts, where a user can draw on the image, as well as give textual inputs alongside, while specifying his/her requirements. Such multi-modal inputs would make the process of query representation very convenient to the end-user, and would provide a rich set of specifications to further accurately retrieve similar floor plan samples. As discussed, in our initial attempts, we have been able to capture the structural details, but we are continuously working towards deducing high-level semantic features from layouts. Such high-level features can be in the form of accessibility of each room, in terms of shortest path of a room to other rooms or to the exit in a floor plan. This scenario can find application in maintaining safety situations in big architectural buildings and retrieving floor plans abiding by the safety standards set by a regulatory body. During my future course of PhD research our aim would be to find such unique semantic features and content while analysis of floor plans, which can find application in real-world scenarios. In parallel, we also plan to deal with basic issues like observing the role of rotation and scale variation while capturing and processing floor plans.

REFERENCES

[1] "Propoerties Online," http://propertiesonline.com/reports/annual-real-estate-trends-report.pdf.

[2] G. Lambert and H. Gao, "Line moments and invariants for real time processing of vectorized contour data," in *ICIAP*, 1995.

[3] A. Dutta, J. Llados, and U. Pal, "Symbol spotting in line drawings through graph paths hashing," in *ICDAR*, 2011.

[4] S. Ablameyko, V. Bereishik, O. Frantskevich, E. Mel'nik, M. Khomenko, and N. Paramonova, "System for automatic vectorization and interpretation of graphic images," *PRIA*, vol. 3, no. 1, pp. 39–52, 1993.

[5] P. Vaxivière and K. Tombre, "Celesstin: Cad conversion of mechanical drawings," *Computer*, vol. 25, no. 7, pp. 46–54, 1992.

[6] S. Ahmed, M. Liwicki, M. Weber, and A. Dengel, "Improved automatic analysis of architectural floor plans," in *ICDAR*, 2011.

[7] S. Ahmed, M. Weber, M. Liwicki, C. Langenhan, A. Dengel, and F. Petzold, "Automatic analysis and sketch-based retrieval of architectural floor plans," *PRL*, vol. 35, pp. 91–100, 2014.

[8] D. Sharma, C. Chattopadhyay, and G. Harit, "A Unified Framework for Semantic Matching of Architectural Floorplans," in *ICPR*, 2016.

[9] ——, "Retrieval of Architectural Floor plans based on Layout Semantics," in *WiCV (CVPR)*, 2016.

[10] 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.

[11] L.-P. de las Heras, O. R. Terrades, S. Robles, and G. Sánchez, "CVC-FP and SGT:A new database for structural floor plan analysis and its groundtruthing tool," *IJDAR*, vol. 18, no. 1, pp. 15–30, 2015.

[12] "ROBIN dataset," https://github.com/gesstalt/ROBIN.git.

[13] D. Sharma, N. Gupta, C. Chattopadhyay, and S. Mehta, "DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans," in *ICDAR*, 2017.

Exploration of novel strategies for Online Writer Identification

Student's name:Vivek Venugopal
Supervisor/s of the thesis: Dr. Suresh Sundaram
University: Indian Institute of Technology Guwahati
Starting date of the PhD: 23 July 2014
Expected finalization date of the PhD: December 2018
Email:v.venugopal@iitg.ernet.in

***Abstract*—Our research is focussed around developing novel strategies for identifying the authorship of online handwritten documents. The directions pursued till date are as follows: (a) Exploration of codebook descriptors - wherein we first adapt the VLAD from the area of object retrieval and highlight on its potential drawback for online writer identification. Subsequently, we improve upon the same by formulating a novel descriptor. (b) Exploitation of the sparse learning framework for online writer identification. In this study, we consider the inclusion of ideas from information retrieval into the sparse representation to formulate a novel descriptor for each document. The dictionary for sparse representation are learnt from a set of histogram based features derived from sub-strokes, pre-segmented from the online trace. The research findings so far have led to the publication of one journal and a conference article.**

I. Short research plan

A. Overview of research topic

The problem of writer identification refers to the task of deciding on the authorship of a piece of handwritten document by using machine learning techniques. The recent advances in technology has enabled devices, wherein the data entry can be made via an electronic pen/stylus. The tip of the stylus records the dynamic information of the trace of the handwriting. The processing of such spatio-temporal data is termed 'online'. Most of the research on online writer identification have focussed on the development of text independent systems, wherein efforts are made to capture the style information of handwriting to identify the writer irrespective of textual content.

Over the last decade, there has been considerable research explorations in the area of online writer identification. A popular approach is that of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [1] adapted from the area of speaker identification. Likewise, techniques from the domain of information retrieval have also been proposed such as the term frequency-inverse document frequency (tf-idf) weighing scheme in [2], [3] and the notion of Latent Dirchlet Allocation in [4], [5]. Other works include those of features to describe the shape primitives, obtained from the handwriting [6], [7], [8], [9].

Lastly, we do mention of deep learning based approaches such as Convolutional Neural Network (CNN) and Recurrent Neural Networks, that have captured the interest of the online writer identification research community [10], [11], [12].

B. Methodology of Research so far

As a first research exploration, we investigated on the viability of deriving descriptors from a pre-learnt codebook generated with k-means algorithm [13]. The motivation for the same stems from the success of codebook descriptors to the problem of object retrieval in image processing.

The codebook comprises a set of code vectors with associated Voronoi cells computed from a clustering algorithm on a set of feature vectors along the online trace. In the context of handwriting, the generated codevectors are representative, on an average, of the frequently recurring writing patterns present among writers. The feature vectors corresponding to the handwriting samples of the same writer would be aligned in a similar relative location with respect to the nearest codevector to which they are assigned in the feature space. This trend may not be prevalent across the feature vectors from the handwritten samples of two different writers. Hence, to capture this information, we explore descriptors constructed from the codevectors for writer identification.

We begin by adapting the so called Vector of Local Aggregate descriptor (VLAD) to writer identification. However, we demonstrate that at times, it cannot effectively discriminate between writers. To overcome this problem, we propose a novel descriptor that improves upon the VLAD formulation. Based on a distance criterion, each feature vector of the test document is assigned to a specific codevector. The proposed descriptors take into consideration the score of each of the attributes in a feature vector with regards of the proximity to their corresponding value in the assigned codevector. This work has been published in the Expert System with Application journal in April 2017 issue. Very recently, we also evaluated the efficacy of this formulation for the task of online signature verification - and have a paper accepted as a poster in ICDAR 2017.

Following the success of sparse representation approach in offline writer identification [14], we conducted a study to demonstrate its effectiveness for online handwritten documents [15]. In this work, the identification was achieved by employing the sparse coefficients in a tf-idf framework. The dictionary for sparse representation was learnt from a set of histogram based features inspired by the Histogram of Oriented Gradient (HOG) used for the application of object detection in the area of computer vision [16]. This proposal has been published as an oral presentation in ICFHR 2016.

C. Future Work

Till date, we have considered the use of hand crafted features for the development of writer descriptors. In future, we would like to address on obtaining descriptors from features learnt by a CNN architecture. We would also compare the efficacy of the writer identification system while employing hand crafted features to that obtained from a CNN. Lastly, we would like to incorporate to combine the convolutional and recurrent neural networks for online writer identification.

REFERENCES

[1] A. Schlapbach, M. Liwicki, and H. Bunke, "A writer identification system for on-line whiteboard data," *Pattern recognition*, vol. 41, no. 7, pp. 2381–2397, 2008.

[2] G. X. Tan, C. Viard-Gaudin, and A. C. Kot, "Automatic writer identification framework for online handwritten documents using character prototypes," *Pattern Recognition*, vol. 42, no. 12, pp. 3313–3323, 2009.

[3] G. Tan, C. Viard-Gaudin, and A. Kot, "Online writer identification using fuzzy c-means clustering of character prototypes," in *ICFHR*, 2008, pp. 475–480.

[4] A. Shivram, C. Ramaiah, and V. Govindaraju, "A hierarchical bayesian approach to online writer identification," *IET Biometrics*, vol. 2, no. 4, pp. 191–198, 2013.

[5] C. Ramaiah, A. Shivram, and V. Govindaraju, "Data sufficiency for online writer identification: A comparative study of writer-style space vs feature space models," in *Pattern Recognition (ICPR), 2014 International Conference on*, 2014, pp. 3121–3125.

[6] B. Li, Z. Sun, and T. Tan, "Hierarchical shape primitive features for online text-independent writer identification," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 986–990.

[7] A. Namboodiri and S. Gupta, "Text independent writer identification from online handwriting," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[8] Z. Sun, B. Li, and T. Tan, "Online text-independent writer identification based on strokes probability distribution function," in *Advances in Biometrics*, 2007, vol. 4642, pp. 201–210.

[9] M. Gargouri, S. Kanoun, and J.-M. Ogier, "Text-independent writer identification on online Arabic handwriting," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, pp. 428–432.

[10] W. Yang, L. Jin, and M. Liu, "Chinese character-level writer identification using path signature feature, dropstroke and deep CNN," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015, pp. 546–550.

[11] ——, "DeepWriterID: An End-to-End Online Text-Independent Writer Identification System," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 45–53, 2016.

[12] X.-Y. Zhang, G.-S. Se, C.-L. Liu, and Y. Bengio, " End-to-End Online Writer Identification With Recurrent Neural Network," *IEEE Trans. Human Machine Systems*, 2017.

[13] V. Venugopal and S. Sundaram, "An online writer identification system using regression-based feature normalization and codebook descriptors," *Expert Systems with Applications*, vol. 72, pp. 196 – 206, 2017.

[14] R. Kumar, B. Chanda, and J. Sharma, "A novel sparse model based forensic writer identification," *Pattern Recognition Letters*, vol. 35, pp. 105 – 112, 2014.

[15] I. Dwivedi, S. Gupta, V. Venugopal, and S. Sundaram, "Online writer identification using sparse coding and histogram based descriptors," *2016 15th International Conference on Frontiers in Handwriting Recognition*, pp. 572–577, 2016.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

Deep Learning for Document Binarization and Segmentation

Student's name: Chris Tensmeyer
Supervisor/s of the thesis: Tony Martinez
University: Brigham Young University
Starting date of the PhD: January 2014
Expected finalization date of the PhD: August 2019
Email: tensmeyer@byu.edu

***Abstract*—Previously, my work has shown that Fully Convolutional Networks (FCN) achieve state-of-the-art performance in binarization tasks for document images. My current work combines an FCN with a Conditional Random Field (CRF) to encourage spatial consistency of the output, and uses graph cut for model inference. One project for future work is to produce realistic synthetic data for training binarization FCNs. Another project for general segmentation is to use a Recurrent Neural Network (RNN) to sequentially regress vertices of the object's bounding polygon. Formulated in this way, a loss function that directly minimizes Intersection over Union can be used for training.**

I. Short research plan

A. Introduction

My research area is Deep Learning with an emphasis on Document Analysis and Recognition (DAR) applications. I have explored a few subtasks in this field including document categorization [1], [2], binarization [3], segmenting page region from background [4], and font classification [5]. My current direction is to focus on binarization and segmentation.

B. Methodology

My approach to binarization is to use a Fully Convolutional Network (FCN) to learn a function that maps an input image $x \in \mathbb{R}^{D \times H \times W}$ to an output probability image $y \in \mathbb{R}^{H \times W}$, where $y_{ij} \in [0, 1]$ is the probability that pixel x_{ij} is foreground, and D is the number of input channels (e.g. 3 for RGB) [3]. While FCNs have been used before in similar tasks, one of the main contributions of this work is to use an FCN architecture that maintains the full input image resolution so that the output probabilities can be properly localized. Another contribution of this work is a loss function for training the FCN that is designed to maximize performance w.r.t. the psuedo F-measure metric for scoring binarizations, which is introduced in [6].

One of my current projects is an extension of [3] that integrates the FCN with a 2-class Conditional Random Field (CRF) to enforce spatial consistency of the output. A CRF is defined by an energy function that assigns a score to every assignment of classes to pixels, X.

$$E(X) = \sum_i E_i(x_i) + \sum_{i,j} E_{ij}(x_i, x_j) \tag{1}$$

where $x_i, x_j \in \{1, 2, \ldots, K\}$ for K classes, $E_i(x_i)$ is the unary energy of pixel i being assigned label x_i, and $E_{ij}(x_i, x_j)$ is the pairwise energy of pixels i, j having assignments x_i, x_j. Finding the optimal segmentation is formulated as

$$X^* = \text{argmin}_X \, E(X) \tag{2}$$

Zheng et al. [7] have considered the multi-class situation ($K > 2$), using a trainable FCN for the unary term in Eq. 1. However, they used approximate inference for solving Eq. 2 because for $K > 2$, Eq. 2 is NP-Hard. Additionally, they restricted the pairwise terms of Eq. 1 to be Gaussian functions of static pixel features (e.g. color and position).

My proposed method parameterizes both the unary and pairwise terms of Eq. 1 using an FCN and uses exact inference to solve Eq. 2 for $K = 2$, which is the case for binarization tasks. Ideally, the FCN would produce values for E_i and E_{ij} such that solving Eq. 2 produces the ground truth labeling. Exact inference for Eq. 2 can be formulated as a graph cut over a graph constructed from E_i and E_{ij}, which in turn can be exactly solved using popular max-flow solvers. In this case, X^* is a discrete quantity and is not differentiable w.r.t. E_i or E_{ij}, which prevents the usual gradient based discriminative training of the model. However, the model can be trained using an Energy Based Learning (EBL) routine [8], where the loss function for the model is some variation of

$$L(X^*; X_G) = E(X_G) - E(X^*) \tag{3}$$

where X_G is the ground truth segmentation. Essentially, minimizing Eq. 3 amounts to lowering the energy of the ground truth segmentation while simultaneously raising the energy of the current optimal predicted segmentation. Note that Eq. 3 is differentiable w.r.t. both E_i and E_{ij}, which allows the FCN to learn these terms in an end-to-end fashion.

For this project, I have not yet determined the best way to parameterize the CRF using the FCN. There also appear to be some learning difficulties, but these could be fixed with pretraining the FCN to find good apriori values for E_i and E_{ij} before joint finetuning of the whole model.

C. Future Work

1) Binarization: One of the challenges of learning based binarization algorithms is the limited amount of high quality training data. This is because it requires several hours to exhaustively annotate a single image with high accuracy.

As shown in [3], variety of data (backgrounds, noise, ink properties) is an important factor in improving performance.

One large possible source of data is to render synthetic images. We can take easily binarized text (e.g. IAMDB of handwritten images) and render it on images of blank documents. In this way, the foreground and background of the synthetic image come from real data and the ground truth is easily obtained.

The difference between this synthetic data and real examples is most noticeable at the boundary between the background and foreground due to the rendering process. Recent work [9] presents the Simulated plus Unsupervised paradigm, where a Refiner network is learned to map images from one domain (e.g. synthetic) to another domain (real data) from unpaired examples in both domains. This is done in an adversarial framework where the refiner is trained to fool a network that attempts to discriminate between real and refined images. Currently, consistency between the synthetic image and the output refined image is only enforced through a small L1 regularization term. I would like to extend this work such that the regularization is based on the ground truth of the synthetic images, so that the refined image can use the same per-pixel ground truth.

2) General Segmentation: Another proposed idea is to perform segmentation by iteratively predicting the vertices of the bounding polygon. While this idea will not work well for binarization (due to a high number of bounding vertices), it could be used for page segmentation or more general object segmentation in natural or medical images.

This proposal is inspired by the Polygon-RNN approach [10], which uses a Recurrent Neural Network (RNN) to sequentially predict the next vertex in a polygon. However, Polygon-RNN treats vertex prediction as a 1-of-N classification task over a downsampled image grid, which limits localization performance. The classification paradigm also means the model does not receive partial credit for being close to the target vertex. Furthermore, the favored metric for such segmentation tasks is Intersection over Union (IoU). If a single vertex is incorrect, it is possible for the impact on IoU to be significant or negligible.

My proposed model would solve these issues by turning vertex prediction into a regression task and training the model to directly minimize the IoU. I have implemented two differentiable neural network layers that take in a 2D grid of non-negative values and output a regressed point.

The first layer is inspired from physics and computes the center of mass of the input grid. The location of the center of mass is continuous w.r.t. each grid point and is differentiable. The downside of this approach is that it does not handle multiple modes in the input data. Furthermore, the more distant an input mass is from the center, the larger the gradient, similar to how the maximum likelihood estimator for the mean of a Gaussian distribution is sensitive to outliers.

The second layer solves these problems by operating on a single mode and directly finding a center value where many inputs have large values. This is done by first defining a continuous unnormalized kernel density function by spreading out the input values over a local neighborhood according to a predefined kernel function. This can be written as

$$v(x,y) = \Sigma_{ij} g_{ij} K(||(x,y) - (i,j)||) \tag{4}$$

where g_{ij} is an input grid value and K is a kernel function. While g is defined at integer values, v can be evaluated at any $x, y \in \mathbb{R}^2$.

The output of the layer is

$$x^*, y^* = \arg\max_{x,y} v(x,y) \tag{5}$$

x^*, y^* can be found by performing gradient ascent over Eq. 4. We estimate an initial starting point by evaluating $v(x,y)$ at each grid point and finding the maximum value. This can be done efficiently with a standard discrete convolution operation.

For the backward step of Eq. 5, we need an expression for $\frac{\partial x^*}{\partial g_{ij}}, \frac{\partial y^*}{\partial g_{ij}}$. I estimate these quantities using a finite difference approach.

Regressing the vertices in this fashion would segmentation models to be trained to directly minimize the Intersection over Union (IoU) metric, which is defined as

$$L(P;G) = \frac{A(P \cap G)}{A(G \cup P)} \tag{6}$$

where P is the predicted polygon, G is the ground truth polygon, and A computes the area of a region. This is desirable because IoU is often used to evaluate segmentation quality.

REFERENCES

[1] C. Tensmeyer, "Confirm: Clustering of noisy form images using robust matching," Master's thesis, Brigham Young University, 2016.

[2] C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," in *Proc. of ICDAR*, 2017.

[3] ——, "Document image binarization with fully convolutional neural networks," in *Proc. of ICDAR*, 2017.

[4] C. Tensmeyer, B. Davis, C. Wigington, I. Lee, and B. Barrett, "Pagenet: Page boundary extraction in historical handwritten documents," *arXiv preprint arXiv:1709.01618*, 2017.

[5] C. Tensmeyer, D. Saunders, and T. Martinez, "Convolutional neural networks for font classification," in *Proc. of ICDAR*, 2017.

[6] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance evaluation methodology for historical document image binarization," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 595–609, 2013.

[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1529–1537.

[8] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, 2006.

[9] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," *arXiv preprint arXiv:1612.07828*, 2016.

[10] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-rnn," *arXiv preprint arXiv:1704.05548*, 2017.