April 24-27

VIENNA

AUSTRIA

Short Papers Booklet

DAS 2018

13th IAPR International Workshop on Document Analysis Systems

13th IAPR International Workshop on Document Analysis Systems

DAS 2018
Short Papers Booklet

À la recherche du nom perdu – Searching for Named Entities with Stanford NER in a Finnish Historical Newspaper and Journal Collection

Teemu Ruokolainen
The National Library of Finland
DH projects
Mikkeli, Finland
firstname.lastname@helsinki.fi

Kimmo Kettunen
The National Library of Finland
DH projects
Mikkeli, Finland
firstname.lastname@helsinki.fi

Abstract — This paper presents work that has been carried out in the National Library of Finland to detect names of locations and persons in a Finnish historical newspaper and journal collection of 1771–1929. Work and results reported in the paper are based on a 500 000 word ground truth (GT) sample of the Finnish language part of the whole collection with different Optical Character Recognition quality.

Named Entity Recognition (NER), search, classification and tagging of names and name like frequent informational elements in texts, has become a standard information extraction procedure for textual data. NER has been applied to many types of texts and different types of entities: newspapers, fiction, historical records, persons, locations, chemical compounds, protein families, animals etc. Performance of a NER system is usually heavily genre and domain dependent. Entity categories used in NER may also vary. The most used set of named entity categories is usually some version of three partite categorization of locations, persons and organizations [1].

In our work we use a standard trainable statistical NER engine, Stanford NER[1]. Taking into account the nature of our data and complexities of Finnish language, our NER results can be considered reasonably good. With our ground truth data we achieve F-score of 0.89 with locations and 0.81 with persons. With re-OCRed Tesseract v. 3.04.01 output the F-score results are 0.79 and 0.72 for locations and persons.

Named Entity Recognition; historical newspapers; Finnish; Stanford NER

I. INTRODUCTION

The free historical 7.36 million page newspaper and journal collection of the National Library of Finland, Digi, is part of the growing global network of digitized newspapers and journals. Historical newspapers and journals are considered now more and more an important source of historical knowledge. As the amount of digitized journalistic information grows, also tools for harvesting the information are needed. Named Entity Recognition has become one of the basic techniques of information extraction from texts since the mid-1990's [1]. In its initial form NER was used to find and mark semantic entities like person, location and organization in texts to enable information extraction related to these kinds of entities. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software [1-2].

Our goal with usage of NER is to provide users of Digi better means for searching and browsing the historical newspapers and journals, i.e. new ways to structure, access and possibly also enhance information. Different types of names, especially person names and names of locations, are used frequently as search terms in different newspaper collections. They can provide also browsing assistance to collections, if the names are recognized and tagged in the newspaper data and put into the index [3]. Thus named entity annotation of newspaper text allows a more semantically-oriented exploration of the contents of a large archive.

We have earlier reported NER results for our collection with several different tools [4]. These tools were mostly rule-based tools for analysis of modern Finnish. None of them had real capability of handling 19th century Finnish with lots of OCR errors. The best results we were able to achieve were F-scores of around 0.60. As we performed the evaluation with our heavily erroneous OCR data, quite low scores were expectable. Nevertheless we gained invaluable experience in usage of NER and setting up an evaluation corpus. We became also much more familiar with our data.

For this evaluation, however, we took a new start. We had now available a 500 000 word token OCRed and manually checked ground truth wordlist for our re-OCR process [5]. This data consists of our current OCR, manually corrected ground truth (GT), and new Tesseract v. 3.04.01 OCR data. Out of the GT data we created a new evaluation and training corpus for NER. The training data was tagged first manually, and then additions were made semi-manually. As our NER tool we used a standard trainable statistical tagger, Stanford NER [6], that has been used earlier for example with Europeana historical newspaper data for Dutch, French and German [3].

II. CREATING THE EVALUATION COLLECTION AND RESULTS OF NER

We annotated first manually 170 pages (248 544 word tokens) out of our 479 page GT OCR collection with name tags of location (LOC) and person (PER). The annotation was performed by one person after discussion of the general principles and also based on our earlier experience with annotation in [4]. Subsequently, we trained a Stanford NER system using these pages, tagged the remaining 101 pages (211 034 word tokens) using the resulting system and manually corrected the automatically annotated pages. Thus our complete data set consists of 170 manually annotated pages (248 544 word tokens) and 101 semi-manually annotated pages (211 034 word tokens). In total there are 10 457 entities of person, and 13 266 entities of location marked in our data.

In addition to manually prepared named entity

[1] https://nlp.stanford.edu/software/CRF-NER.shtml

annotation, our corpus is accompanied with two gazetteers, name lists, which map words into semantically motivated categories. We compiled the gazetteers of person names and locations by combining different open source word lists. Our gazetteer of person names contains about 456 700 entries: about 18 600 of them are names in base form, the rest are automatically generated most important inflected forms of the names. Our gazetteer of locations contains about 333 670 entries. About 20 000 of them are names in base form, the rest are automatically generated most important inflected forms of the names.

In order to carry out the experiments, we divided the annotated data into two non-overlapping sections, training and evaluation sets. In the training set, we included 136 manually annotated pages and all 101 semi-manually annotated pages. The evaluation set consists of the remaining 34 manually annotated pages. The resulting training and evaluation sections contain thus 237 and 34 pages (381 356 and 67 223 word tokens), respectively. Furthermore, we created a second version of the evaluation collection, where the text has been produced by the OCR system Tesseract 3.04.01 [2] instead of manually checked GT (the NE annotations are manually corrected in the OCR set). Obtained results on the ground truth evaluation set are presented in Table 1.

TABLE I. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE GROUND TRUTH EVALUATION SET

Class	Precision	Recall	F1	# found	#gold standard
LOC	0.8872	0.8566	0.8716	1764	1826
PER	0.8408	0.7801	0.8093	1118	1205

Table 2. shows the final NER performance taking into account both errors yielded by the OCR process and the Stanford NER system.

TABLE II. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE OCR EVALUATION SET

Class	Precision	Recall	F1	# found	#gold standard
LOC	0.8527	0.7322	0.7879	1485	1826
PER	0.7856	0.6631	0.7192	1017	1205

In our earlier NER evaluation [4] bad quality OCR, i.e. noisy input, was the main reason for low performance of evaluated several NER tools. Now that we have available a good quality ground truth evaluation collection along with a lower quality re-OCRed version of the same data, we can see more clearly effects of OCR quality on the results. With the GT data performance of persons and locations can now be considered good, as the F-scores are 0.87 and 0.81. As shown in Table 2, realistic achievable new OCR quality impairs results with ca. 9–10% units when compared to GT data NER. As the achieved F-scores with OCR data are as high as 0.79 and 0.72, we consider the results acceptable and useful. We shall start to implement the resulting NER model of places and locations to our on-line presentation system.

[2] https://github.com/tesseract-ocr

III. CONCLUSION

We have reported in this paper usage of a standard trainable statistical NER tool, Stanford NER, for annotation of OCRed Finnish historical newspaper and journal data. We have created an evaluation collection of 67 223 tokens and trained Stanford NER with manually and semi-manually tagged data of 381 356 tokens. The performance we were able to achieve, can be considered reasonably good and useful. With locations and persons F-scores of 0.79 and 0.72 are reached in our re-OCRed output.

Our results show now clearly, what we predicted after our earlier experiments: improved OCR quality data will also improve NER results. We have now available OCR data out of which about 80–90% of the words are recognized by a morphological recognizer; in the old data the percentage was 73% [4]. With worse quality OCR NER results were not useful, with better quality OCR achieved results with persons and locations are on the level that has practical use. This result is in accordance with most of the results in NER of historical or OCRed data. NER experiments with OCRed data in other languages show usually improvement of NER when the quality of the OCRed data has been improved from very poor to somehow better [7].

Our data will be made available during the spring 2018 on the web pages https://digi.kansalliskirjasto.fi/opendata.

ACKNOWLEDGMENT

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020.

REFERENCES

[1] D. Nadeau and S. Sekine, “A Survey of Named Entity Recognition and Classification,” Linguisticae Investigationes, Vol. 30 No. 1, pp. 3–26, 2007.

[2] R. Grishman and B. Sundheim, “Message Understanding Conference-6: A brief history,” Proc. of the Sixteenth International Conference on Computational Linguistics (COLING 1996), vol. 96, pp. 466–471, 1996.

[3] C. Neudecker, “An Open Corpus for Named Entity Recognition in Historic Newspapers,” Proc. of LREC 2016, Tenth International Conference on Language Resources and Evaluation, 2016.

[4] K. Kettunen, E. Mäkelä, T. Ruokolainen, J. Kuokkala, and Laura Löfberg, ”Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910,” Digital Humanities Quarterly 11(3), 2017.

[5] M. Koistinen, K. Kettunen, and J. Kervinen, “How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine,” Proc. of LTC 2017, Nov. 2017, pp. 279–283.

[6] J.R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by Gibbs sampling,” Proc. of the Fourty-Third Annual Meeting on Association for Computational Linguistics (ACL 2005), pp. 363–370, 2005.

[7] T. Packer, J. Lutes, A. Stewart, D. Embley, E. Ringger, K. Seppi, and L.S. Jensen, “Extracting Person Names from Diverse and Noisy OCR Text,” Proc. of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada, 2010.

Bringing Paleography to the Table: Developing an Interactive Manuscript Exploration System for Large Multi-Touch Devices

Vinodh Rajan* and H. Siegfried Stiehl*†
*Research Group Image Processing, iXMan_Lab, Department of Informatics
†SFB 950 Manuscript Cultures in Asia, Africa and Europe
Universität Hamburg, Germany
Email: {sampath, stiehl}@informatik.uni-hamburg.de

***Abstract*—With Document Image Analysis gaining a strong foothold in the domain of paleography and the emergence of Human-Document Interaction as a narrative, we effectively need to reimagine the way scholars in humanities currently interact with digitized manuscripts. We initially evaluate related work in the current context of post-WIMP interfaces and then introduce our proposed system AMAP. With a strong focus on manuscript exploration, it attempts to harness the current state-of-the-art interaction paradigms to develop an intuitive system to engage with the manuscripts.**

***Keywords*-digital paleography; human-document interaction; human-computer interaction; interface design; intelligent interfaces;**

I. Introduction

Human-Document Interaction (HDI) has recently seen a surge in interest within the context of document analysis. With increased mainstream exposure to devices containing interactive multi-touch surfaces and novel input methods, there is a distinct need for domain-specific systems that allow natural interaction with digitized documents, while being both customizable and streamlined for real-world workflows and applications.

II. Background

Manuscript scholars and paleographers who have traditionally interacted with actual physical artifacts are gradually shifting towards interacting with the digitized versions instead. However, the interaction paradigm significantly varies between them. The latter does not afford similar degrees of freedom and intuitiveness. Dealing with collections of images and meta-data spread across files and formats coupled with the usage of multiple tools is also often difficult to manage in terms of workflow. This is further complicated by the fact that researchers frequently use tools that do not reflect their domain. Devices such as touch tables with their large interactive surfaces and multi-touch support provide us with an opportunity to design gesture-driven systems that would allow researchers to virtually engage with digitized manuscripts and visualize information as never before, all the while encouraging real-time collaboration with their colleagues.

III. Related Work

BVREH [1] was one of the earliest attempts to create an integrated workspace for researchers working with manuscripts. It was conceived as a *demonstrator* that enabled researchers to construct personal workspaces, add annotations and perform remote collaboration. The work substantially evolved into VRE-SDM [2], where pre-existing features were further refined and additional functionality such as transcriptions, image processing capabilities and integration to external and internal databases were added. However, the project does not appear to be in development anymore and none of the related files are available to download and use. More recently, VMR CRE [3] supports transcription, basic image manipulation techniques, collation and other relevant project management capabilities as well. There are also projects such as *DigiPal* [4] that provide a CMS-style web interface to upload manuscript images and annotate them based on characters' forms and hands.

All of the systems discussed are browser-based implementations that provide a very traditional Windows-Icons-Menus-Pointers (WIMP) interface to interact with them. This mode of interaction can be often counter-intuitive and inflexible for domain-specific applications such as these. They also do not support extensive image manipulation capabilities and other advanced but relevant features such as word-spotting, writer identification or any form of visualization. Currently, works like [5], [6] and *Turning the Pages* (*https://ceb.nlm.nih.gov/proj/ttp*) already show some interest in applying touch interfaces in our context. But, there are currently no integrated systems for manuscript researchers and paleographers that are capable of taking advantage of the modern capabilities in terms of UI/UX design.

IV. Interactive Exploration System

We conceive the *Advanced Manuscript Analysis Portal* (AMAP) that aims to implement such a modern and intuitive system with a domain-inspired interaction paradigm. It aims to answer the research question *How will scholars in humanities work with digitized manuscripts in the future?* The main motive behind the portal is to allow them to freely explore digital manuscripts with a sense of familiarity and intuitiveness of working with physical artifacts.

A. Objectives

1) Accessible Image Processing: Application of advanced image processing methods currently requires an understanding of various algorithms and the associated parameter regime and, hence, is not very accessible to

users without a computing background. We aim to make it more accessible by allowing users to create complicated processing pipelines through a custom flow-based visual language and get their desired results through iterative experimentation.

2) Extensive Visualization: The system will be built with extensive visualization capabilities that will aid researchers in their day-to-day activities. For instance, appropriately visualizing how the characters differ in their respective feature-space (such as SIFT) can help users in tasks such as writer identification and dating.

3) Enabling Exploration: The system will be equipped with relevant toolsets that will help researchers to formulate new hypotheses through ad-hoc exploration and non-linear workflows in accordance with the user-in-the-loop paradigm. This will also include relevant interfaces to perform tasks like reconstructing ductus and measuring geometric attributes.

4) Integrating Metadata: Apart from the images themselves, information about dates, authors, location, hand descriptors etc. form an integral part of digitized manuscripts. The portal will aim to integrate related metadata into manuscript analyses. This will take forms such as geotagging, sorting and appropriate search interfaces.

5) Seamless Workflow: Considering the difficulty of using ill-fitted multiple tools for various tasks, the portal will strive to provide a seamless workflow through effortless task switching and interplay of data within those tasks. This will involve the ability to annotate, perform layout analysis, manipulate images, create syllabaries and other relevant tasks from the same workspace based on an in-depth requirement analysis.

B. Design Considerations

1) Interaction Paradigm: To implement an interaction paradigm that will reflect the domain of the researchers, we will use a reality-based interaction framework [7] that draws from real-world aspects of the researchers' environment and coupling it with a domain-driven design. We will explore the possibility of using physical-manuscript-based metaphor as a mode of interaction with the manuscript images. This will involve representing/visualizing the individual images as virtual palm-leaves/paper folios to simulate realistic interaction with the objects in a virtual deskspace, enabling actions on manuscripts like rearranging sheets, piling and crumbling/rolling. We will design appropriate multi-touch gestures for interaction and integrate the use of a stylus in a way that will reflect the chosen paradigm. The goal is to model the interaction to create a sense of familiarity through appropriate metaphors relevant to the domain, hence provide researchers with an intuitive system that is easy to understand, use and explore. The system will need to seamlessly blend into their daily workflow without being disruptive.

2) Interface Design: The interface needs to take full advantage of the physical space available in large devices, particularly in terms of actions such as comparing manuscripts/characters in full resolution, image manipulation and annotation tasks. Given the manuscript images will be undergoing several transformations, they will be presented in a layered manner allowing analysis of the manuscripts from multiple perspectives through comparisons and superimpositions. The interface will have a special focus on real-time collaboration with colleagues. We will design an adaptive user interface that will respond to change in form-factors, orientations and capabilities, while also attempting to personalize the layout by continuously learning through user behavior.

V. Current Work

We have developed a demonstrator with a client-server architecture. The backend is an asynchronous server daemon to perform resource-heavy image operations, while the frontend webpage allows users to create a simple image processing pipeline and interact with images through a rudimentary touch interface. We further plan to integrate tools developed by the associated Z03 project[1] within SFB 950 into our system. We plan to develop a working open-source prototype and deploy it at the *Centre for the Study of Manuscript Cultures* (CSMC), University of Hamburg and perform a thorough user evaluation of the system.

VI. Conclusion

We have presented AMAP, a novel work in progress for the interactive exploration of digitized manuscripts. We have discussed the disadvantages of the current systems and have proposed a system that works on an intuitive interaction paradigm and provides an integrated tool to explore manuscripts. Though the work is in its early stages, it will demonstrate how researchers will interact with manuscripts in a post-WIMP future.

References

[1] R. Kirkham, "Building a virtual research environment for the humanities JISC final report," University of Oxford, Tech. Rep., 2007.

[2] A. K. Bowman, C. V. Crowther, R. Kirkham, and J. Pybus, "A virtual research environment for the study of documents and manuscripts," *The Oxford e-Research Conference*, 2010.

[3] T. A. Griffitts and U. B. Schmid, "The virtual manuscript room collaborative research environment," *AIUCD 2016*, p. 61, 2016.

[4] P. A. Stokes, "Digital resource and database for palaeography, manuscripts and diplomatie," *Gazette du livre médiéval*, vol. 56, no. 1, pp. 141–142, 2011.

[5] P. Štorková and J. Kysela, "Tablet as a new interactive tool for education paleography," *Procedia-Social and Behavioral Sciences*, vol. 174, pp. 3164–3169, 2015.

[6] D. Rafiyenko, "Tracing: A graphical-digital method for restoring damaged manuscripts," in *Kodikologie und Paläographie im Digitalen Zeitalter 4*, H. Busch, F. Fischer, and P. Sahle, Eds., 2017, pp. 121–135.

[7] R. J. Jacob, A. Girouard, L. M. Hirshfield, M. S. Horn, O. Shaer, E. T. Solovey, and J. Zigelbaum, "Reality-based interaction: a framework for post-wimp interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2008, pp. 201–210.

[1] https://www.manuscript-cultures.uni-hamburg.de/Projekte_p2_e.html#PBZ

An Efficient Approach for Designing Deep Learning Network on Title Block Extraction for Architecture, Engineering & Construction Documents

Shubham Gupta, Jayanta Mukherjee, Dipali Bhattacharya, Himadri Majumder, Rahul Roy
ARC Document Solutions, Kolkata, India
Email: shubham.gupta@e-arc.com, jayanta.mukherjee@e-arc.com, dipali.bhattacharya@e-arc.com, himadri.majumder@e-arc.com, rahul.roy@e-arc.com

Bidyut B. Chaudhuri
CVPR Unit
Indian Statistical Institute
Kolkata, India
Email: bbcisical@gmail.com

***Abstract*—Till now Architectural, Engineering and Construction industry heavily relies on the paper documents and their management. Digitization of such documents before technical advances was a time intensive manual process. For design documents, the Title block is a unique identifier. This identifier facilitates features like search and auto hyperlinking of documents. In this paper, automation of title extraction has been achieved through OCR (Optical Character Recognition) on the document image. This process involves multiple steps like scanning of documents, automatic title block selection, applying OCR on the block and extraction of relevant information. However, still the user must review all the extracted information to get error-free data. The cumulative results on all steps add to a considerable error rate because of practical challenges. We hereby propose a hybrid system capable of learning in real time using concepts of machine/deep learning applied on the user specific data. This approach has substantially reduced the user intervention, thus improving user experience and efficiency of working on the platform. The applied learning algorithm is based on a simple feed-forward model of neural network designed to take care of the common flaws in the implementation of deep learning that requires large datasets and computational power for training.**

Keywords- Auto Hyperlinking, Deep learning, Engineering drawing management, Hybrid System, Machine Learning, Neural Network, Recommender.

I. INTRODUCTION

For Architectural design documents, OCR is used to automatically extract certain metadata from the uploaded documents. The extracted contents must be reviewed by the user in order to get the 100% accurate data. This leads to a hectic process of reviewing each document metadata. So, we hereby propose a user based customized model which can reduce the number of documents for review and thus enhance the user experience.

The proposed model is further improved to incorporate user specific historical data so that it recognizes individual user pattern and thus provide a more customized reviewing process for each user. This very well synergizes with automatic document hyper-linking and thus improves the efficiency of the entire process [1][2].

II. DATA ANALYSIS

Collected data is categorized in correctable/Non-correctable cases.

1. Incorrect Orientation (Fig 1) and Surplus text (Fig 2)

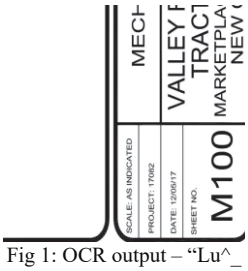


Fig 1: OCR output – "Lu^_ "& User input – "M100-7D"

Fig 2: OCR output – "11-154" & User Input – "EOS 01"

2. User specific Edits /Tagging

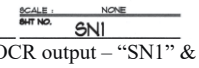


Fig 1: OCR output – "SN1" & User input – "Plan 5 Lot 4 - SN1"

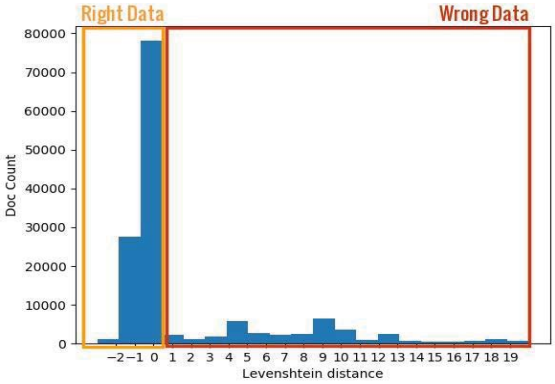


Fig 2: Histogram for data-set points with levenshtein distance

Score -1, -2 are categories where OCR value is correct, but user added or deleted the content respectively.

III. Detector Model

Applied model is a basic artificial neural network with 3 hidden fully connected layers. The input for the model is the feed from traditional OCR System. This feed is encoded using a variation of **one-hot encoding**. The target for the model is to predict the class of the encoded feed based on **Levenshtein** distance between the OCR feed and the corresponding user input.

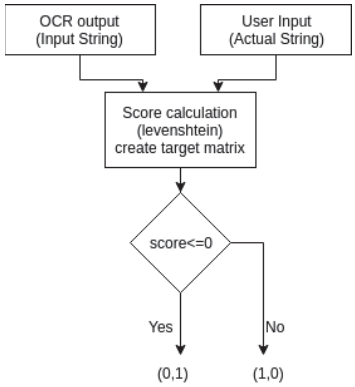


Fig 3: Target Calculation

This target calculation can be optimized using threshold value for a score. For our case we have considered the ideal case.

Model is trained as shown in the flow. The iteration is repeated for all the training data multiple times till the cost stagnates.

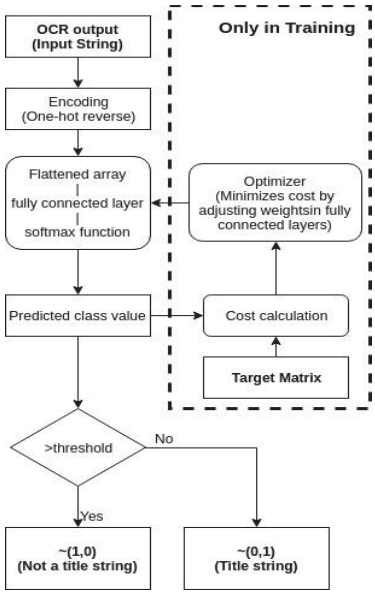


Fig 4: Training Phase

Rule based model was also developed with the following:

1. 10 > Length > 2
2. Should not start with any special character.
3. Should not include any common special character.

On testing the above model, accuracy was 60.8% while accuracy of the Deep learning model was 85%.

IV. Results

Below test results depicts difference between DL (deep learning model) vs rule-based model.

Case 1: Valid AEC Title block

Test string	Correctness from DL model	Correctness from rule based model
PSP L-4.2	100%	0
IMP 7 OF 10	99%	0
09003 - A4-39	100%	0

Case 2: Invalid AEC Title block

Test string	Correctness from model	Correctness from rule based model
T1CHAVEN	0%	100%
TRB.1	15%	100%
21-93	0%	100%

The deep learning model was able to incorporate different patterns in the title while the rule-based model by nature focuses on a very narrow domain of patterns in title block. In addition, false negative cases increased by 120%.

V. User Based Customization

On observing user dataset, we found that user batch edits follow a pattern.

A2.12	A2.12_Bid Thru Add C.pdf
A3.00	A3.00_Bid Thru Add C.pdf
A2.05	A2.05_Bid Thru Add C.pdf
A2.07	A2.07_Bid Thru Add C.pdf
A2.03	A2.03_Bid Thru Add C.pdf
A2.02	A2.02_Bid Thru Add C.pdf

Fig 5: User edits for similar documents

We can use same model to collect all the entries with similar score to group and recommend changes to users. This reduces the reviewing and editing time considerably, thus increases efficiency.

References

[1] P. Banerjee, S. Choudhary, S. Das, H. Majumdar, R. Roy and B. B. Chaudhuri, *"Automatic Hyperlinking of Engineering Drawing Documents,"* 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, 2016, pp. 102-107. doi: 10.1109/DAS.2016.76

[2] P. Banerjee, S. Choudhary, S. Das, H. Majumdar, S. Mukkamala, R. Roy and B. B. Chaudhuri, *"A System for Automatic Navigation in "Architectural and Construction Documents,"* in 14th IAPR International Conference on Document Analysis and Recognition" (ICDAR), November 2017 (accepted).

[3] 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, Pittsburgh, U.S.A., pp.88-100, (1996)

[4] Y Bassil and M Alwani, *"OCR POST-PROCESSING ERROR CORRECTION ALGORITHM USING GOOGLE'S ONLINE SPELLING SUGGESTION", Journal of Emerging Trends in Computing and Information Sciences, Beirut Lebanon, pp.1-4, (2012)*

A High-Performance Document Image Layout Analysis for Invoices

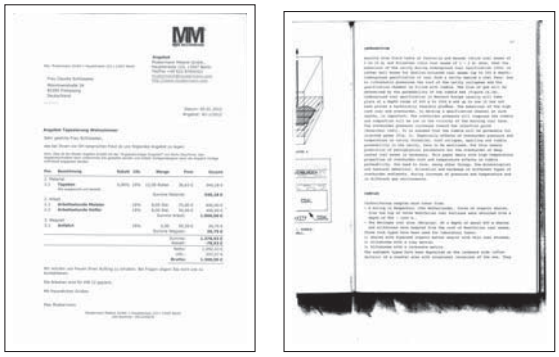
Mohammad Mohsin Reza*, Md. Ajraf Rakib*, Syed Saqib Bukhari, Andreas Dengel
DFKI and University of Kaiserslautern, Germany
{mohammad_mohsin.reza, md_ajraf.rakib, saqib.bukhari, andreas.dengel}@dfki.de

***Abstract*—Layout analysis for document is an important step in OCR pipeline and currently an intensive amount of research is going on to extract searchable full text from scanned images. Invoices are different in nature as compared to pages of books, magazine, loan documents and others, since, there are tables, header, footer, large white spaces, currency, item name, item amount, logo in the invoice. The standard layout analysis proves inefficient on invoices. In this paper we are proposing an advanced layout analysis for invoices that integrate the following steps in the standard layout analysis: removal of table cell lines and merging text lines. Additionally, we integrated the proposed layout analysis for invoices into the anyOCR system, which was mainly developed for both historical as well as contemporary documents from books, magazines etc. In the performance evaluation section, we will compare our advanced layout analysis pipeline with the standard anyOCR [1] pipeline and with a commercial OCR system like ABBYY. Our advanced layout analysis achieved better OCR accuracy as compared to the other mentioned systems.**

I. Introduction

There has been a resurgence of interest in optical character recognition (OCR) in recent years mainly for digitizing document to increase re-usability of information. Automatic data processing plays a vital role in processing lots of documents making our daily life not only easier but also get more benefit from the computerize system. A digital mailroom system is the automation of incoming mail (for example, scanned forms and invoices and digital emails) processes, where structured forms processing is relatively an easier task as compared semi-structured invoices. There are roughly two main tasks to process data from invoices: OCR and Information extraction. While performing end-to-end OCR pipeline for invoices, layout analysis in a most challenging task because of tables, header, footer, large white spaces, currency, item name and amount and logo, which are not commonly present in standard pages form books and magazines. These differences can be seen in Figure 1. Therefore, standard document layout analysis gives inefficient result on invoice as shown in evaluation.

In literature, there are some papers which proposed methods for document layout analysis. Bukhari et al. [2] proposed a layout analysis system for extracting Arabic text-lines from scanned documents written in different languages and styles. They presented the system based on a suitable combination of different well established techniques for analyzing Latin script documents that have proven to be robust against different types of document image degradations. In another paper, Bayer et al. [3] introduced a system that used an OCR tool with FRESCO model to extract particular information from invoices but they did not mention anything about the accuracies of that OCR tool as well as no detail about layout analysis techniques. Tuganbaev et al. [4] used a top-down document analysis structure with the FlaxiCapture technology to capture particular information from invoice. They performed a full-page OCR before applied their technology but also did not focus on the OCR accuracy. Though, the accuracy of the OCR tool is also important for extracting particular information so we were looking for some related paper that talks about some OCR method and its accuracy. But we did not find any particular paper that gives the better explanation for layout analysis of invoices.



(a) (b)

Fig. 1. (a) A sample scanned invoice, (b) A sample scanned standard (book page) document.

In this paper, we introduced an advanced high-performance layout analysis for invoices where we have integrated new methods for different tasks in the layout analysis pipeline of the anyOCR system [1], which is developed for processing pages from historical and contemporary books or magazines. For this purpose, we used a line removal method to remove line-graphics from the table and combined text lines so that information in the table are kept intact row by row. The rest of the paper is organized as follows. After discussing our proposed method in Section II for advance layout analysis of invoices followed by performance evaluation in Section III to compare our result with other systems. Finally in Section IV we conclude our work.

*These two authors contributed equally.

II. A High-Performance Layout Analysis for Invoices:

In order to understand our contribution in this paper, at first we will briefly describe the existing state of the anyOCR system and then we will present our contribution in its layout analysis pipeline.

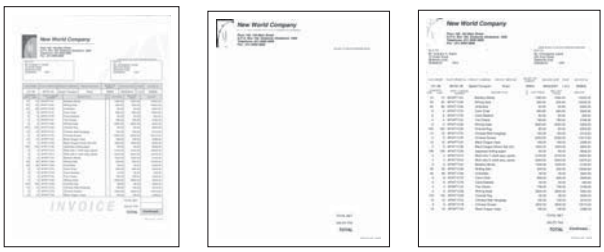
A. The anyOCR System - Overview [1]

The anyOCR component contains a set of document analysis methods that are usually required for a typical end-to-end OCR pipeline for extracting text from a document image. This method includes binarization, text and non-text segmentation, text line extraction, and producing OCR text in hOCR format, where text non-text segmentation and text line segmentation include as a layout analysis pipeline.

B. The Proposed anyOCR System for Invoices

We made some change in existing pipelines for layout analysis of invoices in the anyOCR system [1]. Usually, invoices contain table and most of the table draw with line-graphics that may recognize as a non-text part by existing pipeline which fails to extract text data from invoices. This is one of the first barriers for extracting data from invoices.

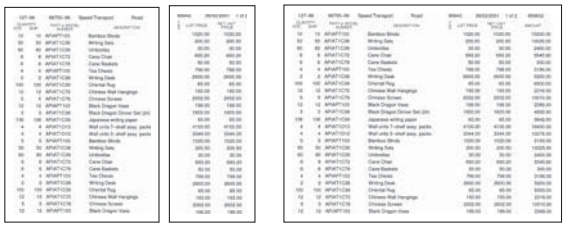
Firstly, we removed all line-graphics from invoices before applying binarization method. The results of an input image after processed by binarization and text non-text segmentation steps without and with our proposed line removal step are shown in Figure 2. One can see the results in this figure with line removal pipeline keep all the text in the image as compared to the existing pipeline.



(a) (b) (c)

Fig. 2. (a) A simple scanned invoice, (b) processing steps: binarization, text non-text segmentation, page frame segmentation, (c) processing steps: a newly introduced line-graphics removal, binarization, text non-text segmentation, page frame segmentation.

The existing text line segmentation method over-segments the columns of a table which may be difficult to synchronize with related data for each item in the table. We would like to intact them in a single line so that item information is extracted row by row which is shown in the Figure 3. Secondly, in order to achieve it, we changed the existing text line extraction method for invoice and merge the text lines segments with each other which are in the same row. Finally, text is recognized for each line in reading order and then saved in the hOCR format.



(a) (b)

Fig. 3. (a) Oversegmentation a table by standard anyOCR pipeline during page segmentation, (b) Our proposed method to keep intact row-wise information.

III. Performance Evaluation

There is no public dataset available for invoices. It is usually a laborious task to create datasets for performance evaluation. However, we have created a dataset of 29 images which we will partly share in public (because around 10 invoices are proprietary). For ground truth text creation, we considered all text entries at same height as a single text line. We compared the final OCR accuracy of the following systems: The standard anyOCR, ABBYY, and the proposed high-performance anyOCR system for invoices. The result are shown in Table 1, where our proposed system achieved the best performance as compared to the other two systems. ABBYY performed well on most of the images. However, for some cases, like Fig 2(a), it completely missed the whole table.

OCR System	**Accuracy**
The anyOCR system	53.95%
ABBYY	76.00%
The Proposed anyOCR system for Invoices	83.34%

TABLE I
The table Showing the OCR accuracy over three different pipelines including proposed one.

IV. Conclusion

The anyOCR is an open-source system which gives very good accuracy for standard document images such as pages from books, magazines and so on. Invoices are naturally different from standard document images because they contain tables, headers, footers. In this paper we integrated additional steps in the existing layout analysis pipeline of the anyOCR system to achieve a high-performance layout analysis for invoices. The proposed system achieved the best performance as compared to not only the existing anyOCR system but also the commercial system ABBYY.

References

[1] S. S. Bukhari, A. Kadi, J. M. Ayman, and A. Dengel, "anyocr: An open-source ocr system for historical archives," in *ICDAR*, 2017.

[2] S. S. Bukhari, F. Shafait, and T. M. Breuel, "High performance layout analysis of arabic and urdu document images," in *ICDAR*, 2011.

[3] T. Bayer and H. U. Mogg-Schneider, "A generic system for processing invoices," in *ICDAR*, 1997.

[4] D. Tuganbaev, A. Pakhchanian, and D. Deryagin, "Universal data capture technology from semi-structured forms," in *ICDAR*, 2005.

Applying Sequence-to-Mask Models for Information Extraction from Invoices

Anoop R Katti, Johannes Hoehne, Steffen Bickel, Jean Baptiste Faddoul
SAP SE
anoop.raveendra.katti@sap.com, johannes.hoehne@sap.com, steffen.bickel@sap.com, jean.baptiste.faddoul@sap.com

***Abstract*—Automatically extracting information from scanned invoices can result in significant time and cost savings in Accounts Payable processing. We present an RNN model for extracting information from scanned invoices. Our model significantly outperforms a Random Forest baseline.**

***Index Terms*—RNN; GRU; bidirectional GRU; character mask**

I. Introduction

Extracting structured information from invoices is a valuable problem. Some examples of structured information are invoice number, amount, date, buyer name and address, vendor name and address, remittance address, items purchased etc. Manually extracting structured information from invoices can be slow, tedious, and error prone. Therefore, automating this process is vital.

Automatic extraction of structured information from invoices can be a challenging task. This is mainly due to the large diversity in the format of the invoices, languages, currencies, taxation rules, and country specific characteristics. In this work, we propose a Recurrent Neural Networks (RNN) based sequence labeling model to extract interesting information from the invoice. We evaluate it on a real world dataset of invoices. We compare it against a Random Forest baseline model and show that the proposed model is superior in terms of accuracy.

II. Related Work

In the recent years, there has been a lot of work on applying deep networks on natural language. RNN have achieved state-of-the-art performance on a number of NLP tasks such as Named Entity Recognition, Translation etc. However, there has been very little work on applying RNNs on invoice data.

Stark et al. [3] formulate invoice extraction as a sequence-to-sequence problem [1] and use encoder-decoder RNN to *translate* the input OCR to output values. However, this results in generation of text outside of the input sequence.

Cloudscan [2], perhaps the closest to our work, formulate information extraction from invoices as sequence labeling where a sequence of word n-grams are labeled as belonging to a field of interest or not. Additionally, word-specific features are employed to boost accuracy. However, invoices often contain words outside of a finite vocabulary. This could be due to proper-names and abbreviations but also due to character flips and unnecessary spaces caused by OCR errors. Contrary to Cloudscan, we operate at the character-level. This allows us to employ simple 1-hot encoding on each character. Not only does this eliminate the need for feature engineering, but it also considerably reduces the length of the feature vector.

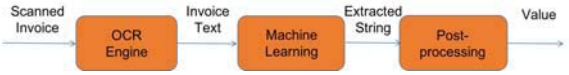


Fig. 1: Processing Pipeline

III. Baseline

The first step in our invoice processing pipeline (see Figure 1) is to extract text from the scanned invoices, by passing the document through an OCR Engine. As a baseline method, we trained a random forest classifier that predicts for each word on the invoice, whether or not it belongs to a certain field. More concretely, each word was encoded as a feature vector that included (A) the target word and its character statistics, (B) the x/y position of the target word (C) the context, i.e. the text above the target word (D) the context left of the target word. A random forest classifier was trained that classified the word as either belonging to the field of interest or not. During inference, the word with highest probability was selected.

IV. Sequence-To-Mask

The text extracted from OCR can be considered as a sequence of input tokens. This makes RNNs an ideal candidate for processing the invoice text.

Extracting structured information from invoices corresponds to extracting fields such as invoice number, amount, date etc. We observe that these values are part of the invoice text inputted to the network. Therefore, given the invoice text as input, we can extract each value from the text by ”highlighting” parts of the input text. This can be formalized by with a binary character-mask that describes for each character, whether or not it contains the target information.

Therefore, we tokenize the input text as characters. To be able to process long character sequences, we use a type of RNN that can retain information over very long sequences called Gated Recurrent Units (GRUs). Each character is fed into the GRU as a one-hot encoded vector. The GRUs are arranged bi-directionally to enable reading in both forward and backward directions. We also stack the GRUs to increase the capacity of the network. Finally, we have a dense layer that predicts a binary mask for each time step at the output. Figure 2 (left) illustrates the network architecture. The binary

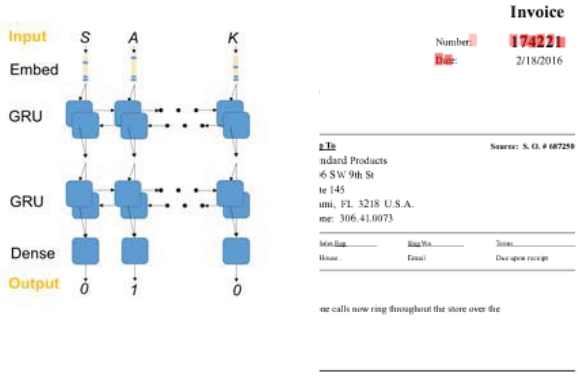


Fig. 2: (a) Network (b) Example prediction

mask at time step t gives the confidence of the character at t belonging to the field being predicted. Figure 2 (right) visualizes confidences over an example invoice as heatmap.

After prediction a post-processing step is executed that normalizes and parses values. For example, amounts are parsed as float and dates are parsed with a standard date parser. The complete pipeline is show in Figure 1.

V. Experiments

In this section, we present the training/test data, the evaluation metric, and the results of our experiments.

In order to train and evaluate our model, we collected a dataset of 12000 invoices. We annotated the bounding boxes on the invoice for each field we aim to extract. We used 10000 invoices for training, 1000 for validation and 1000 for testing.

The model was trained using Stochastic Gradient Descent with adaptive gradient. In order to facilitate efficient training with batching, all invoices are padded or truncated to a fixed length of 1500 characters. In our experiments, we found that at inference time prediction on the complete sequence is not harmed by truncating the sequences at training data. The model is trained for 30 epochs. The time taken to train the model for one field in tensorflow is about 7 hours on a single Tesla M40 GPU.

For our evaluation, we choose to remain as close to the business use-case as possible and therefore compare the final formatted string with the ground-truth value specified by a human expert. For invoice number and purchase order number, we note that only alpha-numeric characters are relevant and therefore, at the time of comparison, ignore all other characters. We compute the accuracy of prediction as the percentage of invoices for which the extracted value is equal to the ground-truth value barring the irrelevant characters as mentioned above. We refer to this as the business metric.

A. Result

Table 1 presents the accuracy numbers obtained according the business metric (described in the previous section). It can be seen that seq2mask performs much better than a random forest based classifier on all fields.

Fields	Random Forest (%)	seq2mask (%)
Number	50.40	74.4
Date	50.80	80.2
Amount	66.40	85.9
Purchase Order No	33,40	68.7

TABLE I: Comparison of extraction results

The proposed model performs best on Invoice Amount. This is perhaps due to the discriminative keyword 'Total' (or its equivalent in different languages) as well as the discriminative location – usually bottom right of the invoice.

For Invoice Number, we found that, while the keyword 'Invoice No.' usually precedes the value, in a number of cases, the value may be preceded by just a 'hash' or nothing at all. This explains a slightly lower accuracy for this field.

We also found that it is important to pick the right threshold for detecting a prediction. This is especially applicable to fields such as Purchase Order Number and Primary Tax Amount which are not always mentioned on the invoice.

VI. Conclusion

We tackle the problem of extracting structured information from invoices. This corresponds to extracting values such as invoice number, date, amount, etc. from invoices. We formulate this as a sequence-to-mask problem and model it with RNNs. Through experiments, we demonstrate the validity of the model and report accuracy numbers that significantly improve over the Random Forest baseline.

A major disadvantage of this method is that the training data cannot be scaled as this requires manual box annotation. This could be approached by directly predicting the desired output text instead of highlighting the given input text. Also, currently the document layout is completely ignored. Incorporating the document layout information in the model could potentially increase the accuracy of the model by a large margin.

References

[1] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014. APA

[2] Palm et. al. 2017. CloudScan - A configuration-free invoice analysis system using recurrent neural networks. https://arxiv.org/pdf/1708.07403.pdf

[3] Stark et. al. 2017. How we use machine learning to interpret bad cell phone images of travel receipts. https://blog.altoros.com/optical-character-recognition-using-one-shot-learning-rnn-and-tensorflow.html

Re-OCR in Action – Using Tesseract to Re-OCR Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals

Kimmo Kettunen
The National Library of Finland
DH projects
Mikkeli, Finland
firstname.lastname@helsinki.fi

Mika Koistinen
The National Library of Finland
DH projects
Mikkeli, Finland
firstname.lastname@helsinki.fi

Abstract— This paper presents work that has been carried out in the National Library of Finland to improve optical character recognition (OCR) quality of a Finnish historical newspaper and journal collection 1771–1929. Work and results reported in the paper are based on a 500 000 word ground truth (GT) sample of the Finnish language part of the whole collection. The sample has three different parallel parts: a manually corrected ground truth version, original OCR with ABBYY FineReader v. 7 or v. 8, and an ABBYY FineReader v. 11 re-OCRed version. Based on this sample and its page image originals we have developed a re-OCRing procedure using the open source software package Tesseract1 v. 3.04.01. Our methods in the re-OCR include image preprocessing techniques, usage of a morphological analyzer and a set of weighting rules for resulting words. Besides results based on the GT sample we present also results of re-OCR for a 10 year period of one newspaper of our collection, Uusi Suometar.

OCR; historical newspapers; Tesseract; Finnish

I. INTRODUCTION

The National Library of Finland has digitized historical newspapers and journals published in Finland between 1771 and 1929 and provides them online [1-2]. The 1920s part of the open collection, 1921–1929, was released in January 2018. The collection contains approximately 7.36 million freely available pages primarily in Finnish and Swedish. The total amount of pages on the web is over 12.8 million, part of them being in restricted use due to copyright reasons. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as *Digi*. An open data package of the collection's newspapers and journals from period 1771 to 1910 has been released in early 2017, years 1911–1920 will be released later [2].

When originally non-digital materials, e.g. old newspapers and books, are digitized, the process involves first scanning of the documents which results in image files. Out of the image files one needs to sort out texts and possible non-textual data, such as photographs and other pictorial representations. Texts are recognized from the scanned pages with Optical Character Recognition (OCR) software. OCRing for modern prints and font types yields usually high quality results, but results of historical document OCRing are still far from that [3].

Newspapers of the 19th and early 20th century were mostly printed in the Gothic (Fraktur, blackletter) typeface in Europe. Fraktur is used heavily in our data, although also Antiqua is common and both fonts can be used in same publication in different parts. It is well known that the Fraktur typeface is especially difficult to recognize for OCR software. Other aspects that affect the quality of OCR recognition are the following [3–4]:

- quality of the original source and microfilm
- scanning resolution and file format
- layout of the page
- OCR engine training
- unknown fonts
- etc.

Due to these difficulties scanned and OCRed document collections have a varying amount of errors in their content. A quite typical example is *The 19th Century Newspaper Project* of the British Library [5]: based on a 1% double keyed sample of the whole collection Tanner et al. report that 78% of the words in the collection are correct. This quality is not good, but quite realistic.

Ways to improve quality of OCRed texts are few, if total rescanning is out of question, as it usually is due to labor costs. Improvement can be achieved with three principal methods: manual correction with different aids (e.g. editing software), re-OCRing or algorithmic post-correction [3]. These methods can also be mixed.

Due to amount of data we have chosen re-OCRing with Tesseract v. 3.04.01 as our main method for improving the quality of our collection. In the rest of the paper we describe the results we have achieved so far.

II. RESULTS

Our re-OCR process has been described more thoroughly in [6–7]. Here we describe it only briefly. The re-OCRing process consists of four parts: 1) image preprocessing of page images using five different techniques, 2) Tesseract OCR 3.04.01, 3) choosing of the best candidate from Tesseract's output and 4) transformation of Tesseract's output to ALTO format. We have developed a new Finnish Fraktur model for Tesseract using an existing German Fraktur model as a starting point.

We have evaluated the results of the re-OCR so far with different methods using our ground truth data of 471 903 words. This parallel data consists of proof read version of the data, current OCR, Tesseract 3.04.01 OCR and ABBYY FineReader v.11 OCR. We performed detailed quality analyses for the results using different ways of evaluation. Kettunen and Pääkkönen [1] have earlier estimated the quality of the whole historical

[1] https://github.com/tesseract-ocr

collection of Finnish with automatic morphological analysis. We applied this quality approximation method now with two morphological analyzers: Omorfi v. 0.3[2] and HisOmorfi, a modified version of Omorfi. Results of analyses are shown in Table 1.

TABLE I. RECOGNITION RATES FOR DIFFERENT COMPARABLE DATA: 471 903 WORDS

	GT	Tesseract 3.04.01	Current OCR	ABBYY FineReader v.11
Omorfi 0.3	81.3%	78.3%	77.1%	85.3 %
HisOmorfi	94.9%	89.9%	81.0%	86.0%

Figures show that the manually edited ground truth version is recognized clearly best, as it should be. Plain Omorfi recognizes Tesseract words slightly better than the words of current OCR, the difference being 1.2% units. The seemingly small difference is caused by the fact that HisOmorfi is used in the re-OCRing process to choose words from output of Tesseract and it favors *w* to *v*. Plain Omorfi does not recognize most of the words that include *w,* but HisOmorfi is able to recognize them, which is shown in the high percentage of Tesseract's HisOmorfi result column. Difference in recognition between current OCR and Tesseract is 8.9% units with HisOmorfi..

When we applied standard measures of recall, precision and F-score to the data, we got recall of 0.72, precision of 0.73 and F-score of 0.73. Combined optimal OCR results of Tesseract and ABBYY FineReader v. 11 would give recall of 0.81, precision of 0.95, and F-score of 0.88**.** The latter figures show that possibility of using several OCR engines would benefit re-OCRing, as has been stated in research literature [3, 8]. Unfortunately we do not have access to several OCR engines in our final re-OCR.

After initial development and evaluation of the re-OCR process with the GT data, we have started final testing of the re-OCR process with realistic newspaper data. We chose for testing *Uusi Suometar*, newspaper which appeared in 1869–1918 and has 86 068 pages. Table 2. shows results of a 10 years' re-OCR of Uusi Suometar.

TABLE II. RECOGNITION RATES OF CURRENT AND NEW OCR WORDS OF *UUSI SUOMETAR* WITH MORPHOLOGICAL ANALYZER HISOMORFI (TOTAL OF 7 937 PAGES)

Year	Words	Current OCR	Tesseract 3.04.01	Gain in % units
1869	658 685	69.6%	86.7%	17.1
1870	655 772	66.9%	84.9%	18.0
1871	909 555	73%	87%	14.0
1872	930 493	76%	88.7%	12.7
1873	889 725	75.4%	87.3%	11.9
1874	920 307	72.9%	85.9%	13.0
1875	1 070 806	71.5%	86%	14.5
1876	1 223 455	72.8%	86.7%	13.9
1877	1 815 635	73.9%	86%	12.1
1878	2 135 411	72%	85.4%	13.4
1879	2 238 412	74.7%	87%	12.3
ALL	13 448 256	73%	86.5%	13.5

As can be seen from the figures, re-OCR is improving the recognition rates considerably and consistently. Minimum improvement is 11.9% units, maximum 18% units. In average the improvement is 13.5% units.

[2] https://github.com/flammie/omorfi

III. CONCLUSION

We have described in this paper results of a re-OCRing process for a historical Finnish newspaper and journal collection. The process consists of combination of five different image pre-processing techniques, a new Finnish Fraktur model for Tesseract OCR enhanced with morphological recognition and rules to weight the result words. Out of the results we create new OCRed data in METS and ALTO XML format that can be used in our docWorks document system.

We have shown that the re-OCRing process yields clearly better results than commercial OCR engine ABBYY FineReader v. 7/8 and v. 11 with our GT data. We have also shown that a 10 year time span of newspaper Uusi Suometar (7937 pages and ca. 13.45 M words) gets significantly and consistently improved word recognition rates for Tesseract output in comparison to current OCR.

We shall continue the re-OCR process by re-OCRing first the whole history of *Uusi Suometar*. Its 86 000 pages should give us enough experience so that after that we can move over to re-OCRing the whole Finnish collection. The GT package data is available on our open data web site https://digi.kansalliskirjasto.fi/opendata.

ACKNOWLEDGMENT

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020.

REFERENCES

[1] K. Kettunen and T. Pääkkönen, "Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means," Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).

[2] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use," D-Lib Magazine, July/August 2016.

[3] M. Piotrowski, Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2012.

[4] R. Holley, "How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," D-Lib Magazine, 15(3/4) 2009 .

[5] S. Tanner, T. Muñoz, and P.H. Ros, "Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive," D-Lib Magazine, (15/8) 2009.

[6] M. Koistinen, K. Kettunen, and J. Kervinen, "How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine," Proc. of LTC 2017, Nov. 2017, pp. 279–283.

[7] M. Koistinen, K. Kettunen, and T. Pääkkönen, "Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing," Proc. of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, May 2017, pp. 277–283.

[8] M. Volk, L. Furrer, and R. Sennrich, "Strategies for reducing and correcting OCR errors," in C. Sporleder, A. van den Bosch, and K. Zervanou, Eds. Language Technology for Cultural Heritage, 2011, pp.3–22.

High-Accuracy Japanese Scene Character Recognition Using Synthetic Scene Characters and Multi-Scale Voting Classifier

Fuma Horie
School of Engineering, Tohoku University
Sendai, Japan
fuma@sc.cc.tohoku.ac.jp

Hideaki Goto
Cyberscience Center, Tohoku University
Sendai, Japan
hgot@cc.tohoku.ac.jp

***Abstract*—Scene character recognition is challenging owing to various factors such as rotation, geometric distortion, uncontrolled lighting, blur, and noise. In addition, Japanese character recognition requires a large number of training data since thousands of character classes exist in the language. In order to enhance Japanese scene character recognition, we have developed a training data augmentation method and a recognition system using multi-scale classifiers. Experimental results show that the multi-scale scheme effectively improves the recognition accuracy.**

Keywords-character recognition; Japanese scene character recognition; synthetic scene data; ensemble voting classifier; multi-scale analysis; Support Vector Machine

I. Introduction

Scene character recognition is challengin owing to some environmental factors such as rotation, geometric distortion, and uncontrolled lighting, and to some blur and noise at image capturing. Japanese scene character recognition requires a lot of character data for training since thousands of character classes exist. Some researchers proposed scene character recognition method using synthetic scene character data (SSD) [1][2]. In our previous work[3], we developed a training data augmentation method and a scene Chinese character recognition method based on the ensemble learning strategy to improve the recognition accuracy.

In this paper, we present a training data augmentation technique using Random Filter and Multi-Scale Voting Classifier for scene character recognition. Experiments using a set of real scene character data confirm the effectiveness of the developed system.

II. Japanese Scene Character Recognition Using Synthetic Scene Character Data

A. SSD Generation Using Random Filter

Scene characters may suffer from some blur, shadow, and noise when they are captured by a camera. Some decorated characters may have enhanced edges and artificial shadows. In this work, we focus on these factors rather than shape variation as character recognition methods are generally designed to be tolerant of some deformation. We generate a 3×3 or 5×5 kernel with uniformly-random numbers. Each element can be positive or negative, and it is normalized so the sum becomes 1.0. Figure 1 shows some images obtained by different kernels.

B. Multi-Scale Resizing

When scene character images are taken by cameras or smart phones, they usually appear in arbitrary sizes after cropping. In [3], Multi-Scale Resizing (MSR) was introduced to design the ensemble learning-based classifier. Two sizes, 32×32 and 64×64, were used. We have added one more scale 16×16 as we expected further accuracy improvements.

Figure 2 shows the schematic diagram of the Multi-Scale Voting Classifier. As for the base classifier, we compare the following configurations:

- NNS: Linear Nearest Neighbor Search
- SVM: Support Vector Machine
- RF: Random Forest [3]
- NNS+SVM: Use both NNS classifiers and SVM classifiers in the voting.

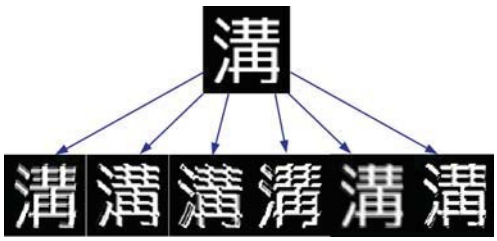


Figure 1. SSD generated using Random Filter.

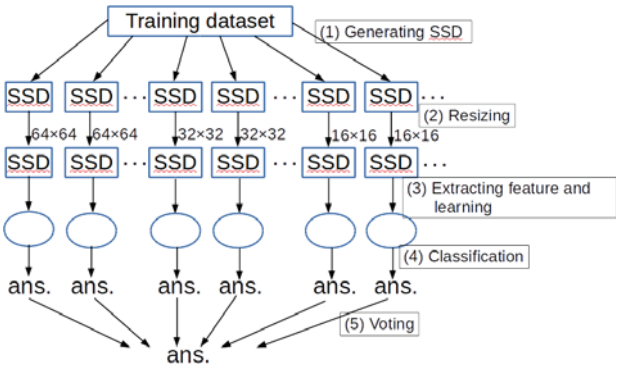


Figure 2. Multi-scale Voting Classifier.

III. Performance Evaluation And Comparison

A. Experiment environment

We compiled a dataset of Japanese scene characters for testing. It consists of Hiragana, Katakana and Kanji (796 images and 373 classes) taken in real scenes. All character images are in color. Regular and bold M+ fonts (3,107 classes, total 6,214 characters) are used for training. The training dataset does not include real scene characters since it is difficult to collect characters of all classes in Japanese. With regard to a feature extraction, we employed the Histogram of Oriented Gradients (HOG), the same one used in [3]. The parameters of HOG feature are shown in Table I.

Table I
Parameter of HOG Feature.

Image size	Cell size	Block size	Orientation	Dimension
16×16	2	8	5	320
32×32	4	8	5	320
64×64	8	8	5	320

B. Effect of Random Filter

Two kinds of SSD generators are compared.

1) morphology operations, Gaussian filter, color change [3]
2) morphology operations, Gaussian filter, Random Filter, color change

By adding the Random Filter, the accuracy has been improved from 56.11 % to 59.76 %.

C. Effect of Multi-Scale Resizing

Two kinds of MSR are compared using NNS (MSR-NNS).

1) 32×32 and 64×64 (15 classifiers for each, total 30 classifiers).
2) 16×16, 32×32 and 64×64 (10 classifiers for each, total 30 classifiers)

All four filters are used in the SSD generation. By adding the scale 16×16, the accuracy has been improved from 57.16 % to 59.76 %. Thus, we use 2) hereinafter.

D. Performance Comparison

We have compared the following methods.

1) NNS (without SSD generator)
2) SVM (without SSD generator)
3) RF (without SSD generator)
4) MSR-RF: (10 classifiers for each, total 30 classifiers)
5) MSR-NNS: (10 classifiers for each, total 30 classifiers)
6) MSR-SVM: (10 classifiers for each, total 30 classifiers)
7) MSR-NNS-SVM: combination of 5) and 6) (total 60 classifiers)

Both of SVM and MSR-SVM use linear SVM.

Table II shows the results as the averages of three time trials. The ensemble of different classifiers contributes to higher accuracy. RF and MSR-RF show worse results probably because of training data shortage. Figure 3 shows some examples of test images.

Table II
Accuracy and Processing Time.

Method	Accuracy [%]	Time [msec]
NNS	52.01	0.39
SVM	47.74	0.77
RF	35.43	0.20
MSR-RF	36.43	8.56
MSR-NNS	59.76	16.97
MSR-SVM	61.06	26.17
MSR-NNS-SVM	**63.19**	46.44

Figure 3. Test image examples.

IV. Conclusion

We have proposed a training data augmentation technique using some image filters and a scene character recognition system based on the ensemble voting scheme. The experiments have shown that the data augmentation method contributes to higher accuracy, especially with the SVM. Our future work includes further analyses of the system using different parameters as well as testing some other classifiers.

References

[1] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in Workshop on Deep Learning, NIPS, 2014.

[2] X. Ren, K. Chen, J. Sun, "A CNN Based Scene Chinese Text Recognition Algorithm With Synthetic Data Engine," arXiv preprint arXiv:1604.01891 (2016).

[3] L. Jiang and H. Goto, "Ensemble Classifier with Dividing Training Scheme for Chinese Scene Character Recognition," Proc. IVCNZ2017.

[4] C. Yi, X. Yang, Y. Tian, "Feature Representations for Scene Text Character Recognition," Proc. ICDAR 2013, pp.907-911, 2013.

Fast Handwritten Chinese Character Recognition Using Convolutional Neural Network and Hierarchical Overlapping Clustering

Soichi Tashima
Graduate School of Information Sciences, Tohoku University
Sendai, Japan
tashima@sc.cc.tohoku.ac.jp

Hideaki Goto
Cyberscience Center, Tohoku University
Sendai, Japan
hgot@cc.tohoku.ac.jp

***Abstract*—Reduction of the computational cost in Handwritten Chinese character recognition is crucial especially on mobile devices with limited processor performance. In this paper, we propose a candidate reduction technique based on the combination of CNN-based feature extractor and the Hierarchical Overlapping Clustering. Experimental results show that the CNN can be successfully combined with our former candidate reduction method and yields 7.73% higher accuracy at 16% faster speed.**

Keywords-Handwritten Chinese Character Recognition; Candidate reduction; Hierarchical Overlapping Clustering; Convolutional Neural Network

I. Introduction

Since a lot of mobile devices with cameras have spread widely, there is a growing demand for applications with Optical Character Recognition (OCR) capabilities. Some improvements in both accuracy and speed are still required especially in Handwritten Chinese Character Recognition (HCCR) on mobile devices with limited processor performance. Nearest Neighbor (NN) search based on the linear search has been popular in character recognition. HCCR tasks require a lot of computation since high-dimensional feature vectors are required and since thousands of character classes exist. In recent years, Convolutional Neural Network (CNN)-based methods have outperformed the conventional ones using empirically- and manually-crafted feature descriptors. Zhang et al. used the direction-decomposed feature, CNN, and adaptation[1]. Liu et al. adopted CNN-based feature extractor in support vector machine[2]. Regarding the speed improvement, Odate and Goto developed a candidate reduction method based on a tree-based dictionary and the Linear Discriminant Analysis (LDA) for HCCR. We call the underlying scheme Hierarchical Overlapping Clustering (HOC).

This paper proposes a candidate reduction technique based on the combination of CNN-based feature extractor and the HOC, and shows the effectiveness through experiments.

II. Hierarchical Overlapping Clustering and CNN-Based Feature Extraction

A. Candidate Reduction Using Hierarchical Overlapping Clustering

Figure 1 shows the schematic diagram of the HOC candidate reduction method proposed by Odate and Goto [3]. This method consists of two stages.

The first stage is binary search in the tree-based dictionary created by the HOC. We use LDA to obtain dimensionality reduction matrix. Each cluster consists of some character classes, and they are clustered into the two child nodes. During the recursive clustering, two techniques have been introduced to improve the performance. The first one is Overlapped Clustering allowing some overlap between two child clusters to prevent the degradation of recognition accuracy. The other one is Outlier-Class Generation to deal with outlier fonts to keep clustering efficiency. There are some parameters controlling the clustering: α controls the amount of overlapping, β controls Outlier-Class Generation, C_R is the threshold ratio of the number of classes in a node to that in its parental node, and H_T is the maximum height of the tree structure. In the recognition process, the feature vector is obtained from the query character image, and the search begins at the root node. The number of character candidates is reduced down to K_1 at most in this stage.

In the second stage, the Linear Search in low-dimensional space (ldLS) is applied to the character candidates in the leaf node. The number of candidates is reduced by picking up the K_2 nearest candidates in n_2-dimensional space, where K_2 and n_2 are constant parameters. The lower dimensionality contributes to faster search. These K_2 candidates are fed to the fine classifier, which is, for example, the linear search using the original dimensionality[3].

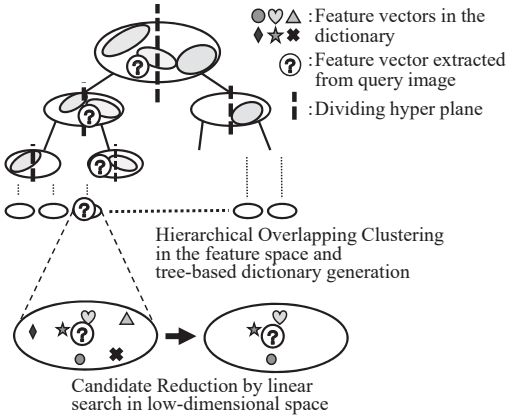


Figure 1. HOC-based candidate reduction method.

B. Feature Descriptors

In [3], 576-dimensional Peripheral Local Moment (P-LM) feature was adopted. P-LM feature is known to be robust to thickness and location changes of character strokes and used to be one of the best handcrafted features designed empirically for Japanese character recognition.

In this paper, we try to extract feature vectors using a pre-trained CNN model. CNN extracts location invariant features while the inputted image goes through some convolutional layers and subsampling layers. In the training phase, the connection weights are updated so the network extracts more effective features. Therefore, it is expected that the values extracted from one of the layers can be used for character recognition features better than the traditional handcrafted ones. Referring to [4], we configure the network model shown in Figure 2. "xCy" means a convolutional layer with x kernels whose size is $y \times y$, "MPy" means a max-pooling layer with kernels whose size is $y \times y$, "xN" means a fully connected layer with x neurons. We employed the output values of the upper FC layer 512N and use them as the feature vectors for character recognition. Thus, we obtain 512-dimensional feature vectors.

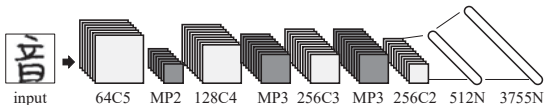


Figure 2. CNN model for feature extraction.

III. Experimental Results

A. Setup

We used CASIA-HWDB 1.1 handwritten chinese character dataset[5], which contains 3,755 classes per set (writer). The first 120 sets are for the training of CNN and constructing dictionary, and the next 80 sets are for evaluation. The computer we used for the experiments has Intel Core i7-2600 (3.40GHz) CPU and 16GB memory. We use the following three measures: accuracy, recognition time, and coverage defined by the percentage of the correct character included in the reduced candidates[3]. The parameters have been set as: $K_1 = 300$, $K_2 = 20$, $C_R = 0.97$, $H_T = 18$, $\alpha = 0.5$, $\beta = 0.3$, and $n_2 = 40$. In addition, we adopted Sequential Similarity Detection Algorithm (SSDA) for accelerating the linear search without any accuracy drop as in[3].

B. Evaluation of feature vector by CNN

We compared the performance of the feature vectors extracted by CNN and P-LM. Each feature is combined with three schemes: simple linear search (LS), HOC alone, and HOC combined with ldLS. Figure 3 shows the comparison results. LS stands for the basic linear nearest neighbor search with the SSDA.

As shown in Figure 3, the CNN feature descriptors work effectively with the HOC-based candidate reduction system. The recognition time has been much reduced from 4.48sec to 0.19sec (23.6 times faster) with mere 0.13% accuracy drop. Compared with the P-LM feature descriptors, the CNN has yielded 7.73% better accuracy and 16% faster speed. These results give support to that the CNN-based feature extractor can be applied to high-speed HCCR method with HOC dictionary.

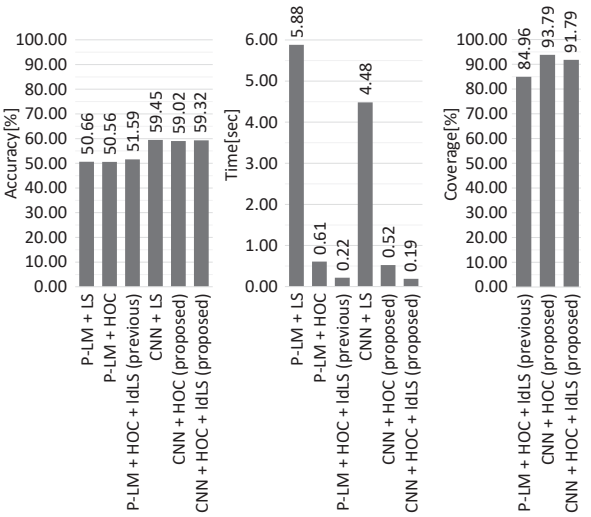


Figure 3. Performance of P-LM and CNN-based feature descriptor in linear search, HOC-based candidate reduction, and HOC-based candidate reduction with ldLS.

IV. Conclusion

This paper has presented a candidate reduction technique based on the Hierarchical Overlapping Clustering for accelerating handwritten Chinese character recognition. The experimental results have shown that the CNN-based feature descriptors can be successfully used in the HOC-based candidate reduction method we proposed before, and outperform the conventional descriptors. We have achieved 23.6 times faster recognition compared with the linear search. Our future work includes further improvements of the accuracy by combining a sophisticated classifier and improving over-all recognition speed.

References

[1] X. Zhang, Y. Bengio, and C. Liu, "Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark," Pattern Recognition, Vol.61, pp.348-360, 2017.

[2] L. Liu, P. Yang, W. Sun, and J. Ma, "Similar Handwritten Chinese Character Recognition Based on CNN-SVM," in Proc. ICGSP'17, pp.16-20, 2017.

[3] R. Odate and H. Goto, "Highly-accurate fast candidate reduction method for Japanese/Chinese character recognition," in Proc. ICIP 2016, pp.2886-2890, 2016.

[4] Y. Tang, L. Peng, Q. Xu, Y. Wang and A. Furuhata, "CNN Based Transfer Learning for Historical Chinese Character Recognition," in Proc. DAS 2016, pp.25-29, 2016.

[5] C. Liu, F. Yin, D. Wang and Q. Wang, "CASIA Online and Offline Chinese Handwriting Databases," in Proc. ICDAR 2011, pp.37-41, 2011.

LSTM Networks for Edit Distance Calculation with Exchangeable Dictionaries

Martin Schall*, Haiyan P. Buehrig†, Marc-Peter Schambach‡ and Matthias O. Franz§
*†§Institute for Optical Systems, University of Applied Sciences Konstanz, Germany
*‡Siemens Postal, Parcel & Airport Logistics GmbH, Konstanz, Germany
*Email: martin.schall@htwg-konstanz.de
†Email: haiyan.buehrig@gmail.com
‡Email: marc-peter.schambach@siemens.com
§Email: mfranz@htwg-konstanz.de

***Abstract*—Algorithms for calculating the string edit distance are used in e.g. information retrieval and document analysis systems or for evaluation of text recognizers. Text recognition based on CTC-trained LSTM networks includes a decoding step to produce a string, possibly using a language model, and evaluation using the string edit distance. The decoded string can further be used as a query for database search, e.g. in document retrieval. We propose to closely integrate dictionary search with text recognition to train both combined in a continuous fashion. This work shows that LSTM networks are capable of calculating the string edit distance while allowing for an exchangeable dictionary to separate learned algorithm from data. This could be a step towards integrating text recognition and dictionary search in one deep network.**

I. Introduction

The string edit distance [1] [2] defines a metric of similarity of two strings. It is the minimum number of character insertion, deletion or replacement operations to transform one string into the other. Information retrieval and document analysis systems use the edit distance for e.g. document retrieval or dictionary search. It is also used for evaluating text recognizers by using it as a measure of the character error rate. Use cases are e.g. the search for address elements in postal and parcel processing, the localization of genome sub-sequences or keyword search in web search engines. Optimized index structures can be used when no two arbitrary strings are compared but a query string with a dictionary of reference strings.

Long Short Term Memory *(LSTM)* networks [3] [4] trained with Connectionist Temporal Classification *(CTC)* [5] [6] produce a sequence of character probabilities while transcribing text from images. This probabilistic output is further decoded to one or more strings. Decoded strings are used for evaluation of the network or in following application steps. A language model can be used to improve decoding of the network output.

Transcription, decoding and dictionary search are often seen as separate steps. We propose to integrate these three steps into one deep LSTM network. This work is a step in this direction by showing that LSTM network can learn to calculate the string edit distance of a one-hot coded string and a dictionary of strings. A one-hot coding of strings is very similar to the probabilistic output of a CTC-trained LSTM network, but values are boolean instead of continuous probabilities. Integration of transcription, decoding and dictionary search in one network could reduce the overall error rate by allowing the network to learn domain specific statistics in all three steps. Also moving decoding and dictionary search into a LSTM network could allow speed improvements by moving the execution to a GPU accelerator.

This work uses an English word corpus [7] derived from the Google Trillion Word Corpus [8] in its experiments.

II. Methodology

Strings used in this work are in English language and between 3 and 10 characters in length. The alphabet is 26 characters in size. Each string is represented as a matrix of size 10×26 with individual characters encoded by a one-hot coding, setting one of the 26 coefficients to one and all others to zero. For example the character *A* is encoded as $[1, 0, \ldots, 0]$, *B* as $[0, 1, 0, \ldots, 0]$ and so on. Strings shorter than 10 characters in length are padded with zero coefficients. Strings are processed by the RNN as sequences of 10 length with 26 features per step.

The network takes two separate inputs. One is the encoded representation of the dictionary strings with the strings concatenated along the feature-dimension. This results in an input of size $|batch| \times 10 \times (26 \times |dictionary|)$ for mini-batch training. Dictionaries are 100 strings each in this work and thus the encoded dictionary is $|batch| \times 10 \times 2600$ in size. Second input is the representation of the query strings with $|batch| \times 10 \times 26$ in size.

The RNN consists of multiple bidirectional [9] LSTM layers with the same number of neurons per layer. The networks task is to process the query string and predict the string edit distances to the dictionary strings as a regression problem. The encoded dictionary is provided as input by concatenating it with the BLSTM input along the feature-dimension. This topology is shown in Figure 1.

Output layer of the RNN is fully connected with ReLU [10] non-linearity. This layer consists of one neuron per string of the dictionary, in our case 100 neurons. These neurons predict the string edit distances between the query string and the dictionary strings. String edit distance is zero or positive

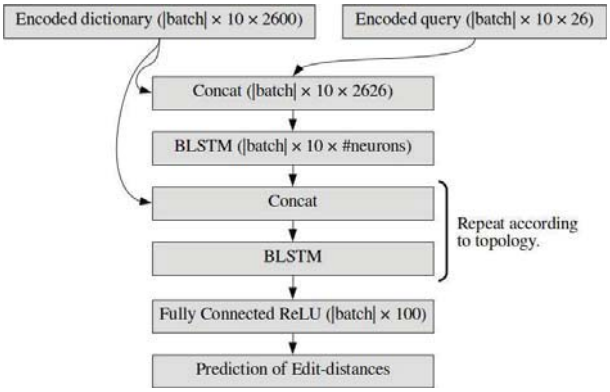


Fig. 1. Network topology for comparing query strings of up to 10 characters length with an alphabet of 26 characters to a dictionary of 100 such strings.

and as such the ReLU non-linearity is capable of predicting it without re-scaling. Loss function for training is the Mean Squared Error *(MSE)* of the predicted and correct string edit distances.

III. Results

Data for training and evaluation was derived from the 20k most frequent English words [7] [8] with a length between 3 and 10 characters, which results in a set of 16968 strings. 1000 of these were used as 10 dictionaries of 100 strings each. 9 dictionaries were for training, the other only for evaluation. A random one of the 9 dictionaries was chosen for each mini-batch during training. 80% of the remaining strings were used as query strings for training and 10% each for validation and evaluation.

Optimization of the network was done using Adam [11] with a learning rate of 0.001 and a mini-batch size of 16. Training was limited to a maximum of 200 epochs. Multiple optimization strategies were evaluated but Adam and mini-batch training produced good and reliable results.

TABLE I
RMSE FOR DIFFERENT NETWORK SIZES WITH UNSHUFFLED DICTIONARIES.

#layers × #neurons	2×30	2×60	3×60	5×200
Test set, unkn. dict.	1.78	1.78	1.56	2.13
Validation set, unkn. dict.	1.78	1.80	1.57	2.14
Training set, unkn. dict.	1.78	1.79	1.57	2.12
Test set, known dict.	0.37	0.30	0.29	0.36
Validation set, known dict.	0.37	0.29	0.29	0.36
Training set, known dict.	0.37	0.29	0.28	0.34

Table I shows the Root Mean Squared Error *(RMSE)* for the described network and experiment. The 10 dictionaries were not shuffled in this experiment and thus the strings remained in the same order within each dictionary for the whole training and evaluation. Much lower RMSE values were achieved for the 9 known dictionaries in comparison to the unknown dictionary.

Table II contains RMSE values for the same experimental set-up, but the dictionaries were randomly shuffled and thus the strings were in random order within their dictionary. Shuffling was done for each mini-batch to reduce the risk of repeating the same dictionary order. Results show a much smaller gap in RMSE between the known and unknown dictionaries.

TABLE II
RMSE FOR DIFFERENT NETWORK SIZES WITH SHUFFLED DICTIONARIES.

#layers × #neurons	2×30	2×60	3×60	5×200
Test set, unkn. dict.	0.86	0.84	0.84	0.84
Validation set, unkn. dict.	0.86	0.84	0.84	0.84
Training set, unkn. dict.	0.86	0.84	0.84	0.84
Test set, known dict.	0.85	0.82	0.82	0.81
Validation set, known dict.	0.85	0.82	0.82	0.81
Training set, known dict.	0.85	0.82	0.82	0.81

IV. Discussion

The conducted experiments are promising and show that LSTM networks are capable of learning to predict the string edit distance while separating dictionary data from the actual algorithm. The achieved RMSE of ≈ 0.8 is not enough to retrieve the correct distance by rounding. It may still enable the use of such networks for applications like decoding of CTC-trained text recognizers. Further studies are necessary to validate the assumptions made about a close integration of CTC-based text recognition and string edit distance calculation in one network.

Acknowledgment

The authors would like to thank the Siemens Postal, Parcel & Airport Logistics GmbH for funding this work.

References

[1] R. A. Wagner and M. J. Fischer, "The String-to-String Correction Problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, 1974.
[2] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory." *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM." *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
[5] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd international conference on Machine Learning*. ACM Press, 2006, pp. 369–376.
[6] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.
[7] "GitHub: Josh Kaufman," https://github.com/first20hours/google-10000-english, accessed: 2017-09-29.
[8] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.
[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2673–2681, 1997.
[10] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013.
[11] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *International Conference on Learning Representations*, pp. 1–13, 2015.

Continuous Competition on Recognition of Documents with Complex Layouts - RDCL

Christian Clausner and Apostolos Antonacopoulos
Pattern Recognition and Image Analysis Research Lab
University of Salford
United Kingdom
www.primaresearch.org

***Abstract*— This paper introduces a continuous competition and the underlying system that enables it based on the ICDAR Competition on Recognition of Documents with Complex Layouts – the most recent being RDCL2017. It is shown how researchers can perform the evaluation of their results using new functionality of the Aletheia system and how the outcome can be published on the competition website for comparison with other evaluated approaches.**

Keywords- performance evaluation; page segmentation; region classification; layout analysis; OCR; recognition; datasets

I. INTRODUCTION

Layout Analysis (Page Segmentation and Region Classification) is a critical step in the recognition workflow. Its performance significantly influences the overall success of a digitisation system, not only in terms of OCR accuracy but also in terms of the usefulness of the extracted information (in different use scenarios).

The aim of the ICDAR Page Segmentation competitions (running since 2001) has been to provide an objective evaluation of methods, on a realistic contemporary dataset, enabling the creation of a baseline for understanding the behaviour of different approaches in different circumstances [1]. The used datasets have been selected from curated repositories [2][3] containing realistic and representative documents. The last edition (RDCL2017 [4]) is based on the same principles established and refined by previous competitions with its focus being on documents with complex layouts.

In addition to having snapshots of evaluation of methods at regular intervals (e.g. at ICDAR) it is important to enable and provide a continuous evaluation facility to track progress in the field and maintain a record of the performance of different approaches over a longer time period. In the rest of this paper, the continuous evaluation system and its use is presented, after an overview of the competition itself and its modus operandi.

II. THE COMPETITION

RDCL has three objectives: 1) comparative evaluation of participating methods on a representative dataset; 2) detailed analysis of the performance in different scenarios; 3) placement of the methods into context by comparing them to commercial and open-source systems.

The initial competition (for ICDAR2017) proceeded as follows. The authors of candidate methods downloaded the *example* dataset (document images and ground truth). The *Aletheia* [5] ground-truthing system and code for outputting results in the required PAGE format [6] were also available. Three weeks before the deadline, participants downloaded the *images* of the *evaluation* dataset. At the closing date, the organisers received both the executables and the results of the candidate methods on the evaluation dataset.

Figure 1. Three images from the example set.

The importance of realistic datasets for meaningful performance evaluation has been discussed and the authors have addressed the issue for contemporary documents by creating the PRImA Layout Analysis dataset with ground truth [2]. For this competition, the evaluation set consists of 75 images selected as a representative sample ensuring a balanced presence of different issues affecting layout analysis and OCR. In addition to the evaluation set, six images were selected as the example set that is provided to the authors with ground truth (Fig. 1).

The ground truth is stored in the PAGE XML format [6]. For each region on the page there is a description of its outline in the form of a closely fitting polygon. A range of metadata is recorded for each different type of region.

III. THE EVALUATION SYSTEM

The performance analysis method [7] consists of two main parts. First, correspondences between ground truth and segmentation result regions are determined. Then, errors are identified, quantified and qualified in the context of use scenarios.

The region correspondence determination step identifies geometric overlaps between ground truth and segmentation result regions. In terms of Page Segmentation, the following situations can be determined: *merger*, *split*, *miss* / partial miss, and *false detection*. In terms of Region Classification, considering also the *type* of a region, *misclassification* can be determined as additional situation. Based on the above, the segmentation and classification errors are *quantified.* The amount (based on overlap area) of each single error is recorded

(raw evaluation data). The raw data (errors) are then *qualified* by their significance using two levels of error significance, expressed by a set of weights, referred to as an *evaluation profile* [7]. Each evaluation scenario has a corresponding profile.

For comparative evaluation, the weighted errors are combined to calculate overall error and success rates.

The complete evaluation procedure has been integrated into the Aletheia Document Analysis System [5]. A dedicated competition dialog (see Fig. 2) guides the user through the required steps, including:

- Downloading the evaluation set images
- Producing segmentation results in PAGE format
- Selecting the image and result folders
- Auto-validating the results (for completeness and XML correctness)
- Selecting one of the predefined evaluation scenarios
- Running the evaluation (takes a few minutes)
- Viewing / exporting results
- Submitting via email (optional)

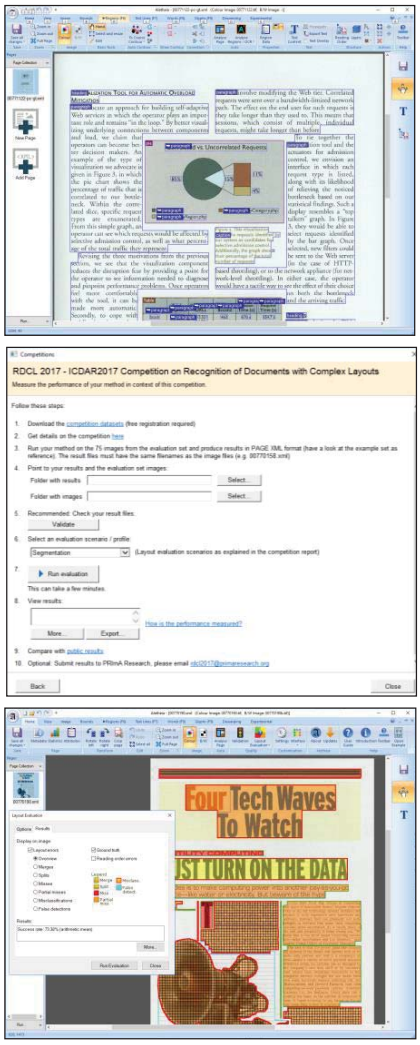


Figure 2. Aletheia with ground truth open (top), competition dialog (middle), and visual evaluation results for one page (bottom)

Detailed information and published results can be found on the competition website [8].

The user's evaluation results are presented in textual form in Aletheia and can be exported as comma-separated values (with per-page figures). The processing is performed locally on the user's system. An evaluation can therefore be repeated as often as required.

Aletheia also allows to evaluate segmentation results for individual pages, giving in-depth visual and textual feedback on different types of errors.

IV. Discussion and Conclusion

The ICDAR competitions provide biennial snapshots of page recognition methods. The continuous RDCL competition builds upon that and adds the possibility for researchers to evaluate their systems at any time. For results to be published on the competition website (primaresearch.org/RDCL2017) the same rigor as in the ICDAR competition is used (validation by organisers). The ground truth of the evaluation dataset and the exact evaluation profile are kept secret for a fairer process. New results will be displayed alongside ICDAR2017 results but labelled clearly as 'new' since the original participants had limited time to finetune their methods.

A limit for how often results can be submitted has not been set but there will be a fair-use policy in place. A short method description and/or reference will be requested for each submission.

Aletheia and the competition dataset are publicly available at primaresearch.org.

References

[1] A. Antonacopoulos, S. Pletschacher, D. Bridson, C. Papadopoulos, "ICDAR2009 Page Segmentation Competition", *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 1370-1374.

[2] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", *Proc. ICDAR2009*, Barcelona, Spain, July 2009, pp. 296-300.

[3] C. Papadopoulos, S. Pletschacher, C. Clausner, A. Antonacopoulos, "The IMPACT dataset of Historical Document Images", *Proc. HIP2013*, Washington DC, USA, August 2013, pp. 123-130.

[4] C. Clausner, A. Antonacopoulos, S. Pletschacher, "ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017", *Proc. ICDAR2017*, Kyoto, Japan, 2017, pp. 1411-1416.

[5] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", *Proc. ICDAR2011*, Beijing, China, 2011.

[6] S. Pletschacher and A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proc. ICPR2008, Istanbul, Turkey, 2010, pp. 257-260.

[7] C. Clausner, S. Pletschacher and A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proc. ICDAR2011*, Beijing, China, Sept 2011.

[8] RDCL competition website: www.primaresearch.org/RDCL2017, January 2018

A Web-Based OCR Service for Documents

Jake Walker, Yasuhisa Fujii, Ashok C. Popat
Google, Inc.
Mountain View, Ca, USA
{jakewalker,yasuhisaf,popat}@google.com

***Abstract*—Google has developed a system capable of high-accuracy OCR in many languages. It is available for general use via the Google Cloud Vision API. This paper outlines the most recent instantiation of the system behind the API: its structure and the functioning and interaction of its components. We explain some design decisions that relate to providing OCR as a high-capacity web service, a scenario which presents specific challenges. Accuracy results are provided using an internal evaluation dataset, comparing against the Tesseract open-source system. Some limitations of the current approach are noted.**

I. Introduction

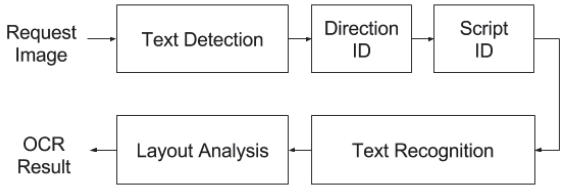
Google has made significant progress developing and integrating OCR capabilities for various internal uses over the past decade. Recently, a version of its OCR system has been made available for external use through the Google Cloud Vision API [1][1]. This OCR system is periodically tested on 232 languages in 30 distinct scripts, achieving state of the art accuracy for most of them on image types ranging from scanned documents to casual photos. Currently, a substantial subset of the internally-supported languages are available through the external API. Previous papers [2]–[4] described algorithms and components of the evolving system as they existed at the time. The present paper provides an up-to-date description or "snapshot" of the current system.

In a production system the goal is not only to optimize accuracy but also to balance considerations such as cost, generalizability, and maintainability. Design decisions informed by these constraints are noted. Acccuracy is compared with Tesseract, a well-known open-source system on an internal dataset. Known limitations of the approach are discussed and possible future directions outlined.

The contribution of this paper is threefold. First, it documents a web service that can serve as a well-performing, always-available benchmark against which emerging research or other production systems may be compared. Second, it provides by example an indication of the capabilities that might be expected when certain design choices are followed, and when training and deployment are carried out in a real-world production setting. Third, it notes current limitations to motivate research in the relevant areas of need.

[1]Google Cloud Vision currently offers a "text detection" mode optimized for photo text recognition and a "document text detection" mode optimized for text recognition on printed documents. This paper describes the system behind the document text detection mode

Fig. 1. The system has 5 stages. Each can be swapped out with a different implementation, enabling quick experimentation.



II. System: design and limitations

Processing happens in stages, as shown in Figure 1. Each stage is accomplished by a corresponding subsystem.

A. Text Detection

A CNN-based subsystem detects and localizes lines of text by generating a pixel-level "heatmap" of text likelihood that is used to generate a set of bounding boxes. Each bounding box corresponds to a single line of detected text. Our system does not currently support curved text detection, and detecting curved text without a performance degradation on linear text detection is an open area of interest for us.

B. Direction Identification

This stage classifies the direction of the characters in each line as one of: North, East, South, or West. (A by-product of this stage is to filter out some lines that were erroneously detected as text.) A single character direction is allowed per text line, while multiple lines in the same document can have different directions. A further heuristic is used to bias toward consistent directions among the lines in each page [4].

C. Script Identification

This stage identifies the main script (writing system) of the text in the line [4]. The current system assumes that each line has a single dominant script, but allows multiple lines in the same document to have different scripts. Heuristics similar to those used in direction identification are applied to reduce the frequency of spurious script classifications, by biasing decisions towards scripts detected elsewhere on the page. Multiple scripts on the same line are currently not supported, but remain an open area of interest for us. A unified recognition model across multiple scripts would potentially solve this problem, but we have not yet achieved a unified model without significant accuracy loss.

D. Text Recognition

This stage transcribes the line image into a sequence of Unicode codepoints; in this sense, it can be regarded as the main OCR step. Characters consist of one or more codepoint. A log-linear framework is used, combining an inception style optical model and a N-gram character-based language model [4] as the major inputs. A custom decoding algorithm designed for OCR is applied [3].

Using the output of this stage, language identification is performed for each line and made available as an annotation to inform client applications, or optionally, to inform another round of language-specific OCR: By default, a single model is used per script, rather than per language[2]. In general, better accuracy would result if language-specific models were used, which is the approach followed in various internal deployments of the system. For the web-based service, the decision to use script-specific rather than language-specific models was motivated by practical concerns: doing so reduces the footprint and maintenance cost of the production system. For specific languages, this decision may be overridden; this is currently done only for Vietnamese, which although based on Latin script is distinct enough in its orthographic and linguistic characteristics to warrant specialized models.

Similarly, while some gains in accuracy could be achieved using a bidirectional LSTM on top of the CNN, those gains have proven small, perhaps because the combination of an inception-based CNN and a character-based language model is already strong enough to exploit much of the available contextual information.

E. Layout Analysis

Layout analysis refers to inferring structure, and it generaly includes the determination of reading order; distinguishing among title, headings, footers, and page numbers; classifying entities such as figures, halftone images, and tables; etc. Our system restricts layout analysis to the segmentation of recognized text into blocks and paragraphs, with more sophisticated, data-driven, and trainable approaches to layout analysis left as an area of open interest. We employ a simple bottom-up clustering procedure to identify blocks: each block is initialized to be a single line, then nearest blocks are recursively merged until a heuristic criterion is met. Within blocks, paragraph boundaries are inferred by detecting indentations.

F. General considerations

To enable broad language coverage a hybrid training regime is used that involves synthetic and real data, the latter comprising both unlabeled and labeled instances. To create the synthetic data, source text collected from the web is digitally typeset in various fonts, then subjected to realistic degradations [3].

Currently, text lines are used as the basic unit throughout the system. Like other design choices, this too represents a tradeoff involving several factors: ease and effectiveness of parallelization, the ability to handle mixed-language inputs of varying granularity, and the ability to exploit linguistic context.

[2]Because of the prevalence of Latin characters in lines that are not primarily Latin script, all our models support detection of Latin characters even if no Latin-character N-grams are included in the character-based language model

TABLE I
PERFORMANCE COMPARISON: GOOGLE OCR VS. TESSERACT 4.00.00α

Language	Books			Web		
	#Lines	N-CER [%]		#Lines	N-CER [%]	
		Tesseract	Google		Tesseract	Google
Arabic	946	14.0	4.8	4208	54.8	19.4
English	1000	1.0	0.6	4868	44.0	15.6
Hindi	1067	5.4	2.5	3726	49.3	20.6
Japanese	773	28.0	4.9	3256	57.5	17.1
Russian	864	1.7	1.2	3883	36.2	16.7

III. PERFORMANCE

Table I presents accuracy figures compared with Tesseract version 4.00.00α [5] on an internal dataset consisting of line images in five languages, sampled over two domains: Google Books and general text-bearing Web images. Character error-rate (CER) is conventionally defined as edit distance divided by reference length, scaled by 100 to allow interpretation as a percentage-like figure. Here, a modified version is used: normalized CER (N-CER), which does not penalize certain substitutions, such as those involving hyphens and dashes. For both domains and for all five languages, accuracy of the Google system is higher than that of Tesseract, a well-known and widely used open-source OCR system.

IV. CONCLUSION

A Web-based OCR system was described, and design decisions discussed. The authors would be interested in correspondence about how the system might be improved, particularly for applications that support digital scholarship and cultural preservation efforts.

ACKNOWLEDGMENT

Many people have contributed in various ways to this work, including: Hartwig Adam, Ash Hurst, Jonathan Baccash, Alessandro Bissacco, Karel Driesen, Reeve Ingle, Sameer Kulkarni, Michalis Raptis, and Thatcher Ulrich.

REFERENCES

[1] G. C. Platform, *Document Text Tutorial*, Accessed Jan. 24, 2018. [Online]. Available: https://cloud.google.com/vision/docs/fulltext-annotations

[2] D. Genzel, A. C. Popat, R. Teunen, and Y. Fujii, "HMM-based Script Identification for OCR," in *Proceedings of the 4th International Workshop on Multilingual OCR*, ser. MOCR '13. New York, NY, USA: ACM, 2013, pp. 2:1–2:5.

[3] Y. Fujii, D. Genzel, A. C. Popat, and R. Teunen, "Label transition and selection pruning and automatic decoding parameter optimization for time-synchronous viterbi decoding," in *Proceedings of the 13th International Conference on Document Analysis and Recognition*, 2015.

[4] Y. Fujii, K. Driesen, J. Baccash, A. Hurst, and A. C. Popat, "Sequence-to-label script identification for multilingual OCR," in *Proceedings of the 14th International Conference on Document Analysis and Recognition*. IEEE, Nov. 2017.

[5] R. Smith, *Tesseract version 4.00.00 alpha*, 2017 (accessed January 24, 2018). [Online]. Available: https://github.com/tesseract-ocr/tesseract

Word-Hunter: Speeding up the Transcription of Manuscripts via Gamesourcing

Jialuo Chen, Alicia Fornés, Joan Mas, Josep Lladós
Computer Vision Center - Computer Science Department
Universitat Autonoma de Barcelona, Spain
jialuo.chen@e-campus.uab.cat, {afornes,jmas,josep}@cvc.uab.es

Joana Maria Pujadas
Centre for Demographic Studies
Universitat Autonoma de Barcelona, Spain
jpujades@ced.uab.es

***Abstract*—Nowadays, there are still many handwritten historical documents in archives waiting to be transcribed and indexed. Since manual transcription is tedious and time consuming, the automatic transcription seems the path to follow. However, the performance of current handwriting recognition techniques is not perfect, so a manual validation is mandatory. Given that crowdsourcing is usually boring, we propose experiences based in gamification to increase the interest of users. Concretely, we propose to validate the automatic transcription via gamesourcing through an application for Android mobile devices.**

***Keywords*-handwritten documents; crowdsourcing; gamification;**

I. Introduction

Despite the efforts in the last decades, the amount of historical manuscripts that have not yet been transcribed is still huge [2]. Consequently, they have not been properly indexed and their contents are not available through searches. Since manual transcription requires enormous human efforts, the trend is to go towards an automatic transcription. However, the existing handwriting recognition systems still need more development before trusting a completely automatic transcription. For this reason, semi-assisted approaches based on handwriting recognition[1] [6] and keyword spotting [8] have been investigated.

In the last years, the crowdsourcing [11] paradigm has shown to be an interesting alternative. The key idea of crowdsourcing is to split the work in a big amount of micro-tasks and distribute them among many users. However, even though the collaborative transcription via crowdsourcing using web-based interfaces [3] or mobile applications [1], the transcription is tedious, and many transcribers loose interest after a while.

For the above reasons, we propose to speed up the transcription of historical document collections based on two aspects: automatic transcription, and manual validation through gamesourcing (understood as crowdsourcing via gamification). Firstly, when the automatic transcription is quite accurate, the time spent by the user to validate and correct errors is lower than manually transcribing from scratch. Secondly, gamification, defined as the application of game-design elements and principles in non-game contexts, has demonstrated to engage and keep the interest of users, also in crowdsourcing activities [7], such as the $Digitalkoot$[2] transcription games at $Facebook$.

[1]http://transcriptorium.eu/
[2]http://www.digitalkoot.fi/

In summary, our gamesourcing application for Android devices is used to only validate the automatic transcription, minimizing the human effort, and still engaging the users.

II. Handwritting Recognition method

The handwriting recognition system that has been used for transcription of historical manuscripts assumes that the document images have been preprocessed, and the word images have been already segmented. The transcription system is based on Pyramidal Histogram of Characters (PHOC), Convolutional Neural Networks (CNN) and Bi-directional Long Short-Term Recurrent Neural Networks (BLSTM-RNN). Concretely, we adapt the PHOC attribute embedding to sequence learning, in order to move from word classification to continuous handwriting recognition.

This approach is divided into two parts. The first stage corresponds to CNNs that embeds small windows of text into the PHOC space. In this embedding, each attribute represents the presence of a character in a specific part of the word. For this purpose, we use the PHOCNet, but in our case, we embed small windows of text (patches), instead of the whole word. The second stage corresponds to the sequence transcription, which is performed over embedded text patches. The sequence of embeddings are recognized using BLSTM-RNNs[5]. Concretely, we use a two-layer BLSTM network that performs the sequence recognition and outputs the transcription. In this case, we do not use any dictionary or language model. For further details, the reader is referred to [10].

III. Gamesourcing mobile application

The application is based on a client-server architecture. The client part, i.e. the mobile device, takes care of the graphical interface and the interaction with the user. The server is formed by four different parts: The processing, the database and interaction, and the image repository. The processing part is devoted to the image processing algorithms used to obtain the word images and their transcriptions. The interaction with the database is done with a PHP server that controls the creation of users, the transcription results and the golden tasks. We use a MySql server to define the database. The image repository contains the images that will be shown to the player and downloaded to the mobile device.

In the client part, the Android application shows several word images and their corresponding transcriptions (these transcriptions have been provided by the handwriting

recognition system). The player has to select, for each word, the correct transcription among several possible transcriptions, all in a limited time. In spite of showing several transcriptions for each word, it might be possible that none of these transcriptions is correct. For this reason, there is an option named "none of these" (see Fig.1).

Figure 1. Android gamesourcing application.

In order to make it more fun, we give points for each correct answer. Since the goal is to validate the automatic transcription, we do not know if the answer of the player is correct or not. For this reason, we use golden tasks (words with a known transcription) to help the player learn how to correctly validate the transcriptions at the beginning, and also, to avoid that the player chooses random transcriptions. In these cases, when the player selects the wrong answer, there is a penalty in the score.

At each level, the player must validate the transcription of several words. There is a countdown timer, and if the user can correctly validate all words within the given time, then, the player goes to the next level. At higher levels, the amount of golden tasks decreases. Finally, the players can create an account to save their best scores and compete with others in a raking.

IV. Document Images

Although our architecture is generic and any kind of document collection could be used, we have chosen population documents. Population sources allow the study of the demographic behaviour and the understanding of the social and economic evolution of the past. In nominative sources, one of the most relevant keywords to index are the names and surnames. For this reason, we have selected surnames from the marriage records of the Barcelona Cathedral [9] and census records of Sant Feliu del Llobregat[3]. For training the handwriting recognition system, we have used the training set of the ICDAR-IEHHR competition [4].

V. Conclusion

In this paper we have shown that the transcription of historical documents can be speed up through gamesourcing. Our next steps will be selecting a group of users to study the playability, the learning curve, the time spent in the game, and the validation accuracy.

[3]http://dag.cvc.uab.es/xarxes

Acknowledgment

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the Ramon y Cajal Fellowship RYC-2014-16831, the CERCA Programme / Generalitat de Catalunya, and RecerCaixa (XARXES, 2016ACUP-00008), a research program from Obra Social "La Caixa" with the collaboration of the ACUP.

References

[1] A. Amato, A. Sappa, A. Fornés, F. Lumbreras, and J. Lladós, "Divide and conquer: Atomizing and parallelizing a task in a mobile crowdsourcing platform," in *Int. ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2013, pp. 21–22.

[2] V. Bachi, A. Fresa, C. Pierotti, and C. Prandoni, "The digitization age: Mass culture is quality culture. challenges for cultural heritage and society," in *Digital Heritage: Progress in Cultural Heritage. Documentation, Preservation, and Protection. EuroMed. Lecture Notes in Computer Science*, vol. 8740. Springer, 2014, pp. 786–801.

[3] A. Fornés, J. Lladós, J. Mas, J. M. Pujades, and A. Cabré, "A bimodal crowdsourcing platform for demographic historical manuscripts," in *International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 103–108.

[4] A. Fornés, V. Romero, A. Baró, J. I. Toledo, J. A. Sanchez, E. Vidal, and J. Lladós, "Competition on information extraction in historical handwritten records," in *International Conference on Document Analysis and Recognition*, 2017, pp. 1389–1394.

[5] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 5, pp. 855–868, 2009.

[6] D. Martín-Albo, V. Romero, A. H. Toselli, and E. Vidal, "Multimodal computer-assisted transcription of text images at character-level interaction," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 5, 2012.

[7] B. Morschheuser, J. Hamari, and J. Koivisto, "Gamification in crowdsourcing: a review," in *49th Hawaii International Conference on System Sciences*, 2016, pp. 4375–4384.

[8] A. Santoro, C. De Stefano, and A. Marcelli, "Assisted transcription of historical documents by keyword spotting: a performance model," in *International Conference on Document Analysis and Recognition*, 2017, pp. 971–976.

[9] G. Thorvaldsen, J. M. Pujadas-Mora, T. Andersen, L. Eikvil, J. Lladós, A. Fornés, and A. Cabré, "A tale of two transcriptions. machine-assisted transcription of historical sources," *Historical Life Course Studies*, vol. 2, pp. 1–19, 2015.

[10] J. I. Toledo, S. Dey, A. Fornés, and J. Lladós, "Handwriting recognition by attribute embedding and recurrent neural networks," in *International Conference on Document Analysis and Recognition*, 2017, pp. 1038–1043.

[11] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *IEEE third International Conference on Privacy, security, risk and trust (PASSAT), and IEEE third International Conference on Social Computing (Socialcom)*. IEEE, 2011, pp. 766–773.


CVL
Computer Vision Lab
TU
WIEN


IAPR
APA
IT
LABS
NAVER LABS EUROPE