HYBRID MODEL OF POST-PROCESSING TECHNIQUES FOR ARABIC OPTICAL CHARACTER RECOGNITION

IMAD QASIM HABEEB

DOCTOR OF PHILOSOPHY

UNIVERSITI UTARA MALAYSIA

2016

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

IMAD QASIM HABEEB

calon untuk Ijazah PhD
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"HYBRID MODEL OF POST-PROCESSING TECHNIQUES FOR ARABIC OPTICAL CHARACTER RECOGNITION"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***28 Julai 2016.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***July 28, 2016.***

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Dr. Haslina Mohd**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Shahnorbanun Sahran**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Dr. Faudziah Ahmad**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Shahrul Azmi Mohd Yusof**	Tandatangan ***(Signature)***
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Dr. Yuhanis Yusof**	Tandatangan *(Signature)*

Tarikh:
(Date) ***July 28, 2016***

Permission to Use

In presenting this thesis in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for the scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Pengecaman aksara optik (OCR) digunakan untuk mengeluarkan teks yang terkandung di dalam sesuatu imej. Salah satu fasa dalam OCR ialah prapemprosesan dan ianya membetulkan kesalahan teks yang terasil dari OCR. Kaedah berbilang output dalam OCR mengandungi tiga proses iaitu: pembezaan, penjajaran dan pengundian. Teknik pembezaan yang sedia ada mengalami kehilangan ciri-ciri penting kerana ia menggunakan N-versi imej sebagai input. Dalam pada itu, teknik penjajaran yang terdapat dalam kajian adalah berdasarkan penghampiran manakala proses pengundian adalah tidak peka kepada konteks. Kekangan-kekangan ini mengakibatkan kadar ralat yang tinggi dalam OCR. Kajian ini telah mencadangkan tiga teknik pembezaan, penjajaran dan pengundian yang ditambahbaik untuk mengatasi kekurangan yang telah dikenalpasti;. Kesemua teknik ini kemudiannya digabungkan dalam satu model hibrid yang boleh mengecam aksara optik dalam Bahasa Arab. Setiap teknik yang dicadangkan telah dibandingkan dengan tiga teknik berkaitan yang sedia ada secara berasingan. Ukuran prestasi yang digunakan adalah kadar ralat perkataan (WER), kadar ralat aksara (CER) dan kadar ralat bukan perkataan (NWER). Keputusan eksperimen menunjukkan pengurangan relatif kadar ralat dalam semua ukuran untuk teknik-teknik yang telah dinilai. Secara yang serupa, model hibrid juga telah memperolehi nilai WER, CER dan NWER yang lebih rendah iaitu sebanyak 30.35%, 52.42% dan 47.86% apabila dibandingkan dengan tiga model relevan yang sedia ada. Kajian ini menyumbang kepada domain OCR kerana model hibrid yang dicadangkan bagi teknik pasca pemprosesan boleh membantu pengecaman teks Bahasa Arab secara automatik. Oleh itu, ia akan menjurus kepada capaian maklumat yang lebih baik.

Kata Kunci: Pengecaman aksara optic Bahasa Arab, teknik pasca pemprosesan, OCR berbilang ouput.

Abstract

Optical character recognition (OCR) is used to extract text contained in an image. One of the stages in OCR is the post-processing and it corrects the errors of OCR output text. The OCR multiple outputs approach consists of three processes: differentiation, alignment, and voting. Existing differentiation techniques suffer from the loss of important features as it uses N-versions of input images. On the other hand, alignment techniques in the literatures are based on approximation while the voting process is not context-aware. These drawbacks lead to a high error rate in OCR. This research proposed three improved techniques of differentiation, alignment, and voting to overcome the identified drawbacks. These techniques were later combined into a hybrid model that can recognize the optical characters in the Arabic language. Each of the proposed technique was separately evaluated against three other relevant existing techniques. The performance measurements used in this study were Word Error Rate (WER), Character Error Rate (CER), and Non-word Error Rate (NWER). Experimental results showed a relative decrease in error rate on all measurements for the evaluated techniques. Similarly, the hybrid model also obtained lower WER, CER, and NWER by 30.35%, 52.42%, and 47.86% respectively when compared to the three relevant existing models. This study contributes to the OCR domain as the proposed hybrid model of post-processing techniques could facilitate the automatic recognition of Arabic text. Hence, it will lead to a better information retrieval.

Keywords: Arabic optical character recognition, Post-processing techniques, Multiple outputs of OCR.

Acknowledgement

Each part of this study is guided, inspired, and supported by many people. Firstly, I would like to thank all the members of my family especially my mother for their unconditional support. My goal would not be achieved without them. The most important support and guidance were from my research supervisors Dr. Shahrul Azmi Mohd Yusof and Assoc. Prof. Dr. Yuhanis Binti Yusof. Thank you very much for your great help and support. It is an honor for me to do a research under your supervisions. I would like to thank all the academic and technical staff in Utara Universiti Malaysia for their help in the study process and providing all the excellent facilities. Finally, I would like to thank the Ministry of Higher Education and Scientific Research in Iraq for financial sponsorship.

Table of Contents

List of Tables

List of Figures

Glossary of Term

Symbol: represents the smallest meaningful unit in a writing system, such as character, number, comma, signs, etc.

Token: a sequential group of symbols not containing any spaces. It consists of a small number of symbols.

String: a sequential group of symbols. It can consist of a large number of symbols including spaces.

Word: a token exists in the specific language.

Cursive Token: a token has a group of characters joined together.

Non-word error: occurs when the word produced from the OCR process does not exist in the language resource.

Real word error: occurs when the word produced from the OCR process exists in the language resource, but it does not match with the source text.

Wrong-word: also known as an incorrect word. It refers to either non-word error or real word error.

Document Image: represents any image containing a text.

Model: a symbolic representation of concepts. It can be a schematic model or mathematical.

Lexicon: a list of words that belongs to a specific language. It does not contain any information to describe the words.

List of Abbreviations

OCR	Optical character recognition
HR	Handwriting recognition
MO	Multiple outputs of OCR
LD	Levenshtein distance
PCA	ProbCons alignment
SWA	Smith–Waterman alignment
LDB	Levenshtein distance with backtrack
WER	Word error rate
CER	Character error rate
NWER	Non-word error rate
EDT	Enhanced differentiation technique
ASW	Alignment by using words separation
VCI	Voting by using context information of sentences
CGLL	Candidates' list generation by using N-gram language model and LD
HMNL	A hybrid of MO, N-gram language model, and Levenshtein distance.
MOUMT	Multiple outputs using multiple threshold values
MOUMS	Multiple outputs using multiple scanning
MOUMO	Multiple outputs using multiple OCR systems
PPS	Post-processing stage

CHAPTER ONE
INTRODUCTION

1.0 Background

An optical character recognition, commonly referred to as OCR, is used to extract and recognize texts within images (Bassil & Alwani, 2012c). Several commercial OCR systems are currently available for various purposes, such as mail sorting systems, plate number recognition systems (Singh, Bacchuwar, & Bhasin, 2012). Figure 1.1 shows the input and output of an OCR system.

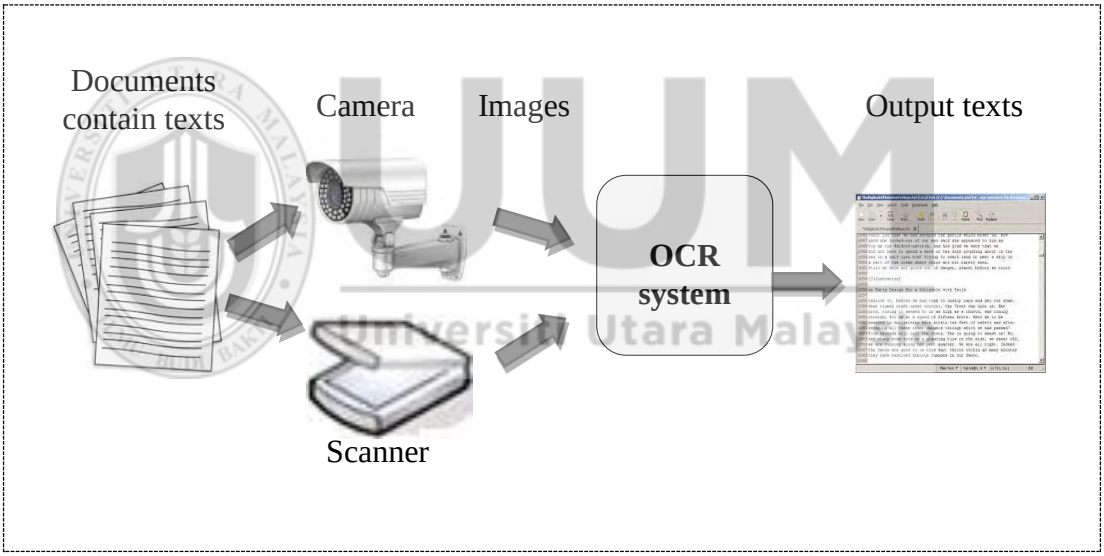

Figure 1.1. The input and output of an OCR system

There are four categories of OCR systems (El-Mahallawy, 2008). The first category is based on the type of input to these systems: offline or online. The second category depends on the mode of writing: handwritten or machine printed. The third category depends on the connectivity of a text: isolated symbols or cursive words. The last category depends on font restrictions: single font or Omni-font (Al-Badr &

The contents of the thesis is for internal user only

REFERENCES

AbdelRaouf, A., Higgins, C. A., Pridmore, T., & Khalil, M. (2010). Building a multi-modal Arabic corpus (MMAC). *International Journal on Document Analysis and Recognition (IJDAR), 13*(4), 285-302.

Abdulkader, A. E., & Casey, M. R. (2015). Efficient identification and correction of optical character recognition errors through learning in a multi-engine environment: Google Patents.

Abulnaja, O. A., & Batawi, Y. A. (2012). Improving Arabic Optical Character Recognition Accuracy Using N-Version Programming Technique. *Canadian Journal on Image Processing and Computer Vision, 3*(2), 44-46.

Ahmad, I., Mahmoud, S. A., & Fink, G. A. (2016). Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models. *Pattern recognition, 51*, 97-111.

Akhter, S., & Roberts, J. (2006). *Multi-core programming* (Vol. 33): Intel press Hillsboro.

Akila, G., El-Menisy, M., Khaled, O., Sharaf, N., Tarhony, N., & Abdennadher, S. (2015). Kalema: Digitizing Arabic Content for Accessibility Purposes Using Crowdsourcing. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 9042, pp. 655-662): Springer International Publishing.

Al-Badr, B., & Mahmoud, S. A. (1995). Survey and bibliography of Arabic optical text recognition. *Signal processing, 41*(1), 49-77.

Al-Masoudi, A. F. R., & Al-Obeidi, H. S. R. (2015). Smoothing Techniques Evaluation of N-gram Language Model for Arabic OCR Post-processing. *Journal of Theoretical and Applied Information Technology, 82*(3), 432-439.

AL-Shatnawi, A. M., AL-Salaimeh, S., AL-Zawaideh, F. H., & Omar, K. (2011). Offline arabic text recognition–an overview. *World of Computer Science and Information Technology Journal (WCSIT), 1*(5), 184-192.

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation, 49*(3), 721-751. doi: 10.1007/s10579-014-9284-1

Al-Zaydi, Z. Q., & Salam, H. (2015). Multiple Outputs Techniques Evaluation for Arabic Character Recognition. *International Journal of Computer Techniques (IJCT), 2*(5), 1-7.

Al Azawi, M. (2015). *Statistical Language Modeling for Historical Documents using Weighted Finite-State Transducers and Long Short-Term Memory*. (PhD dissertation), Technical University of Kaiserslautern, Kaiserslautern, Germany.

Al Azawi, M., & Breuel, T. M. (2014). *Context-dependent confusions rules for building error model using weighted finite state transducers for OCR post-processing.* Paper presented at the Proceeding of the 11th IAPR International Workshop on Document Analysis Systems (DAS) Loire Valley, France.

Alex, B., Grover, C., Klein, E., & Tobin, R. (2012). *Digitised Historical Text: Does it have to be mediOCRe?* Paper presented at the Proceeding of the 11th Conference on Natural Language Processing (KONVENS), Vienna, Austria.

Aljarrah, I., Al-Khaleel, O., Mhaidat, K., Alrefai, M. a., Alzu'bi, A., & Rabab'ah, M. (2012). Automated System for Arabic Optical Character Recognition with Lookup Dictionary. *Journal of Emerging Technologies in Web Intelligence, 4*(4), 362-370.

Alkhalifa, M., & Rodríguez, H. (2009). *Automatically extending NE coverage of Arabic WordNet using Wikipedia.* Paper presented at the Proceeding of the 3rd International Conference on Arabic Language Processing (CITALA2009), Rabat, Morocco.

Alobaedy, M. M. T. (2015). *Hybrid Ant Colony System Algorithm For Static And Dynamic Job Scheduling In Grid Computing.* (PhD thesis), Universiti Utara Malaysia, Kedah, Malaysia.

Andoni, A., & Krauthgamer, R. (2012). The smoothed complexity of edit distance. *ACM Transactions on Algorithms (TALG), 8*(4), 44.

Attia, M., Rashwan, M., & Khallaaf, G. (2002). *On stochastic models, statistical disambiguation, and applications on Arabic NLP problems.* Paper presented at the Proceedings of the 3rd Conference on Software Language Engineering (CLE'2002), Cairo, Egypt.

Attia, M., Toral, A., Tounsi, L., Monachini, M., & van Genabith, J. (2010). *An automatically built Named Entity lexicon for Arabic.* Paper presented at the Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010) Valletta, Malta.

Attia, M. E. (2000). *A large-scale computational processor of the Arabic morphology.* (Master thesis), Cairo University, Cairo, Egypt.

Badawi, E.-S. M. (1996). *Understanding Arabic: essays in contemporary Arabic linguistics in honor of El-Said Badawi*: American Univ in Cairo Press.

Bard, G. V. (2007). *Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric.* Paper presented at the Proceedings of the fifth Australasian symposium on ACSW frontiers, Darlinghurst, Australia.

Barnes, D. N. (2011). *The Text Contains its Own Lexicon: Extracting a Spelling Reference in the Presence of OCR Errors.* (Master dissertation), The Open University, Milton Keynes, United Kingdom.

Bassil, Y., & Alwani, M. (2012a). Context-sensitive Spelling Correction Using Google Web 1T 5-Gram Information. *Computer and Information Science, 5*(3), 37-48.

Bassil, Y., & Alwani, M. (2012b). Ocr context-sensitive error correction based on google web 1t 5-gram data set. *arXiv preprint arXiv:1204.0188*.

Bassil, Y., & Alwani, M. (2012c). Ocr post-processing error correction algorithm using google online spelling suggestion. *Journal of Emerging Trends in Computing and Information Sciences, 3*(1), 90-99.

Batawi, Y., & Abulnaja, O. (2012). Accuracy Evaluation of Arabic Optical Character Recognition Voting Technique: Experimental Study. *IJECS: International Journal of Electrical & Computer Sciences, 12*(1), 29-33.

Boyell, R. L., & Ruston, H. (1963). *Hybrid techniques for real-time radar simulation.* Paper presented at the Proceedings of the November 12-14, 1963, fall joint computer conference (AFIPS '71), Las Vegas, USA.

Cai, X. (2013). *Approximate Sequence Alignment.* (Master thesis), Louisiana State University, Louisiana, USA.

Daðason, J. F. (2012). *Post-Correction of Icelandic OCR Text.* (Master thesis), University of Iceland, Reykjavik, Iceland.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171-176.

Dehkordi, Y. H. (2014). *Incorporating User Reviews as Implicit Feedback for Improving Recommender Systems.* (Master thesis), University of Victoria, Victoria, Canada.

Do, C. B., Mahabhashyam, M. S., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research, 15*(2), 330-340.

El-Mahallawy, M. S. M. (2008). *A large scale HMM-based omni front-written OCR system for cursive scripts.* (PhD thesis), Cairo University, Cairo, Egypt.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP), 8*(4), 14.

Golding, A. R., & Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine learning, 34*(1), 107-130.

Goswami, R., & Sharma, O. (2013). A Review on Character Recognition Techniques. *International Journal of Computer Applications, 83*(7), 19-23.

Govindan, V., & Shivaprasad, A. (1990). Character recognition—a review. *Pattern recognition, 23*(7), 671-683.

Habash, N., & Roth, R. M. (2011). *Using deep morphology to improve automatic error detection in Arabic handwriting recognition.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, USA.

Habeeb, I. Q., Yusof, S. A., & Ahmad, F. B. (2014). Two Bigrams Based Language Model for Auto Correction of Arabic OCR Errors. *International Journal of Digital Content Technology and its Applications, 8*(1), 72 - 80.

Hadj Ameur, M. S., Moulahoum, Y., & Guessoum, A. (2015). Restoration of Arabic Diacritics Using a Multilevel Statistical Model. In A. Amine, L. Bellatreche, Z. Elberrichi, J. E. Neuhold & R. Wrembel (Eds.), *Computer Science and Its Applications* (pp. 181-192). Saida, Algeria: Springer International Publishing.

Herceg, P., Huyck, B., Johnson, C., Van Guilder, L., & Kundu, A. (2005). *Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents.* Paper presented at the Proceedings of the International Society for Optical Engineering (SPIE) on Visual Information Processing, Florida, USA.

Howell, D. C. (2012). *Statistical methods for psychology* (8th ed.): Cengage Learning.

Islam, A., & Inkpen, D. (2009). *Real-word spelling correction using Google Web 1T n-gram with backoff.* Paper presented at the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE2009), Dalian, China.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.): Pearson Education India.

Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K., & Ward, N. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 2): MIT Press.

Just, W. (2001). Computational complexity of multiple sequence alignment with SP-score. *Journal of computational biology, 8*(6), 615-623.

Kai, N. (2010). *Unsupervised Post-Correction of OCR Errors.* (Diploma thesis), Leibniz University, Hannover, Germany.

Kanoun, S., Alimi, A. M., & Lecourtier, Y. (2011). Natural language morphology integration in off-line Arabic optical text recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41*(2), 579-590.

Kenter, T., Erjavec, T., & Fišer, D. (2012). *Lexicon construction and corpus annotation of historical language with the CoBaLT editor.* Paper presented at the Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012), Avignon, France.

Khorsheed, M. S. (2002). Off-line Arabic character recognition–a review. *Pattern analysis & applications, 5*(1), 31-45.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226-239.

Knopp, J. (2010). *Classification of named entities in a large multilingual resource using the Wikipedia category system.* (Master thesis), University of Heidelberg, Heidelberg, Baden-Württemberg, Germany.

Kolak, O., & Resnik, P. (2005). *OCR post-processing for low density languages.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, Canada.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR), 24*(4), 377-439.

Lee, Y.-S., & Chen, H.-H. (1996). Analysis of error count distributions for improving the post-processing performance of OCR. *Communication of Chinese and Oriental Languages Information Processing Society, 6*(2), 81-86.

Lopresti, D., & Zhou, J. (1997). Using consensus sequence voting to correct OCR errors. *Computer Vision and Image Understanding, 67*(1), 39-47.

Lund, W. B. (2014). *Ensemble Methods for Historical Machine-Printed Document Recognition.* (PhD dissertation), Brigham Young University, Utah, USA.

Lund, W. B., Kennard, D. J., & Ringger, E. K. (2013a). *Combining multiple thresholding binarization values to improve OCR output.* Paper presented at the Proceedings of the International Society for Optical Engineering (SPIE) on Document Recognition and Retrieval XX, San Francisco, California.

Lund, W. B., Kennard, D. J., & Ringger, E. K. (2013b). *Why multiple document image binarizations improve OCR.* Paper presented at the Proceedings of the Workshop on Historical Document Imaging and Processing (HIP 2013), Washington, USA.

Lund, W. B., & Ringger, E. K. (2009). *Improving optical character recognition through efficient multiple system alignment.* Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Austin, USA.

Lund, W. B., & Ringger, E. K. (2011, 18-21 Sept. 2011). *Error Correction with In-Domain Training Across Multiple OCR System Outputs.* Paper presented at the

Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China.

Lund, W. B., Ringger, E. K., & Walker, D. D. (2014). *How well does multiple OCR error correction generalize?* Paper presented at the Proceedings of Document Recognition and Retrieval XXI (DRR 2014), San Francisco, USA.

Lund, W. B., Walker, D. D., & Ringger, E. K. (2011). *Progressive alignment and discriminative error correction for multiple OCR engines.* Paper presented at the Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China.

Ma, D., & Agam, G. (2012). *Lecture video segmentation and indexing.* Paper presented at the Proceedings of the International Society for Optical Engineering (SPIE) on Document Recognition and Retrieval XIX, California, USA.

Ma, D., & Agam, G. (2013). *A super resolution framework for low resolution document image OCR.* Paper presented at the Proceedings of the International Society for Optical Engineering (SPIE) on Document Recognition and Retrieval XX, California, USA.

Magdy, W., & Darwish, K. (2008). Effect of OCR error correction on Arabic retrieval. *Information Retrieval, 11*(5), 405-425.

Mai, B. Q. Q., Huynh, T. H., & Doan, A. D. (2014). *A study about the reconstruction of remote, low resolution mobile captured text images for OCR.* Paper presented at the Proceeding of the International Conference on Advanced Technologies for Communications (ATC 2014), Saigon, Vietnam.

Muaz, A. (2011). *Urdu Optical Character Recognition System* (Master thesis), National University of Computer & Emerging Sciences, Islamabad, Pakistan.

Naseem, T. (2004). *A Hybrid Approach for Urdu Spell Checking.* (Master thesis), National University of Computer & Emerging Sciences, Islamabad, Pakistan.

Naseem, T., & Hussain, S. (2007). A novel approach for ranking spelling error corrections for Urdu. *Language Resources and Evaluation, 41*(2), 117-128. doi: 10.1007/s10579-007-9028-6

Navarro, G. (2001). A Guided tour to approximate string matching. *ACM Computing Surveys (CSUR), 33*(1), 31-88.

Notredame, C. (2002). Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics, 3*(1), 131-144.

Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseract: A case study. *International Journal of Computer Applications, 55*(10), 50-56.

Pervez, M. T., Babar, M. E., Nadeem, A., Aslam, M., Awan, A. R., Aslam, N., . . . Waheed, U. (2014). Evaluating the Accuracy and Efficiency of Multiple Sequence Alignment Methods. *Evolutionary bioinformatics online, 10*, 205-217.

Pratt, W. K. (1991). *Digital image processing*: John Wiley & Sons, Inc.

Raaid, A. F., & Rafid, H. S. (2015). Performance Evaluation of Smoothing Techniques for Arabic Character Recognition. *International Journal of Research in Information Technology (IJRIT), 3*(11), 22-28.

Ramanan, M., Ramanan, A., & Charles, E. (2014). *A performance comparison and post-processing error correction technique to OCRs for printed Tamil texts.* Paper presented at the Proceeding of the 9th International Conference on Industrial and Information Systems (ICIIS) Gwalior, India.

Rardin, R. L., & Uzsoy, R. (2001). Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics, 7*(3), 261-304.

Saber, S., Ahmed, A., Elsisi, A., & Hadhoud, M. (2016). Performance Evaluation of Arabic Optical Character Recognition Engines for Noisy Inputs. In T. Gaber, A. E. Hassanien, N. El-Bendary & N. Dey (Eds.), *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt* (Vol. 407, pp. 449-459): Springer International Publishing.

Sattar, S. A. (2009). *A Technique for the Design and Implementation of an OCR for Printed Nastaliue Text.* (PhD thesis), NED University of Engineering & Technology, Karachi, Pakistan.

Shaalan, K., Samih, Y., Attia, M., Pecina, P., & van Genabith, J. (2012). Arabic Word Generation and Modelling for Spell Checking. *Language Resources and Evaluation (LREC)*, 719-725.

Shafii, M. (2014). *Optical Character Recognition of Printed Persian/Arabic Documents.* (Doctoral dissertation), University of Windsor, Ontario, Canada.

Shahrour, A., Khalifa, S., & Habash, N. (2015). *Improving Arabic Diacritization through Syntactic Analysis.* Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal.

Shannon, C., & Weaver, W. (2002). *A Mathematical Theory of Communication*: University of Illinois Press.

Silfverberg, M., & Rueter, J. (2015). *Can Morphological Analyzers Improve the Quality of Optical Character Recognition?* Paper presented at the Proceeding of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015), Tromsø, Norway.

Singh, A., Bacchuwar, K., & Bhasin, A. (2012). A Survey of OCR Applications. *International Journal of Machine Learning and Computing (IJMLC), 2*, 314-318.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). *OCR of historical printings of Latin texts: problems, prospects, progress.* Paper presented at the Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, Madrid, Spain.

Strohmaier, C., Ringlstetter, C., Schulz, K. U., & Mihov, S. (2003). *Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary.* Paper presented at the Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), Edinburgh, UK.

Taghva, K., & Stofsky, E. (2001). OCRSpell: an interactive spelling correction system for OCR errors in text. *International Journal on Document Analysis and Recognition, 3*(3), 125-137.

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR errors *Language Technology for Cultural Heritage* (pp. 3-22): Springer press.

Vrandecić, D., Sorg, P., & Studer, R. (2011). *Language resources extracted from Wikipedia.* Paper presented at the Proceeding of the sixth international conference on Knowledge capture (K-CAP '2011), Banff, AB, Canada.

Vu Hoang, C. D., & Aw, A. T. (2012). *An unsupervised and data-driven approach for spell checking in Vietnamese OCR-scanned texts.* Paper presented at the Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, Avignon, France.

Watson, J. C. (2007). *The phonology and morphology of Arabic*: Oxford university press.

Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., . . . Rashwan, M. (2006). *Building annotated written and spoken Arabic LR's in NEMLAR project.* Paper presented at the Proceeding of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

Zribi, C. B. O., & Ahmed, M. B. (2003). *Efficient automatic correction of misspelled Arabic words based on contextual information.* Paper presented at the Proceeding of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003), Oxford, UK.