ARTICLE

Development of a Gold-standard Pashto Dataset and a Segmentation App

Yan Han and Marek Rychlik

ABSTRACT

The article aims to introduce a gold-standard Pashto dataset and a segmentation app. The Pashto dataset consists of 300 line images and corresponding Pashto text from three selected books. A line image is simply an image consisting of one text line from a scanned page. To our knowledge, this is one of the first open access datasets which directly maps line images to their corresponding text in the Pashto language. We also introduce the development of a segmentation app using textbox expanding algorithms, a different approach to OCR segmentation.

The authors discuss the steps to build a Pashto dataset and develop our unique approach to segmentation. The article starts with the nature of the Pashto alphabet and its unique diacritics which require special considerations for segmentation. Needs for datasets and a few available Pashto datasets are reviewed. Criteria of selection of data sources are discussed and three books were selected by our language specialist from the Afghan Digital Repository. The authors review previous segmentation methods and introduce a new approach to segmentation for Pashto content. The segmentation app and results are discussed to show readers how to adjust variables for different books. Our unique segmentation approach uses an expanding textbox method which performs very well given the nature of the Pashto scripts.

The app can also be used for Persian and other languages using the Arabic writing system. The dataset can be used for OCR training, OCR testing, and machine learning applications related to content in Pashto.

BACKGROUND

The OCR technology for printed modern Latin scripts is a largely solved problem, as both character and word accuracies typically reach greater than 95%. Most well-known commercial OCR systems include ABBYY, OmniPage, and Adobe Acrobat OCR engine (licensed from IRIS), while open source systems have Tesseract, OCRopus, and Kraken. OCR technology for other languages and scripts, including Arabic scripts and Traditional Chinese, is still not satisfactory despite the fact that OCR research on these languages has been ongoing since the 1980s. An East Asian librarian in 2019 wrote to the author:

> I am just back from the annual AAS (Association for Asian Studies) and CEAL (Council on East Asian Libraries) meetings. This year, Prof. Peter Bol of Harvard hosted a 2-day digital tech expo there to promote digital humanities . . . I spent 1 day on the DH sessions, where scholars constantly mentioned Chinese OCR as a conspicuous and serious block on their path to assessing "digitized" textual collections. If you and your team succeed, it will surely help the EAS scholarly community a lot.[1]

Yan Han (yhan@email.arizona.edu) is Full Librarian at the University of Arizona Libraries. **Marek Rychlik** (rychlik@math.arizona.edu) is Full Professor at the Department of Mathematics, University of Arizona.

https://doi.org/10.6017/ital.v40i1.12553

Sturgeon, who has directed the Chinese Text Project since 2005, stated that OCR of premodern Chinese texts presents challenges distinct from OCR of modern documents and premodern documents in other languages, because training data is typically not available and a natural approach to improving accuracy is to train using data extracted from real images of text in the same historical writing style.[2] Sturgeon utilized both imperfect OCR software and allowed users to manually key in corresponding text via a crowdsourcing approach to gradually improve the quality of transcriptions.[3]

In 2018, the authors received a grant award from the National Endowment for the Humanities (NEH) to develop OCR and a software prototype for an open-source global language databank for Pashto and Traditional Chinese. Activities included fundamental research and software implementation of new OCR technology for the two languages. For the past two years, we have been engaged in all aspects of OCR research in Pashto, Persian, and Chinese scripts, including assessing current technology and systems, reviewing and building datasets, and researching and implementing segmentation algorithms and machine learning models involving neural networks.

Languages, Scripts, and Writing Systems

People in the world read, write, and speak a handful of major languages. Of those, reading and writing is accomplished through the use of several types of scripts: Latin, Chinese, Arabic, and Devanagari. Languages and scripts are very complex topics in regard to origin, structure, and use. They evolve due to influencing and being influenced by each other. A script is defined as "a collection of letters and other written signs used to represent textual information in one or more writing systems," where a writing system is a common communication method to allow people to exchange information through a medium such as paper.[4]

The first requirement in a writing system is letters or other written signs. A common writing system can use an alphabet, syllabary, or logography. Specifically, the Latin and Arabic writing systems use alphabets, where an alphabet is a standardized set of letters. Combination of letters makes a word. Another approach is to use a logogram. Chinese characters (including Japanese Kanji and Korean Hanja) are logograms. In the alphabet and syllabic systems, individual characters represent sounds only, while in the logographic system each logogram represents a word or a phrase.

One script, such as Latin and Arabic, may be used for several different languages, while some languages use several scripts. Latin script is used in Western Europe, most of Eastern Europe, and across North and South America. Arabic script was adopted by the West Asian, Middle Eastern, and near African regions. In contrast, the Japanese use three scripts: the hiragana and katakana syllabaries and the kanji logogram.

The next critical feature in a writing system is the order in which to read and write. A writing system has two directions: horizontal and vertical. Almost all writing systems are written vertically from top to bottom (TTB). Bottom-to-top (BTT) writing systems do exist. The Philippines traditional scripts, the Tagalog (Baybayin), Hanunóo, Buhid, and Tagbanwa are in limited use today. They are written from BTT.[5] Within the TTB method, four possibilities exist:

1. Left to right (LTR) first and TTB: This method refers to writing a horizontal line starting from the top left of a page, continuing to the right, and returning to the next line all the way from top to bottom. The Latin writing system uses this variation. The current Chinese writing system uses this order as well.

2. Right to left (RTL) first and TTB: This method refers to writing a horizontal line starting from the top right of a page, continuing to the left, returning to the next line and all the way from top to bottom. Arabic writing systems, such as Arabic, Persian, and Pashto, use this order.
3. TTB first and RTL: This method refers to writing a vertical line starting from top right of a page, continuing to the bottom, and returning to the next line all the way from right to left. This method was widely used in Traditional Chinese (before the 1950s) and traditional Japanese materials for thousands of years. It is still used in Chinese calligraphy, and occasionally can be found in materials published in Chinese.
4. TTB first and LTF: Rarely used by a writing system. One of the examples is the Manchu script.[6]

The nature of the scripts and the writing systems may require different algorithms and considerations when we deal with OCR technology, including preparing datasets, segmentation, and performing OCR in computer vision.

Pashto

Pashto (پښتو), alternatively spelled as Pushto, Pukhto, or Pakhto, historically as Afghani (افغاني), is one of the two official languages of Afghanistan (the other is Dari/Farsi/Persian). It is also spoken as a regional language in Pakistan. Pashto is spoken by 40 to 60 million people in Afghanistan and Pakistan.[7] The Arabic script writing system is used for writing Arabic, Persian, and Pashto languages in a cursive style. Arabic, Persian, and Pashto are totally different languages, though they use almost the same alphabets within the same writing system. The Pashto alphabet is a modified form of the Arabic alphabet. It consists of 45 letters and four diacritic marks and includes all 28 letters from the Arabic alphabet. The Pashto alphabet includes all 32 letters from the Persian alphabet, of which 28 letters are from the Arabic alphabet. The romanization of Pashto consists of several standards including the American Library Association (ALA) and Library of Congress (LC) ALA-LC Romanization, BGN/PCGN, DIN 31635, ISO233, and ArabTex. Details of romanization of Pashto letters with their initial, medial, final forms, and the ALA-LC rules are available at Library of Congress's website.[8]

The Need for Datasets

The authors are currently engaging in OCR research, and have applied machine learning (ML) models and methods such as convolutional neural networks (CNN) and recurrent neural networks (RNN). The advance of ML models and multiple methods has achieved great improvements in many fields. For instance, the most well-known event in ML occurred when an AI program named AlphaGo defeated the World Go champion in 2015. Open-source OCR systems Tesseract and OCRopus both released their OCR systems using the RNN models in 2014 and 2018. These models and methods rely heavily on datasets for training, improvement, and evaluation. Similarly, AlphaGo uses datasets for training and evaluations. Good and comprehensive datasets are critical to the success of an ML model and/or method. The most well-known dataset is the MNIST database which contains a training set of 60,000 images and a test set of 10,000 images (28 × 28 pixels) of handwritten digits (0–9). The dataset is widely used for training and testing in ML as the gold-standard dataset for ML techniques and pattern recognition.

Related Datasets

Currently, few Pashto datasets are available as open access. While there are other Pashto datasets mentioned in the literature, we have not found one that provides a one-to-one mapping of line images to texts.

A search on GitHub has one result showing a raw text dataset containing content in Pashto scraped from the web. However, this dataset is of little use in the case of training ML models for OCR, because it has no corresponding text. The Computer Science Department of the National University of Computer and Emerging Sciences (NUCES) Peshawar campus has been working on Pashto OCR since 2006, and its research has created a Pashto image-to-ligature dataset titled FAST-NU dataset, containing 4,000 images of 1,000 unique ligatures in a variety of font sizes.[9] The creators of this dataset have kindly sent us the Pashto image-to-ligature dataset. A recent paper discussed the use of deep learning architectures for OCR in Pashto with the development of a bigger dataset based on the FAST-NU dataset including contours, negative, and rotated images.[10]

Ali developed a database recording Pashto digits from 25 male and 25 female native Pashto speakers for automatic speech recognition. Unfortunately, the authors had difficulties in downloading this dataset.[11]

Khan et. al. designed a database encompassing a total of 4,488 images (102 distinguishing samples for the 44 Pashto letters). This approach is very close to that of the FAST-NU dataset.[12] We are not sure if they are very similar, as we have not found a way to download and evaluate the dataset.

Another article describes offline Pashto OCR using ML which tested more than 5,000 images in the dataset.[13] The article describes its "extraction of lines containing Pashto content," but these "lines containing Pashto content" have no specific resource or link to check.

Rawan and Han compiled a Pashto–English Dictionary, which is open accessible through its website and an Android app.[14] In the past decade of working with Afghan materials, Han and Rawan found several existing Pashto language dictionaries online but encountered several issues related to standardized spelling, pronunciation, romanization/transliteration, and limited content. This improved dictionary contains over 12,000 entries of Pashto words; each entry has a Pashto word and corresponding English meanings. The Pashto–English dictionary has been created with the following objectives in mind: a) standardized spelling and vocabulary, b) standard pronunciation, and c) standardized romanization with the ALA–LC romanization scheme. Other published Pashto dictionaries either use one of the above or a combination of a few romanization systems. This dataset is available for noncommercial use upon reasonable request.

Two datasets but in different languages (Arabic and Persian) were produced by the Open Islamicate Texts Initiative, available in Github (https://github.com/OpenITI/OCR_GS_Data).[15] Both Arabic and Persian datasets have scans of original books from the premodern and corresponding texts.[16] For example, its Persian datasets came from page images from three Persian books. These pages were segmented into separated line images and the line images were transcribed with corresponding Persian texts.

BUILDING A PASHTO DATASET

Our dataset creation methodology consists of three phases:

- The first is to select Pashto publications from our largest digital Afghan collections. The focus was to have a language specialist who selected publications varying in fonts, original quality, and publication years.
- The second phase is to use our segmentation app to produce line images from page images of the selected titles. Because of the nature of Pashto alphabets, we took a different segmentation approach involving expanding textboxes. This approach produced positive outcomes.
- The final phase is to generate gold-standard text from corresponding line images involving human key-in and final review. We originally hoped that OCR generated text could increase productivity. Unfortunately, the text produced from the current open-source OCR system Tesseract 4.x was not useful. A Persian Ph.D. student was hired to complete the one-to-one key-in. Finally, the author and his colleague reviewed the dataset.

Data Source

Rawan and Han at the University of Arizona Libraries have been collaborating with the Afghanistan Centre at Kabul University (ACKU), the de facto National Library of Afghanistan. The purpose of the 13-year-long collaboration is to preserve and provide open access to Afghanistan's unique materials from the ACKU's physical collections. Initially funded by a grant of $350,000 from the National Endowment for the Humanities (NEH) for the period of 2008 to 2012, the project digitized 200,000 pages of materials from the modern period. The project continues to receive support from the University of Arizona and the ACKU. The ACKU's permanent collection is the most extensive in the region covering a time of war and social upheaval in the country, with most of the documents in the principal languages of Pashto, Dari (Persian), and English with a variety of formats such as monographs, series, reports, yearbooks, videos, and newspapers. In addition, Rawan and Han also pursued related Afghani scholars' collections including those of Ludwig W. Adamec and M. Mobin Shorish. A repository (www.afghandata.org) has been openly accessible containing these unique materials dating from the 1950s to the present. The repository has grown from the initial 200,000 pages to 2 million, and is the biggest digital repository in the world covering Afghanistan and its region with more than 200,000 active users viewing 400,000 pages per year.

The wealth of the materials in terms of content, formats, and sources of information makes them undoubtedly the ultimate source of information for the studies of Afghanistan and its region. From a data scientist's point of view, the repository is a treasure trove for big data and ML purposes because it consists of a diversity of content from many sources in a variety of formats and document layouts.

Selection

The selected books, published in 1986, 2002, and 2006 respectively, vary in fonts, printing, and digitization quality. Ms. Rawan, a language specialist, selected ten Pashto books from the Afghan Digital Repository. Due to the limited funding available, only three books were used as the source for the dataset. More titles can be added if additional funding is available in the future.

1. "په حالنامه کی دبایزید روښان عرفانی اوفلسفی څیره /څیړونکی محمد اکبر کرګر." (*Mystic and Philosophic Profile of Bayazid Roshan as Reflected in Halnama*), published in 2006 and digitized in 400 dpi in grayscale (www.doi.org/10.2458/azu_acku_bp189_5_bay29_pay23_1385).
2. "لیکنه او څیړنه معصومه رضاء سید" (*Women in Life*), published in 2002 and digitized in 600 dpi in black and white (www.doi.org/10.2458/azu_acku_bp173_4_ray62_1381).
3. "تعلمیم القران او دینیات ..." (*Teaching the Qur'an and Theology*), published in 1986 with lower quality printing in a different Pashto font digitized in 600 dpi in black and white (www.doi.org/10.2458/azu_acku_bp45_tay67_1365).

Image Processing Algorithms and the Segmentation App

In a general traditional OCR system, the workflow of recognition consists of the following stages: preprocessing, document layout analysis, page segmentation, classification, and postprocessing. In our research, we refer to segmentation as the process of partitioning a digital image into one or more information blocks. The segmentation app is used during the preprocessing and segmentation stages. Sturgeon's paper discussed major methods for preprocessing and character segmentation.[17] Multiple papers discussed various methods to do Arabic or Pashto text segmentation, including

- horizontal projection[18]
- baseline [19]
- template matching [20]
- contour analysis [21]
- zoning, [22] and
- a combination of one or more above methods such as contour analysis and template matching.[23]

These methods have certain issues when dealing with letters with dots on the top or the bottom, and diacritics, specific to Pashto scripts, as the Pashto alphabet contains more letters and diacritics than its counterparts in Arabic and Persian. In addition, noise from original low-quality printing and digitization creates additional barriers. Ullah et. al.[24] briefly mentioned text area detection and segmentation with the detection and removal of diacritics. Their segmentation goes from line segmentation using the horizontal projection, to word, and then to character level progressively. The letters (e.g., ښ, ټ, څ) are sensitive to noise randomly appearing in page images. Our method has proven to be successful in getting accurate character and line segments with the benefits of simplicity and program efficiency. Details of discussion of the method are beyond the scope of this paper and shall be discussed in another paper.

The author and a postdoctoral researcher created the code to identify Pashto/Persian text lines from page images, where the page images are from our digitization master files. Our method takes a different approach from the above segmentation methods. Algorithms and specific properties related to the characteristics of Pashto letters have been implemented. We called it the "expanding textbox" method, which calculates the overlapping ratio of one textbox with the others and merges them based on a confidence level controlled by users. The confidence level of overlapping ratio is controlled by properties such as TextBox, OverlapType, OverlapThreshold, MaxDiacriticsSize, and MinLineHeight. To achieve segmentation, the app is also a specific image processing program that contains common preprocessing algorithms such as binarization.

All commercial and open-source OCR systems give users few choices in page segmentation. We believe that the availability of flexible adjustments unique to the Pashto/Persian/Arabic alphabets allows users to achieve accurate results based on analysis of our largest collections of Pashto materials. Our huge collections of printed materials spanning the period from the 1950s to the 2010s were published by governments, non-profit organizations, local companies, and individuals. These materials were printed in a diverse range of fonts and printing quality. The app has unique features to allow users to adjust several variables to ensure that they have accurate segmentation. Segmentation parameters such as vertical expansion and horizontal expansion (see fig. 1) can be adjusted to expand the line vertically and/or horizontally. Our experiments show that typically vertical expansion is set to −0.15 and horizontal expansion is set to 5 for most of the page images from our collections. However, both variables are subject to change if lines are not segmented correctly. Figures 2 and 3 show a real-life example of the different values in the vertical expansion (set to 0.20) to get all of the correct lines. Users can adjust these variables to achieve desired outputs if diacritics and lines are not recognized correctly.

The app was programmed using MatLab, which can run on MatLab or run independently if packaged with MatLab. The app can be exported to other platforms and run in batch mode if needed. The app has a simple GUI (see fig. 1) providing a preview of expanded ligatures, Expanded diacritics, lines of text, and binarized image windows. This allows users to adjust segmentation variables and verify results before outputting. Figure 4 demonstrates an example of lines of text preview. When satisfied, users can output these lines as images (one image per line from a page image). These line images are ready for OCR or manual transcription.

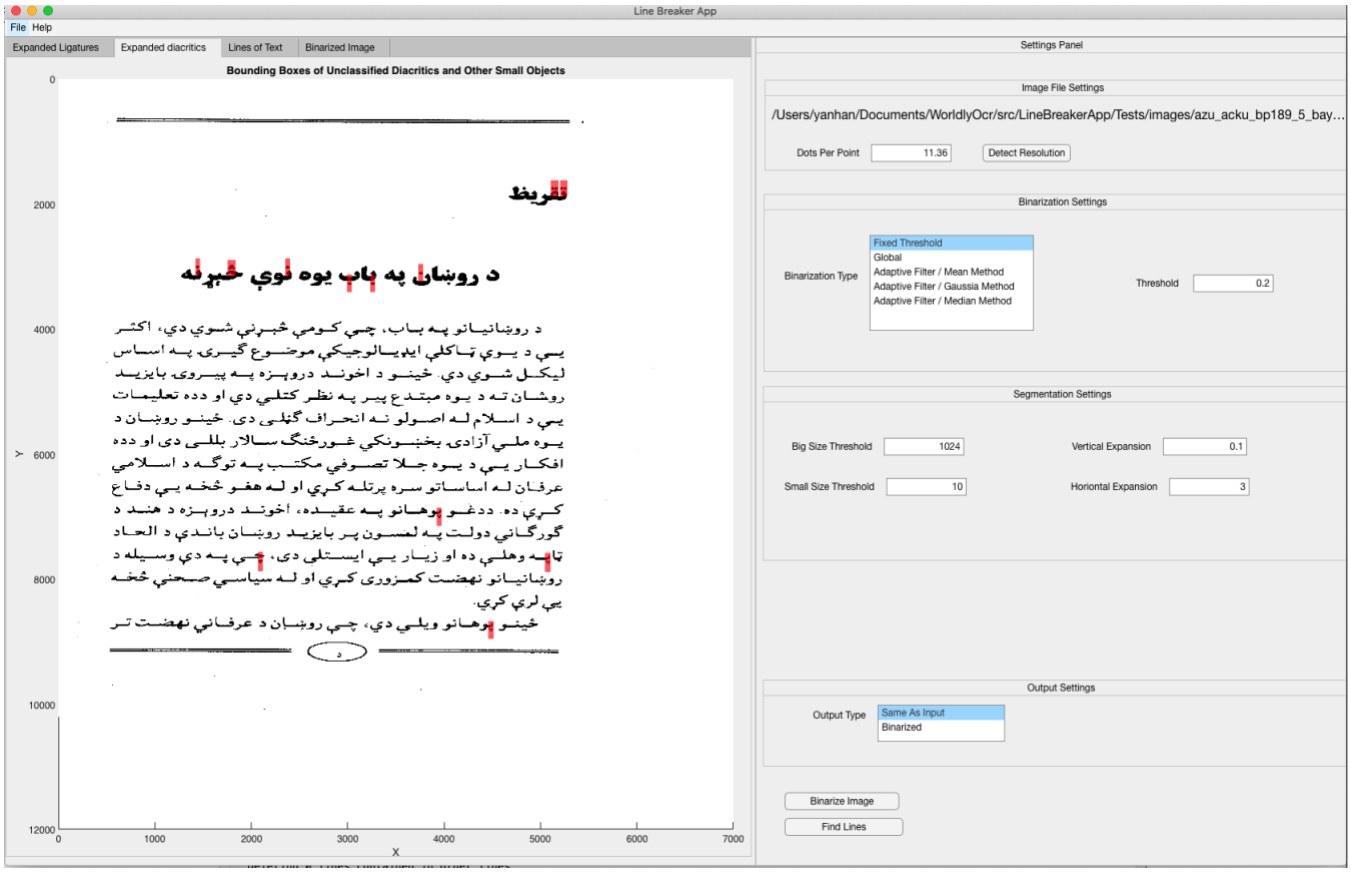

Figure 1. Expanded diacritics (highlighted in red) and the app GUI.

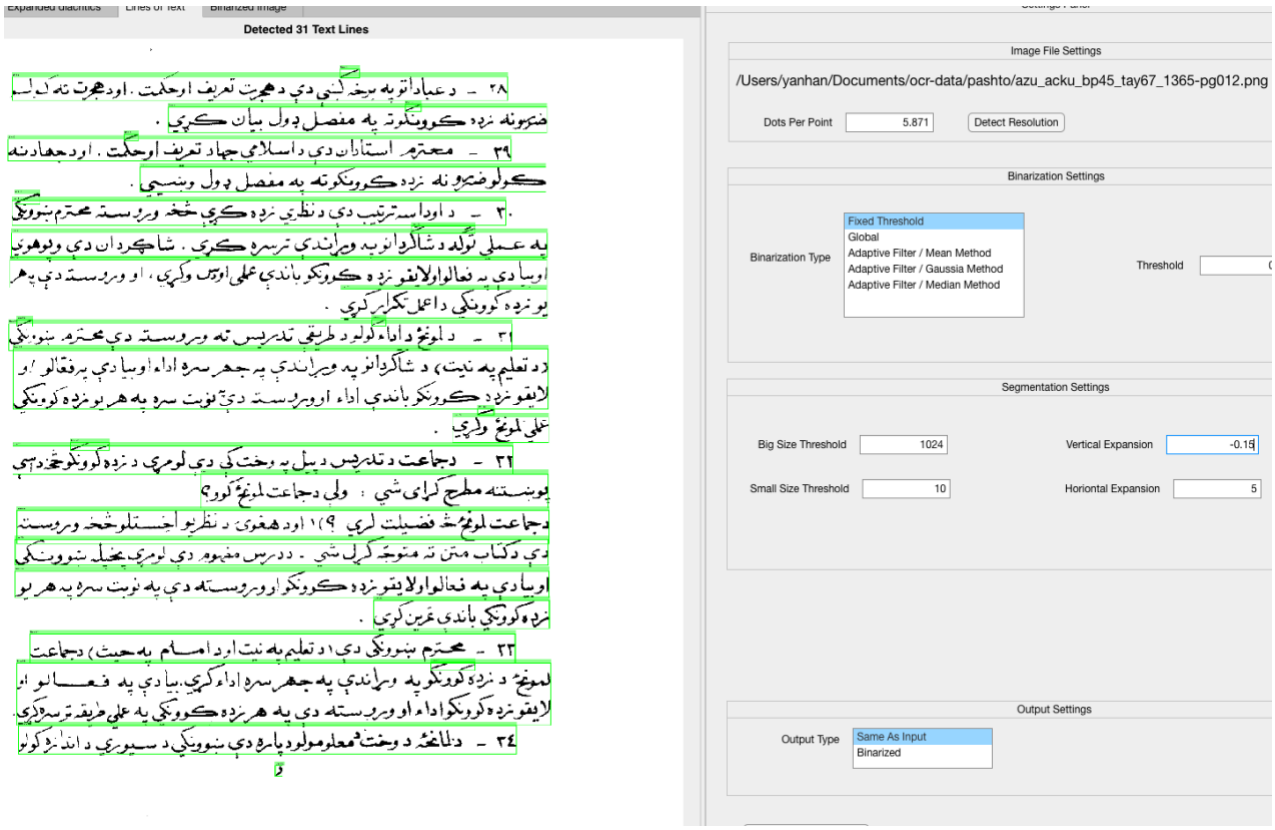

Figure 2. Vertical expansion set as −0.15 missing two lines.

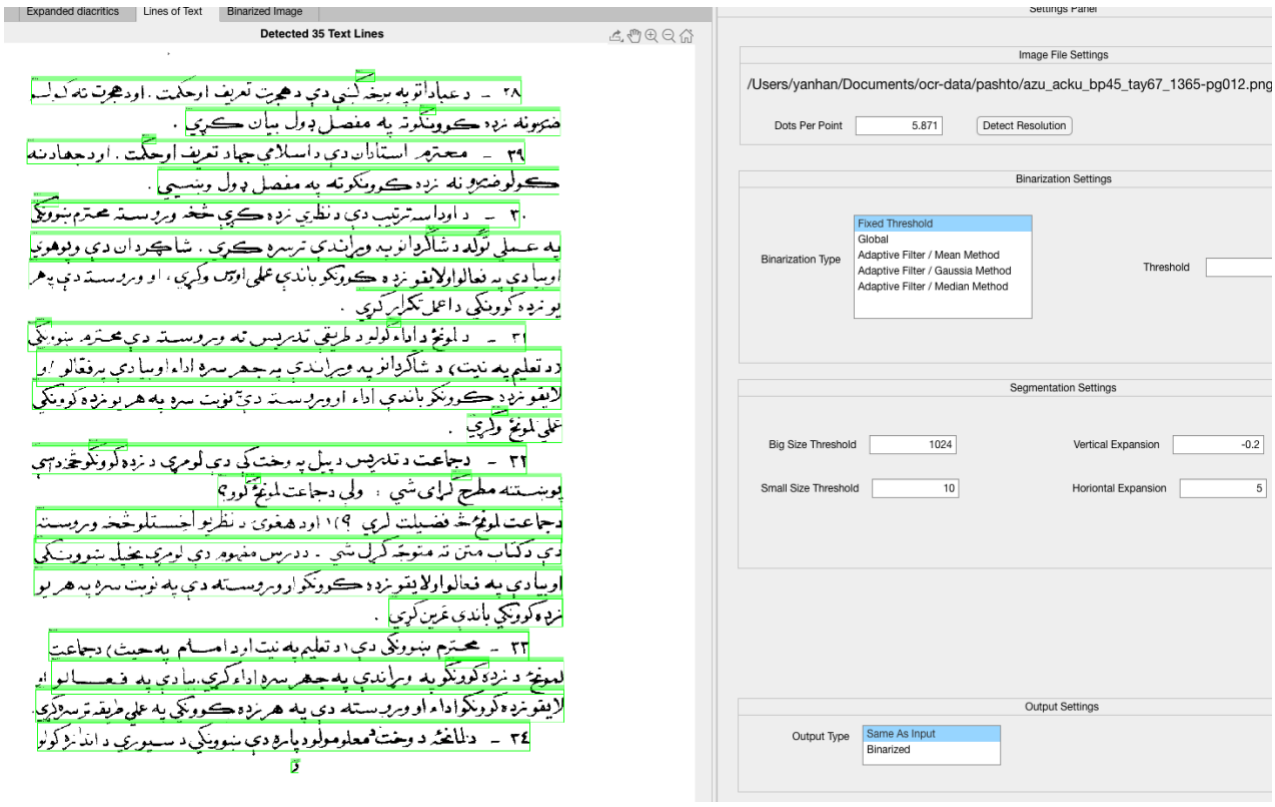

Figure 3. Vertical expansion set as −0.20 producing correct results

Figure 4. Text lines identified (lines in green).

Finalizing the Dataset

To build a truly 100% accurate dataset, we have a language specialist who keyed in, verified, and double-checked the corresponding Pashto texts. A Ph.D. student from the School of Middle Eastern and North African Studies (with Persian language fluency) was hired to complete this task.

Initially, we tried to OCR these line images by using the open-source system Tesseract 4.x with the hope that its output would speed up the key-in process. Unfortunately, the majority of the OCR results from these line images was not usable. To ensure that the dataset has the gold-standard one-to-one mapping of a line image to a line text, the Ph.D. student keyed in Pashto texts line by line by viewing every individual line image. Figure 5 shows a sample line image and its text. Finally, Ms. Rawan and the author Han reviewed these line images and their corresponding texts.

The dataset is organized in a hierarchical structure consisting of directories, where each directory contains line subdirectories which hold line images and its texts. The dataset is openly available at GitHub (https://github.com/yhan818/Pashto-Dataset).

Figure 5. Sample line image and corresponding text.

DISCUSSION

The nature of the scripts and the writing systems may require different algorithms and considerations when we deal with OCR technology, including preparing datasets, segmentation, and performing OCR in computer vision. In our research in specific languages, we have tested this app with documents in Pashto, Persian/Dari and Arabic with successful results. Our textbox extension method should work for any language using the Arabic writing system beyond these above scripts.

During our research, we are clearly aware of the following limitations of the OCR technology, techniques and systems:

1) Lack of high accuracy in segmentation:
   a) While it is true that OCR on the character/word accuracy of the Latin scripts can exceed 95% accuracy, one shall not believe that the accuracy of a document after OCR will be at the same level. Depending on the nature of a document, segmentation accuracy varies among documents. OCRing documents in simple layout (e.g., a monograph without columns and tables) generally reaches high accuracy, while OCRing documents with complex layouts (e.g., newspapers and scientific articles) generates poor results.
   b) We have tested multiple popular commercial and open-source OCR systems specifically in the area of segmentation. On several samples, every OCR system failed completely. In other words, the text output of every OCR system is nonsense. In some cases, only ABBYY recognized columns correctly; the remaining systems unexpectedly transposed

text columns, which means potential indexing and searching errors, although the character and word accuracy reached 95% accuracy.

c) We argue that segmentation accuracy shall be added as one of the most important evaluation criteria.

2) To date almost all OCR technology and systems are limited to text only:
   a) Missing information in other formats: We agree that plain text in the writing systems is the most commonly used and a very important communication method. However, almost all materials contain information in other formats (e.g., illustrations, figures, tables, and formulas) that may be very difficult to describe in text. An individual page from a monograph, journal article, or newspaper may contain information in other formats beyond text. Such information can be a table, mathematical formula, figure, picture, or drawing. One company, Mathpix, recently started to provide OCR on simple tables and mathematical formulas for a fee.
   b) Missing semantic information: In addition, current OCR simply outputs plain text, ignoring existing semantic information (e.g., bold highlighted text and section/subsection headings in different fonts and sizes). OCR refers to character recognition, which limits its own scope in theory. In current practices, semantic information is totally ignored by every OCR system. Scant research has been carried out for them too.

CONCLUSION AND FUTURE WORK

So far, we have created a Pashto dataset containing 300 line images and their corresponding text from three Pashto monographs published in 1986, 2002, and 2006, respectively. The dataset is openly available as a gold-standard Pashto dataset from real books. When future funding is received, we will add more data to this dataset.

The segmentation app produces accurate line images from page images for Pashto and Persian content. It will work for other languages using the Arabic writing system. Potential users of our prototype software will find it is relatively easy to modify with little knowledge of the underlying technology in other programming languages such as Java. In addition, researchers who understand linear algebra in which MATLAB is used can modify the code for their needs.

We are also using this dataset to train and evaluate our current OCR algorithms with RNN and other ML models. An initial report of our research and results can be found at arxiv.org.[25] The authors will report and update future research results and available datasets via conferences and formal publications.

The authors would like to thank the National Endowment for the Humanities for its grant (PR-263939-19) to our project Development of Image-to-text Conversion for Pashto and Traditional Chinese. The authors would like to thank Riaz Ahmad and Saeeda Naz for providing the NUCES FAST ligature dataset. The authors would also like to thank Atifa Rawan, Sayyed M. Vazirizade, and Sharam Parastesch for their valuable contributions. Ms. Rawan selected the sample Pashto manuscripts and reviewed the lines. Dr. Vazirizade worked on segmentation algorithms and code. Ph.D. student Sharam Parastesch keyed in and verified the dataset.

GLOSSARY OF TERMS

- **Alphabet**: An alphabet is a standardized set of letters and symbols. The most popular are the Latin alphabet (a–z) and the Arabic alphabet.
- **Classification**: In machine learning, classification is to assign a sample to one or more classes by supervised learning from examples.
- **Dataset**: A set of data. A dataset can be in a variety of forms or formats (e.g., text, images, audio, videos, 3D objects, GPS data, machine learning data), from one table, to a collection of images of handwritten digits (e.g., MNIST database), to a collection of metadata of a digital repository (e.g., arXiv dataset).
- **Document layout analysis**: In the OCR technology, document layout analysis is the process of identifying the layout and categorizing the information blocks in a digital image of a document. The goal is to segmentalize one information block from the other and arrange these information blocks in the correct reading order.
- **Language**: A structured system of communication; the system of linguistic signs or symbols considered in the abstract (as opposed to speech).
- **Left to right (LTR) first and top to bottom (TTB) writing direction**: Writing a horizontal line starting from the top left of a page, continuing to the right, and returning to the next line all the way from top to bottom. The Latin writing system uses this writing direction. The current Chinese writing system uses this order as well.
- **Logogram**: A written character that represents a word or phrase. The most popular are Chinese (simplified and traditional) characters, kanji (Japanese), and hanja (Korean).
- **Optical Character Recognition (OCR)**: Conversion of an image consisting of text (printed or handwritten) into digital text.
- **Page segmentation**: Segmentation process for a scanned page in a digital image file format.
- **Right to left (RTL) first and top to bottom (TTB) writing direction:** Writing a horizontal line starting from the top right of a page, continuing to the left, returning to the next line and all the way from top to bottom. The Arabic writing systems such as Arabic, Persian, and Pashto use this order.
- **Script**: "[A] collection of letters and other written signs used to represent textual information in one or more writing systems."[26]
- **Segmentation**: Segmentation is the process of partitioning a digital image of a document into multiple segments where each segment consists of a set of pixels. It aims at separating the digital image into one or more information blocks, where each information block contains logical information separated from the other information block. These information blocks shall be arranged in the correct reading order. (see document layout analysis)
- **Textbox**: In an OCR system, a textbox is a box with (x,y) (identified in the computer source code) that contains one or more characters.
- **Top to bottom (TTB) first and left to right (LTR) writing direction:** Writing a vertical line starting from the top left of a page, continuing to the bottom, and returning to the next line all the way from left to right. This method is rarely used by any writing system.
- **Top to bottom (TTB) first and right to left (RTL) writing direction:** Writing a vertical line starting from the top right of a page, continuing to the bottom, and returning to the next line all the way from right to left. This method was widely used in Traditional Chinese (before 1950s) and traditional Japanese materials for thousands of years. It is still used in Chinese calligraphy, and occasionally can be found in materials published in Chinese.
- **Writing system**: A common communication method to allow people to exchange information through a medium such as paper.

ENDNOTES

[1] Lu Gan, email message to author, March 25, 2019.

[2] Donald Sturgeon, "Large-scale Optical Character Recognition of Pre-modern Chinese Texts," *International Journal of Buddhist Thought and Culture* 28, no. 2 (2018): 11–44, https://dsturgeon.net/papers/large-scale-chinese-ocr.pdf.

[3] Donald Sturgeon, "Digitizing Premodern Text with the Chinese Text Project," *Journal of Chinese History* 4, no. 2 (2020): 486–98, https://doi:10.1017/jch.2020.19.

[4] "*Glossary of Unicode Terms*," The Unicode Consortium, last updated May 20, 2020, http://www.unicode.org/glossary/.

[5] *The Unicode Standard Version 13.0—Core Specification: Chapter 17: Indonesia and Oceania* (The Unicode Consortium: Mountain View, CA, 2020), https://www.unicode.org/versions/Unicode13.0.0/ch17.pdf#G26723.

[6] Britta-Maria Gruber and Wolfgang Kirsch, "Writing Machu on a Western Computer (an interim report)," *Saksaha: A Journal of Manchu Studies*, 3, (1998): https://doi.org/10.3998/saksaha.13401746.0003.008.

[7] Herbert Penzl, *A Grammar of Pashto: A Descriptive Study of the Dialect of Kandahar, Afghanistan.* (New York: Ishi Press, 2009).

[8] Library of Congress, *Pushto Romanization Tables* (2013), https://www.loc.gov/catdir/cpso/romanization/pushto.pdf.

[9] Riaz Ahmad et al., "Robust Optical Recognition of Cursive Pashto Script Using Scale, Rotation and Location Invariant Approach," *PLOS ONE* 10, no. 9 (September 14, 2015): e0133648, https://doi.org/10.1371/journal.pone.0133648.

[10] Shizza Zahoor et al., "Deep Optical Character Recognition: A Case of Pashto Language," *Journal of Electronic Imaging* 29, no. 02 (March 4, 2020), https://doi.org/10.1117/1.JEI.29.2.023002.

[11] Zakir Ali et al., "Database Development and Automatic Speech Recognition of Isolated Pashto Spoken Digits Using MFCC and K-NN," *International Journal of Speech Technology* 18, no. 2 (June 2015): 271–75, https://doi.org/10.1007/s10772-014-9267-z.

[12] Sulaiman Khan et al., "KNN and ANN-Based Recognition of Handwritten Pashto Letters Using Zoning Features," *International Journal of Advanced Computer Science and Applications* 9, no. 10 (2018), https://doi.org/10.14569/IJACSA.2018.091069.

[13] Sultan Ullah et al., "Offline Pashto OCR Using Machine Learning," in *2019 7th International Electrical Engineering Congress (IEECON)*, (Hua Hin, Thailand, 2019): 1–4, https://doi.org/10.1109/iEECON45304.2019.8938859.

[14] Atifa Rawan and Yan Han, *The Pasto-English Dictionary* (2014), http://www.pashtoenglish.org.

[15] Open Islamicate Texts Initiative, OPEN ISLAMICATE TEXTS INITIATIVE (OPENITI): Creating the digital infrastructure for the study of the premodern Islamicate World (2016), https://iti-corpus.github.io/.

[16] Matthew Thomas Miller, Maxim G. Romanov, and Sarah Bowen Savant, "Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans," *International Journal of Middle East Studies* 50, no. 1 (February 2018): 103–9, https://doi.org/10.1017/S0020743817000964.

[17] Sturgeon, "Large-scale Optical Character Recognition of Pre-modern Chinese Texts," 11–44.

[18] Mohamed Attia and Mohamed El-Mahallawy, "Histogram-Based Lines and Words Decomposition for Arabic Omni Font-Written OCR Systems; Enhancements and Evaluation," in *Computer Analysis of Images and Patterns*, ed. Walter G. Kropatsch, Martin Kampel, and Allan Hanbury, vol. 4673, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer Berlin Heidelberg, 2007), 522–30, https://doi.org/10.1007/978-3-540-74272-2_65; Mahmoud A. A. Mousa, Mohammed S. Sayed, and Mahmoud I. Abdalla, "Arabic Character Segmentation Using Projection Based Approach with Profile's Amplitude Filter," *ArXiv:1707.00800 [Cs]*, July 3, 2017, http://arxiv.org/abs/1707.00800.

[19] Atallah Al-shatnawi and Khairuddin Omar, "Methods of Arabic Language Baseline Detection—The State of Art," *International Journal of Computer Science and Network Security* 8, no. 10 (October 2008); Tarik Abu-Ain et al., "A Novel Baseline Detection Method of Handwritten Arabic-Script Documents Based on Sub-Words," in *Soft Computing Applications and Intelligent Systems*, ed. Shahrul Azman Noah et al., *Communications in Computer and Information Science* 378 (Springer: Berlin, Heidelberg, 2013), 67–77, https://doi.org/10.1007/978-3-642-40567-9_6; Saeeda Naz et al., "Challenges in Baseline Detection of Arabic Script Based Languages," in *Intelligent Systems for Science and Information*, ed. Liming Chen, Supriya Kapoor, and Rahul Bhatia, *Studies in Computational Intelligence* (Springer International Publishing, 2014), 542: 181–96, https://doi.org/10.1007/978-3-319-04702-7_11.

[20] Majid Ziaratban and Karim Faez. "A Novel Two-Stage Algorithm for Baseline Estimation and Correction in Farsi and Arabic Handwritten Text Line," in *2008 19th International Conference on Pattern Recognition*, Tampa, FL, USA: IEEE, 2008: 1–5, https://doi.org/10.1109/ICPR.2008.4761822.

[21] Safwan Wshah, Zhixin Shi, and Venu Govindaraju, "Segmentation of Arabic Handwriting Based on Both Contour and Skeleton Segmentation," in *2009 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain: IEEE, 2009: 793–97, https://doi.org/10.1109/ICDAR.2009.152; Yusra Osman, "Segmentation Algorithm for Arabic Handwritten Text Based on Contour Analysis," in *2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE)*, Khartoum, Sudan: IEEE, 2013: 447–52, https://doi.org/10.1109/ICCEEE.2013.6633980.

[22] Khan et al., "KNN and ANN-Based Recognition of Handwritten Pashto Letters Using Zoning Features."

[23] Abdelhay Zoizou, Arsalane Zarghili, and Ilham Chaker. "A New Hybrid Method for Arabic Multi-Font Text Segmentation, and a Reference Corpus Construction." *Journal of King Saud University—Computer and Information Sciences* 32, no. 5 (June 2020): 576–82, https://doi.org/10.1016/j.jksuci.2018.07.003.

[24] Ullah, "Offline Pashto OCR Using Machine Learning."

[25] Marek Rychlik et al., "Development of a New Image-to-Text Conversion System for Pashto, Farsi and Traditional Chinese," *ArXiv:2005.08650 [Cs]*, May 8, 2020, http://arxiv.org/abs/2005.08650.

[26] "*Glossary of Unicode Terms*," http://www.unicode.org/glossary/.