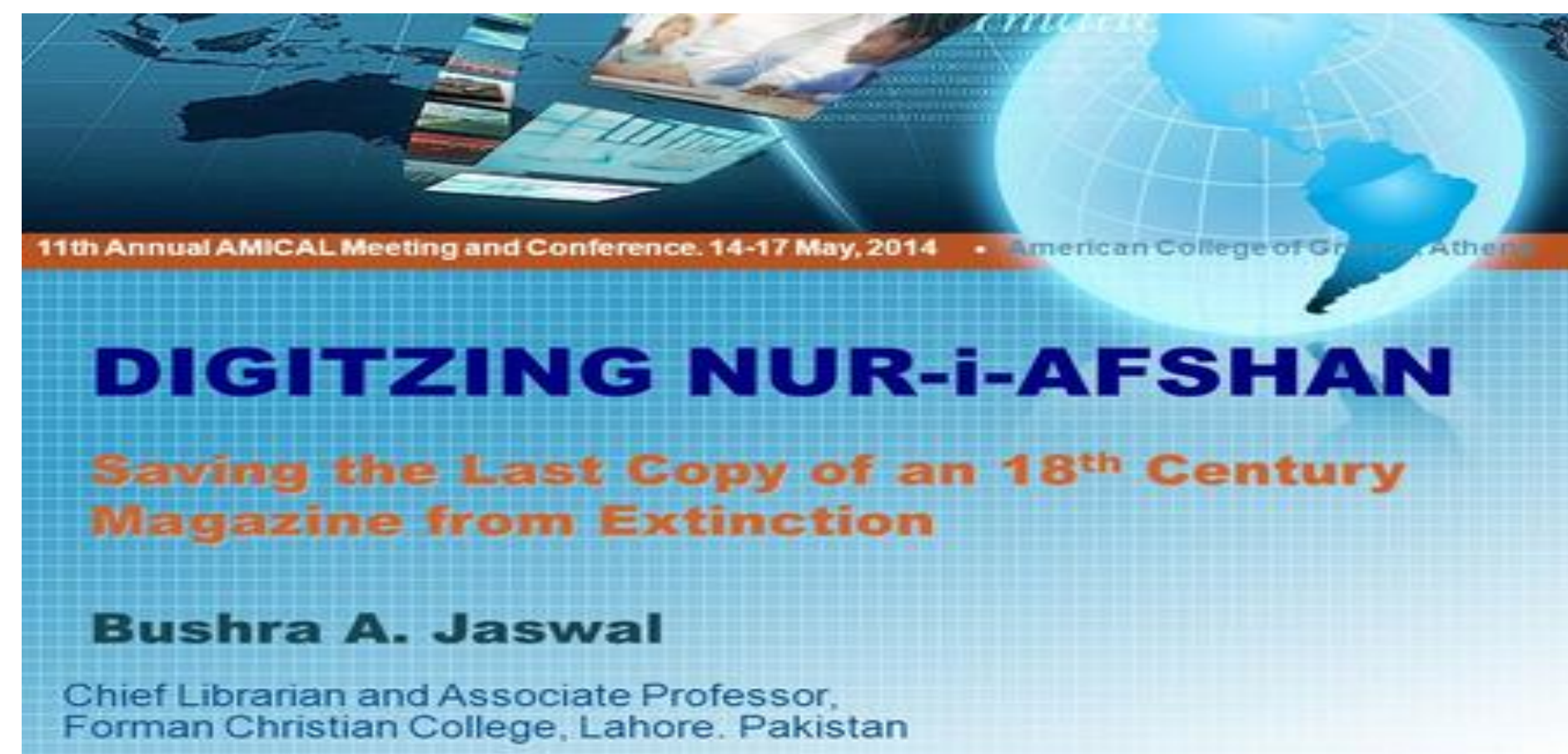


Need for Urdu OCR

The “Dire Need” that initiated a search for a robust Urdu OCR System for Urdu surfaced during the online publishing (as part of Terms of Grant), of the digitized outcome of the ‘Nur-i-Afshan Digitization Project’, presented in 11th AMICAL Conference in 2014 at Athens, Greece.

Metadata Indexing of the digitized corpus of this Urdu Newspaper, which had published from 1873-1944 (71 years), could have been saved, if the whole corpus could be made searchable in Urdu Full-Text. An Urdu OCR System was the solution.



Latin vs Arabic Script OCR

Optical Character Recognition (OCR)

OCR – is a technique used to recognize language characters from hard format on scanned images of printed and handwritten text, by computer systems, to convert into soft, computer readable or machine-encoded and editable text format. OCR converts pictorial data into text and enables the fastest mode of data input.

Latin vs Arabic Script:

Character recognition, is easier with **Latin Scripts** with isolated characters (non-cursive) like **English** as any font of individual characters script is easily recognizable. **Significant advances** have been made for successful OCR for this style.

However, developing OCR for **Arabic like Scripts** including **Urdu**, with cursive characters is still in its initial stage as recognizing the individual character shapes, is more complicated for such intermingled scripts, are **not much worked**

Issues & Challenges of Urdu OCR

Highly complex structure, bi-directionality, and compound in nature, make the Urdu language too complex to obtain accurate OCR results.

Additionally, due to its cursive features, more characters, dots, diacritics, large set of characters for recognition, more base shape group characters, placement of dots, ambiguity between the characters and ligatures with very slight difference, context sensitive shapes, ligatures, noise, skew and fonts in Urdu OCR.

Collaboration for Urdu OCR - Methodology & Procedures

Samples for Urdu OCR & Invitation to collaborate for Pilot Run

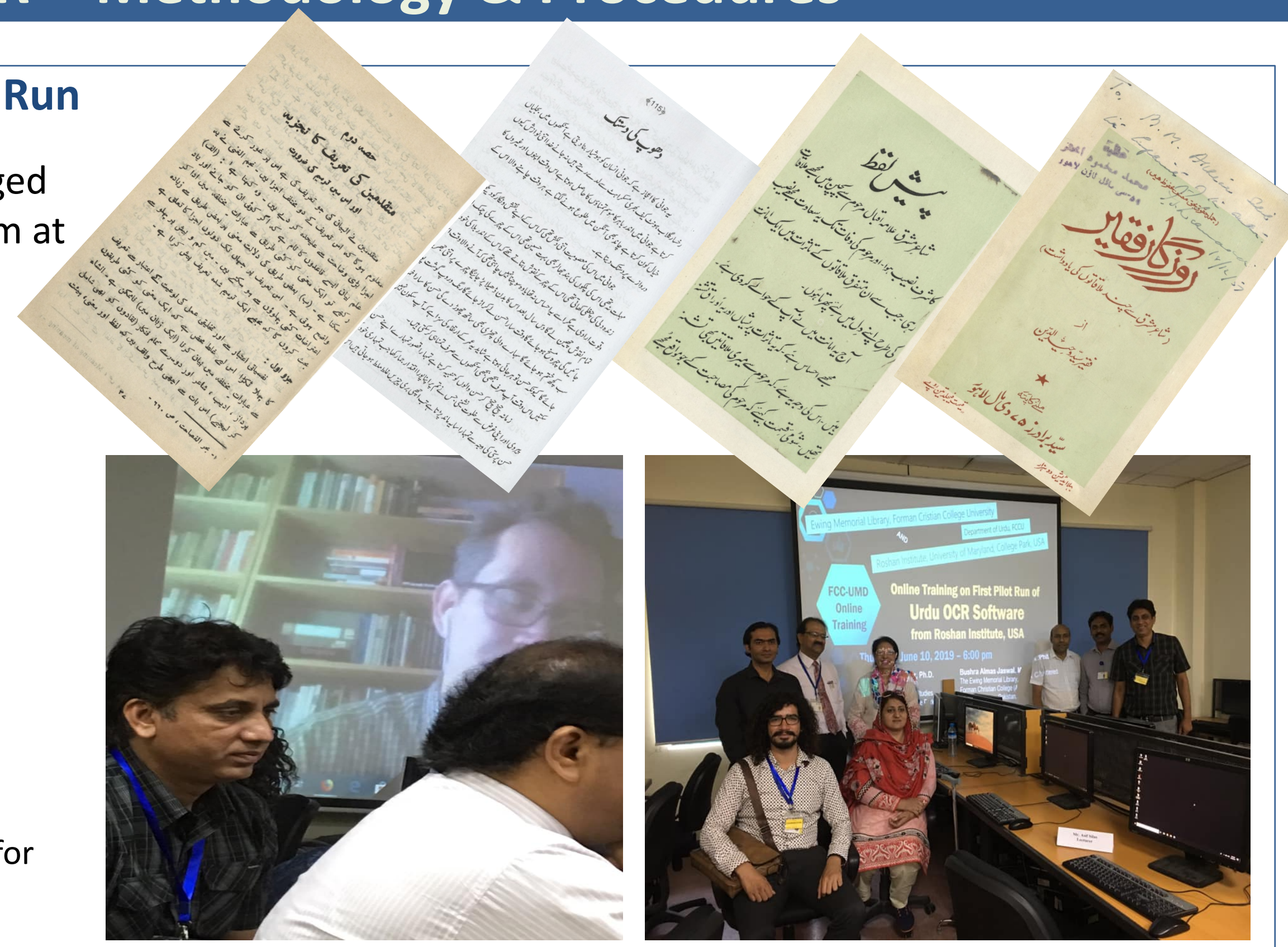
Sample pages of Urdu Books printed in various Urdu Fonts, exchanged for finalization of the first sample run on OCR CorpusBuilder platform at Maryland University.

Training Data Builders from FCC Deptt. of Urdu Language:

1. Professor Tahir Masood
2. Mr. Aneel Samuel, Asstt. Professor
3. Dr. Abida Batool, Associate Professor
4. Mr. Shahzad Ansar, Asstt. Professor
5. Mr. Johnson Rehmat, Lecturer
6. Muhammad Azeem Siddique, Lecturer
7. Mr. Asif Silas, Lecturer

Holding Online Training Workshop

Dr. Matthew Miller held an online training for the new CorpusBuilder Platform for building TrainingData Platform for Urdu OCR System.



Search for Solution

An OCR System for Arabic & Persian? Why not Urdu!



Important New Developments in Arabographic Optical Character Recognition (OCR)

Romanov, M., Miller, M. T., Savant, S. B., & Kiessling, B. (2017). Important New Developments in Arabographic Optical Character Recognition (OCR). *arXiv preprint arXiv:1703.09550*.

The **OpenITI** (Open Islamicate Texts Initiative) team has achieved Optical Character Recognition (OCR) accuracy rates for classical Arabic-script texts in the high nineties, based on tests of seven different Arabic-script texts. These accuracy rates are produced using an open-source OCR software, called **Kraken** thus enabling to make this **Arabic-script OCR technology freely available** to the broader Islamic, Persian, and Arabic Studies communities.

Urdu Language

Urdu is the national language of Pakistan. There are about **60 to 80 million speakers** of Urdu language, which is ranked as **fifth most spoken language with 4.7 percent of the total world population**. Urdu is spoken in South Asia vastly in Pakistan and India. **Huge amount of valuable Urdu literature from history, philosophy to sciences**, is vanishing into useless form because it has not been digitized till now. Mostly **Urdu speakers can only read and write Urdu language** and very rare data is available for them on internet and in digitized form.

Collaboration for Urdu OCR – 2017 till date



This article initiated a search for the **Project KITAB (Knowledge Information Technology and the Arabic Books)** and start of the collaboration for URDU OCR.

2017 July– A note in Comments area of the KITAB Website: “Arabic and Persian OCR System is a commendable initiative. I think, the same system can work for Urdu, just by adding few more alphabets. Would you be interested in adding one more language with little more effort?”

2018 March – Sara Savant– email reply: “Thank you for this note through the KITAB website. It should definitely work for Urdu. We are in the process of finishing a beta version of our OCR pipeline and would be happy with you to test it. If you kindly get back when it is ready sometime in the late summer.

2018 July – Matthew Miller – replied: “If you are still interested in **running a small Urdu OCR pilot on our new CorpusBuilder platform**--- we can get you set up with a testing account on the new system”. It was a YES from us – So after few more emails, this was the start of our collaboration with Roshan Institute, University Of Maryland.

2019 -Andrew W. Mellon Foundation grant for project, "**Open Islamicate Texts Initiative (OpenITI) Arabic-script OCR Catalyst Project**" (**AOCP**)," which will focus on the improvement of Arabic-script OCR and digital text production.

Future Directions

Andrew W. Mellon Foundation Grant for OpenITI:

To focus on the improvement of Arabic-script OCR and digital text production. This author has been invited to a **Workshop, at the University of Maryland, College Park** on January 29th-30th, 2020.

One key component of this larger project is an **Experts Workshop** that will bring together OCR researchers and OCR project teams from the computer science, library, and Digital Humanities domains to share their **latest research, project developments**, and **unmet (individual and institutional) OCR user needs**.

The workshop will include three formats of structured interaction between the invited participants:

- (1) project presentations and Q/A;
- (2) hands-on mini-workshops on particular OCR tools;
- (3) open brainstorm sessions.

The goals of this workshop are to:

- (1) raise awareness among all communities of the diversity of OCR research and OCR project needs;
- (2) test and evaluate existing OCR solutions for their limitations and potential for reuse or expansion; and, most importantly,
- (3) develop project partnerships and plans for strategically and collectively advancing the field of OCR for academic audiences.

This is a major development as well as a new hope for this URDU Language OCR Project which, when successful, has a potential of impacting a much wider segment of Arabic-like languages communities in AMICAL, including Urdu speakers. describe how scholars can benefit by having those books OCR'ed and uploaded to your digital collections.

AMICAL members institutions in Arabic-speaking region may collaborate to work with such initiatives, if interested.

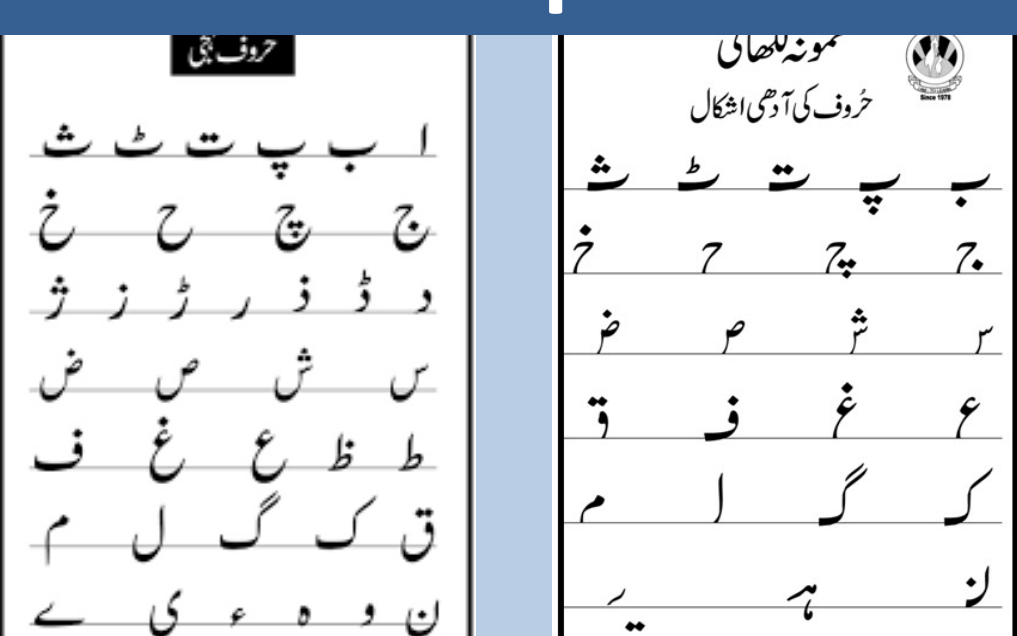
Contact Information

Bushra Almas Jaswal

Chief Librarian & Associate Professor
Ewing Memorial Library, Forman Christian College,
Lahore. Pakistan

Email: bushrajaswal@fccollege.edu.pk
Website: <http://library.fccollege.edu.pk/>
Phone: +92 300 511 7435

Full & half Alphabets



classification of shapes

Mem bers	Class	Members	Class	Members	Class
ا	ا	آ	ا	آ	ا
ب	ب	ب	ب	ب	ب
پ	پ	پ	پ	پ	پ
ت	ت	ت	ت	ت	ت
ث	ث	ث	ث	ث	ث
ج	ج	ج	ج	ج	ج
چ	چ	چ	چ	چ	چ
ح	ح	ح	ح	ح	ح
خ	خ	خ	خ	خ	خ
د	د	د	د	د	د
ذ	ذ	ذ	ذ	ذ	ذ
ر	ر	ر	ر	ر	ر
ز	ز	ز	ز	ز	ز
س	س	س	س	س	س
ش	ش	ش	ش	ش	ش
ص	ص	ص	ص	ص	ص
ض	ض	ض	ض	ض	ض
ط	ط	ط	ط	ط	ط
ظ	ظ	ظ	ظ	ظ	ظ
ف	ف	ف	ف	ف	ف
ق	ق	ق	ق	ق	ق
ک	ک	ک	ک	ک	ک
گ	گ	گ	گ	گ	گ
ل	ل	ل	ل	ل	ل
م	م	م	م	م	م
ن	ن	ن	ن	ن	ن
و	و	و	و	و	و
ی	ی	ی	ی	ی	ی

Diacritics above & below

Short Vowel Name (Diacritics)	Diacritics above or below letter 'ب' (sounds B)	Pronunciation
Fatha	اَ	/ba/
Damma	اُ	/bu/
Kasra	اِ	/bi/
Tanween Alfath	اَاَ	/baa/
Tanween Aldam	اُاُ	/buu/
Tanween Alkahr	اِاِ	/bii/
Sokan	ا	/bi/

Features of URDU writing

Techniques of Urdu OCR

3,4 & 5 letter ligatures

3-letter ligatures	چا ب پ ت ش
4-letter ligatures	چا ب پ ت ش
5-letter ligatures	چا ب پ ت ش

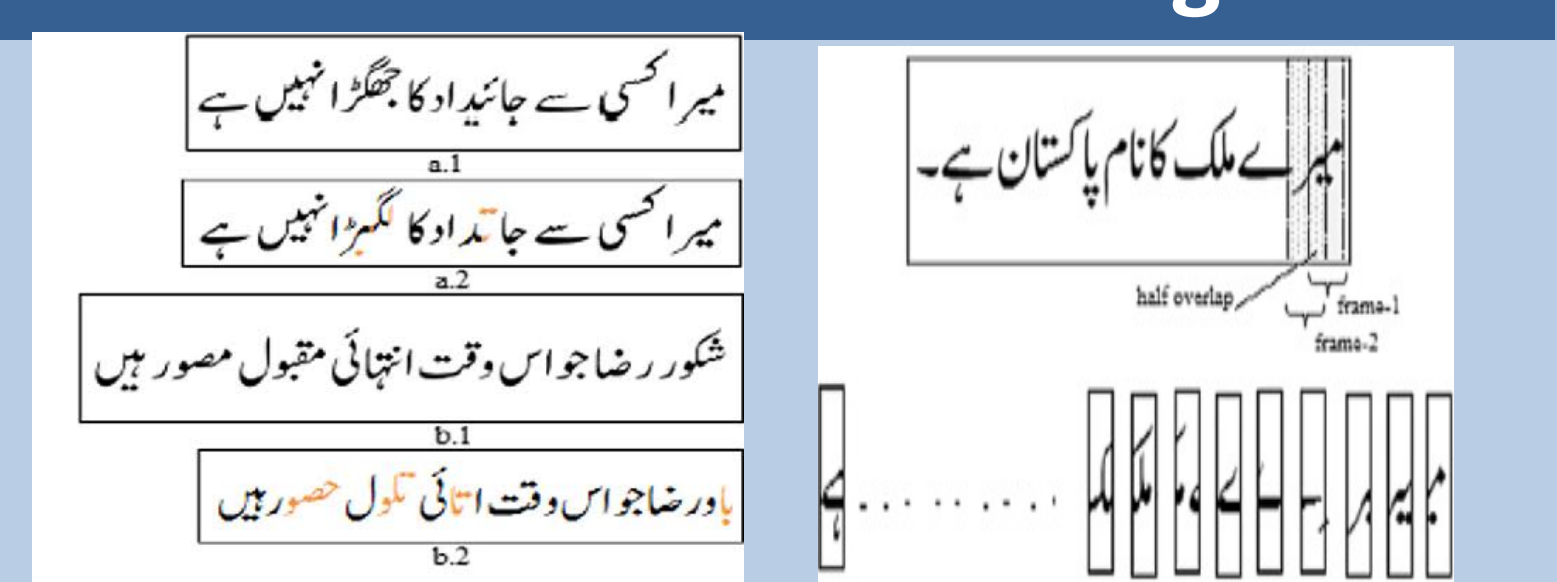
Segmentation

Segmentation-free OCR for Arabic scripts

ل + م + ع + ط + ی = لمعطی

- Nastalique writing: Classify ligatures instead of individual characters.
- Over 20,000 valid ligatures in the Urdu language.
- Issues in the preprocessing, with difficulty in feature extraction & classification.

Multi-dimensional recognition





Abstract

This session describes an ongoing collaborative project between the Ewing Memorial Library at Forman Christian College University, in Lahore, Pakistan and the Roshan Institute for Persian Studies at the University of Maryland in College Park, USA, for development of an Optical Character Recognition (OCR) Software for Urdu language.

The Roshan Institute had successfully developed an OCR System for Persian Language. As the Arabic, Persian and Urdu languages share the similar script and alphabets, this author contacted them with the idea of being a strong chance that the same system, with some additions and changes, could be adapted to the Urdu Language OCR.

Matthew Thomas Miller, Associate Director responded positively. The emails communications started more meaningfully in July 2018. Exchanging various samples of digitized Urdu script further supported the possibility of a test-run on the first OCR pilot corpus-builder platform. Due to the required Urdu Language expertise, it was decided, that the training data be developed by local language experts.

By engaging the faculty from Department of Urdu at FCCU, the project is now ready to embark on the task. The scan from selected Urdu Books are ready for upload to Roshan Institute platform. An online training session has been by Dr. Matthew Thomas Miller has been organized for the faculty volunteers and by the time of this presentation next year, I hope that the project will move further and there would be more development to report.

This is the first Urdu OCR initiative which is being planned to be working for all Fonts and writing Styles of Urdu Language. If successful, it may largely impact the information practitioners in library, research, journalism and publishing sectors.

This session has multiple dimensions of alignment with the topics in AMICAL criteria. It directly relates with the topic of Digital collections but it is also relevant with Digital pedagogy & digital humanities and also falls under Interdisciplinary and inter-institutional collaborations

The session may be interesting for the AMICAL Institutions working with digitization of documents in Urdu-like languages. It will also help the authors to learn about similar language OCR needs or to learn about similar efforts going on among AMICAL Institutions.

Interested AMICAL members may create a discussion group for sharing information or working together.

Bushra Almas Jaswal

Chief Librarian & Associate Professor

Ewing Memorial Library, Forman Christian College, Lahore. Pakistan

Email: bushrajaswal@fccollege.edu.pk

Phone: +92 300 511 7435

Muhammad Imtiaz Bhatti,

Metadata Librarian,

Ewing Memorial Library, Forman Christian College,

Lahore. Pakistan

Email: imtiazahmed@fccollege.edu.pk