Framework of Urdu Nastalique Optical Character Recognition System

*Qurat-ul-Ain Akram Sarmad Hussain Farah Adeeba Shafiq-ur-Rehman Mehreen Saeed
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
**ainie.akram@kics.edu.pk, firstname.lastname@kics.edu.pk*

Abstract

The development of Urdu Nastalique Optical Character Recognition (OCR) is a challenging task due to the cursive nature of Urdu, complexities of Nastalique writing style and layouts of Urdu document images. In this paper, the framework of Urdu Nastalique OCR is presented. The presented system supports the recognition of Urdu Nastalique document images having font size between 14 to 44. The system has 86.15% ligature recognition accuracy tested on 224 document images.

1. Introduction

Urdu belongs to Arabic script which is cursive in nature. Urdu has an extended character set shown in Figure 1, and additional aerab which are normally used for pronunciation [1]. One or more characters of Urdu are joined together to form ligature [2]. A ligature has a base stroke called RASM or main body and secondary strokes called IJAM or diacritics. Based on the shape similarity of RASM, Urdu ligatures are divided into different classes. Different ligatures are joined together to form Urdu words. In Urdu, spaces are not properly used to define Urdu word boundary [3, 4].

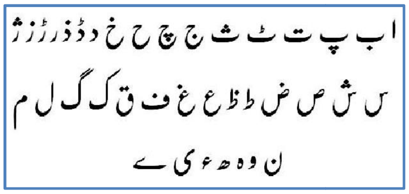
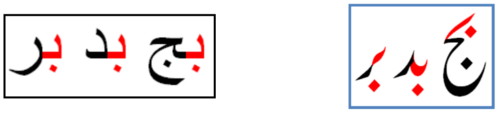


Figure 1. Urdu character set

Nastalique writing style is normally used to write Urdu books, magazines and newspapers. Nastalique is written diagonally, which results in vertical overlapping of characters and ligatures. This characteristic adds complexity in Urdu document image segmentation. Naskh writing style which is used to write Arabic text, has four unique shapes for a character. Unlike Naskh, Nastalique has contextual character shaping [5] as can be seen in Figure 2. Some cases of contextual shaping are highlighted with red color.



(a) Character shaping in Naskh writing Style

(b) Contextual Character shaping in Nastalique writing Style

Figure 2. Contextual character shaping of ب character highlighted with red color

In Nastalique, rules for the placement of Nuqtas and diacritics are complex, which are based on the contextual existence of characters in a ligature. This characteristic also adds additional complexity for text image segmentation especially during marks association with the respective character/ligatures. In Nastalique, some of the characters and diacritics have same shapes but are different in size. The examples of diacritics and main body confusions are listed in Table 1.

Table 1. Diacritics and main bodies confusion [6]

Ligature	Main body shape	Confusing diacritic	Diacritic shape (enlarged)
بر		Diacritic of ئ	
با		Diacritic of	
ط		Diacritic of ٹ	

A ligature has variation of thick-thin strokes which introduces complexity in the pre-processing module especially during binarization of Urdu document images. The examples of ligatures indicating thick-thin transitions are given in Figure 3.

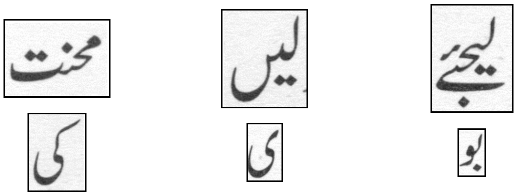


Figure 3. Examples of thick-thin stroke variation across characters in a ligature

2. Literature Review

Due to the complexities of Nastalique, limited effort has been carried out for the development of complete Optical Character Recognition (OCR) system for Nastalique writing style. In this section current state of the art for the development of Urdu OCR is discussed. Normally, OCR has three modules; (1) preprocessing, (2) classification and recognition, and (3) post-processing.

Preprocessing module deals with the processing of an input image to improve its quality and to segment image into different areas. The relevant information from these areas is extracted which is used in classification and recognition, and post-processing modules. The binarization system for Urdu document images is developed by modifying the existing binarization algorithm to address the Nastalique complexities [7]. The evaluation of binarization algorithm for Urdu document images is also devised in this study. Shafait et al. [8] apply some of the existing pre-processing techniques on Urdu Nastalique document images to segment the page into columns and text lines. The system is tested on 25 images scanned from different magazines, poetry books, text books, digest and newspapers. The reported accuracies of the system are 91.45% , 92.31 %, 80.63%, 90.07% and 72.16% for text books, poetry books, digests, magazines and newspapers respectively. The projection profile method is also used to segment the Urdu document image into text lines [9]. In addition, some heuristics are also used to improve the line segmentation results. The main body and diacritics of ligatures are extracted using bounding box information, and association of diacritics with the respective main body is done using overlapping information. The reported line segmentation accuracy is 100% tested on 20 images scanned from three poetry books. The diacritics association accuracy is 94% tested on synthesized data of 3,655 ligatures at 36 font size. Bukhari et al. [10] present layout analysis of Arabic and Urdu document images having multiple layouts. The existing systems for other languages are tweaked for segmentation of Arabic and Urdu document images. The system is tested on 25 Arabic and 20 Urdu document images. The reported text and non-text segmentation accuracy of the system is 99% for Arabic test data. The text line extraction accuracies are 96% and above 92% for Arabic and Urdu document images respectively.

Classification and Recognition module has two phases. The first phase called training phase deals with the classification of character/ligature shapes into different classes based on the shape similarity. The features of each class are extracted and used as input to a classifier for training. In the recognition phase, the features of input shape are computed and recognized using the trained classifier. For the recognition of cursive script, the classification and recognition module is developed using two approaches; (1) Ligature-based classification and recognition, and (2) Segmentation-based classification and recognition.

In ligature-based classification and recognition, the ligature as a whole is used for the classification and recognition. The ligature-based recognition of Nastalique main bodies is done using structural features which are classified using neural network [11]. Sabbour and Shafait [12] use shape context features of contours of the main body and diacritics for the recognition of Urdu and Arabic text. The reported accuracy of the system for Urdu is 91% tested on synthesized data. Javed et al. [13] divide ligature stroke into smaller windows and extract features for recognition using HMMs as classifier. The system is tested on synthesized data at 36 font size and has 92% recognition accuracy. Lehal and Rana [14] perform different experiments on different feature sets and classifiers for the recognition of Nastalique ligatures. The test data contains 4,380 images of 2,190 main body classes and 1,700 images of 17 diacritics classes. The DCTs as feature set with SVM performs well among others features and classifiers. The system has 98.01% main body recognition accuracy and 99.91% diacritics recognition accuracy. The ligature-based classification and recognition of Nastalique main bodies using Tesseract is also reported [6]. The Tesseract engine is modified to improve the main body recognition accuracy. The reported accuracy of the system is 97.87% for 14 font size and 97.71% for 16 font size, tested on separate test data of 22,125 instances of 1,475 main body classes for 14 and 16 font sizes.

The segmentation-based classification and recognition system deals with the segmentation of main body stroke into smaller parts which can be characters. The segments are then used for the classification and recognition. A segmentation-based technique for the recognition of handwritten Nastalique text is presented by Safabakhsh and Abidi [15]. The Fourier descriptor, structural and discrete features are extracted from the segmented primitives and are classified using continuous-density variable-duration HMMs. The reported accuracy of the system is 96.8%. Javed and Hussain [16] segment the ligature into smaller primitives using branch points information of thin main body stroke. The DCTs features of the segments are computed and classified using HMMs. The sequence of recognized primitives is used to recognize the ligature. The system is tested on synthesized data at 36 font size, which contains 1,692 high frequency ligatures of six character classes. The reported main body recognition accuracy of the system is 92.73%. This approach is extended by Muaz [17] for the recognition of ligatures of all 21 character classes. The recognition accuracy of the system is 92.19% tested on 2,494 ligatures synthesized at 36 font size. The bidirectional LSTM networks are used for the recognition of printed Nastalique text [18]. The synthesized dataset having text lines is used in this approach. The pixel level information extracted from the sliding window is used as a feature. This system has 94.85% character recognition accuracy tested on 2,003 text line images. Rashid et al. [19] develop the system for the recognition of multi script documents using Convolutional Neural Networks (CNNs). The reported script recognition accuracy of the system is above 95% tested on Greek-Latin, Arabic-Latin and Antiqua-Fraktur document images. Naz et al. [20] evaluate state of the art techniques of preprocessing, feature extraction and classification and recognition for Urdu, Pashto and Sindhi document images having Nastalique and Naskh writing styles.

The post-processing phase deals with the formation of words and sentences using recognized characters /ligatures sequences. This module deals with the word segmentation, spell checker and POS tagger etc. sub-modules, and outputs the correct sequence of words to form sentences. In Urdu, development of the word segmentation system is challenging due to the inconsistent use of space. Durrani and Hussain [21] develop a rule-based word segmentation system which automatically defines the word boundaries of the input Urdu text. A statistical word segmentation system for Urdu text corpus is also developed [4]. The ligature N-grams and word N-grams are computed from the corpus to statistically compute the best sequence of words from the sequence of ligatures. The system has 96% word formation and 67% sentence formation accuracy.

3. Methodology

In this paper, the framework of Urdu Nastalique OCR system is presented. During books survey, it has been analyzed that the font size of text books ranges from 14 to 44 where 14 and 16 font sizes are used for normal text and remaining font sizes are used in headings of Urdu text books. The presented OCR supports the recognition of Urdu document images scanned from different books and magazines having font sizes between 14 to 44. The architecture of Urdu Nastalique OCR is illustrated in Figure 4. Urdu Nastalique OCR has three main modules; (1) Preprocessing, (2) Classification and Recognition, and (3) Post-processing modules. To cover the range of 14 to 44 font sizes, four different recognizers are developed at 14, 16, 22 and 36 font sizes. To support the reasonable accuracy of normal text recognition, the recognizers at 14 and 16 font sizes are developed separately. The remaining font sizes normally appear in headings, therefore to cover this range one recognizer is developed at 22 font size and the other recognizer is developed at 36 font size. The details of each module are discussed in subsequent sections.

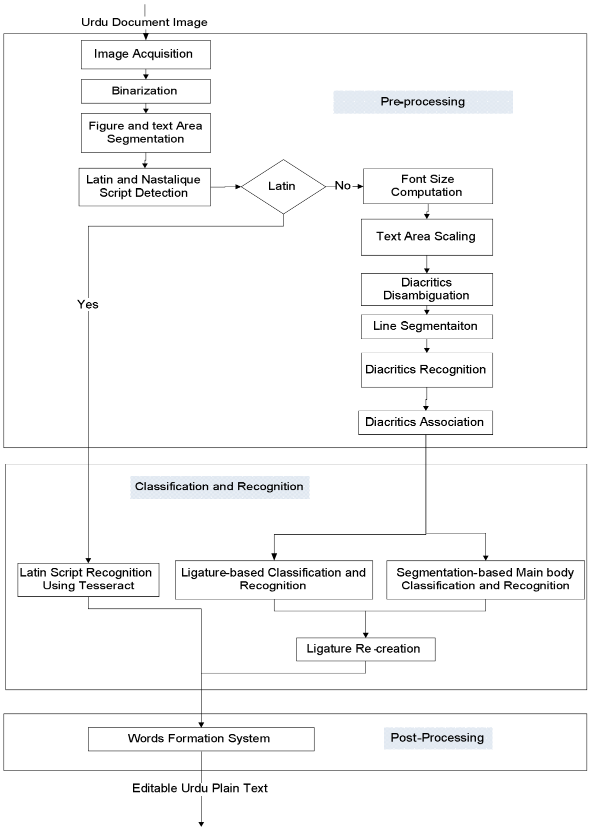

Urdu Document Image
Image Acquisition
Binarization
Pre-processing
Figure and text Area Segmentation
Latin and Nastalique Script Detection
Latin
No
Font Size Computation
Text Area Scaling
Yes
Diacritics Disambiguation
Line Segmentaiton
Diacritics Recognition
Diacritics Association
Classification and Recognition
Latin Script Recognition Using Tesseract
Ligature-based Classification and Recognition
Segmentation-based Main body Classification and Recognition
Ligature Re-creation
Words Formation System
Post-Processing
Editable Urdu Plain Text

Figure 4. Architecture diagram of Urdu Nastalique OCR

3.1. Preprocessing

In Preprocessing module, the binarization of Urdu document images is performed using [7]. The layout extraction is performed which segments the Urdu document images into figure and text areas. The extracted text areas are then sequenced into column(s) according to the reading order. The output of the layout extraction is given in Figure5, the figure areas are marked with red rectangle and text areas are marked with green rectangle.

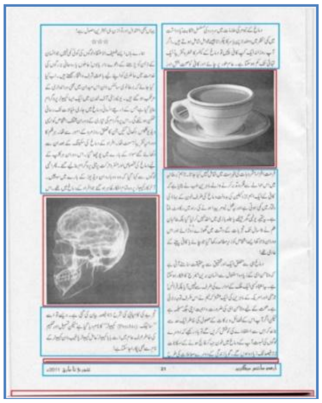
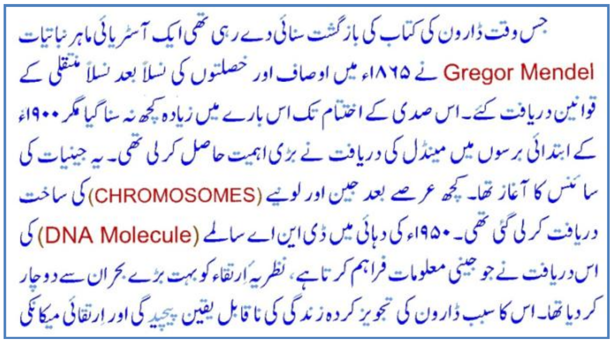


Figure 5. Output of image segmentation into figures and text areas

Each text area of the document image is processed to mark the Latin and Nastalique text. Script detection sub-module processes each connected component and marks script identity either Latin or Nastalique. Figure 6 shows the sample output of script detection system, the Nastalique script is marked with blue color and Latin script is marked with red color.

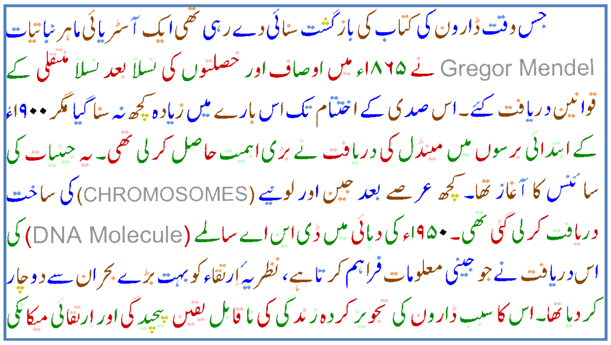


جس وقت ڈارون کی کتاب کی بازگشت سنائی دے رہی تھی ایک آسٹریائی ماہر نباتیات
Gregor Mendel نے ۱۸۶۵ء میں اوصاف اور خصلتوں کی نسلاً بعد نسلاً منتقلی کے
قوانین دریافت کیے۔ اس صدی کے اختتام تک اس بارے میں زیادہ کچھ نہ سنا گیا مگر ۱۹۰۰ء
کے ابتدائی برسوں میں مینڈل کی دریافت نے بڑی اہمیت حاصل کرلی تھی۔ یہ جینیات کی
سائنس کا آغاز تھا۔ کچھ عرصے بعد جین اور لونیے (CHROMOSOMES) کی ساخت
دریافت کرلی گئی تھی۔ ۱۹۵۰ء کی دہائی میں ڈی این اے سالمے (DNA Molecule) کی
اس دریافت نے جو جینی معلومات فراہم کرتا ہے، نظریۂ ارتقاء کو بہت بڑے بحران سے دوچار
کردیا تھا۔ اس کا سبب ڈارون کی تجویز کردہ زندگی کی ناقابل یقین پیچیدگی اور ارتقائی میکانکی

Figure 6. Output of script detection

The Nastalique text is further processed to compute the font size of the text area. Based on the computed font size, the text area is resized to the nearest font size on which pivot line segmentation and recognizer are developed. The text areas having font size between 18 to 20 are scaled up to 22 font size and text areas of 24 to 28 font sizes are scaled down to 22 font size. In the same way, the text areas of 30 to 34 font sizes are scaled up to 36 font size and 38 to 44 font sizes are scaled down to 36 font size.

The connected components of the resized text areas are disambiguated as diacritics or main bodies. The main bodies are used to form text line. As some of the diacritics are confused in shape with some of the main bodies, but are different in sizes (see Table 1). Therefore, separate recognizers of diacritics and main bodies are developed. For better results of diacritics association with respective main body, the diacritics recognition is performed in pre-processing module. Correct diacritics association eventually affects the ligature recognition accuracy. After diacritics recognition, the association of diacritics with the respective main body is done. In addition, the positional information of the diacritics with respect to main body such as above, below or middle of the main body is also computed. This information is used in ligature string creation sub-module of classification and recognition. The diacritics association output is shown in Figure 7. The blue and brown colors are used to show the alternating ligatures in one line and red and green colors are used to show the alternating ligatures in next text line. The dark color indicates the main body and light color indicates the associated diacritics of the respective main body. The Latin script is not processed to form the text lines. However, the positional information in the respective text line is also maintained so that its recognized text can be output in proper location in the text line. The Latin text in Figure 7 is highlighted with gray color.



جس وقت ڈارون کی کتاب کی بازگشت سنائی دے رہی تھی ایک آسٹریائی ماہر نباتیات
Gregor Mendel نے ۱۸۶۵ء میں اوصاف اور خصلتوں کی نسلاً بعد نسلاً منتقلی کے
قوانین دریافت کیے۔ اس صدی کے اختتام تک اس بارے میں زیادہ کچھ نہ سنا گیا مگر ۱۹۰۰ء
کے ابتدائی برسوں میں مینڈل کی دریافت نے بڑی اہمیت حاصل کرلی تھی۔ یہ جینیات کی
سائنس کا آغاز تھا۔ کچھ عرصے بعد جین اور لونیے (CHROMOSOMES) کی ساخت
دریافت کرلی گئی تھی۔ ۱۹۵۰ء کی دہائی میں ڈی این اے سالمے (DNA Molecule) کی
اس دریافت نے جو جینی معلومات فراہم کرتا ہے، نظریۂ ارتقاء کو بہت بڑے بحران سے دوچار
کردیا تھا۔ اس کا سبب ڈارون کی تجویز کردہ زندگی کی ناقابل یقین پیچیدگی اور ارتقائی میکانکی

Figure 7. Output of diacritics association

3.2. Classification and Recognition

Open source Tesseract engine [22] for Latin is used for the recognition of marked Latin script. The recognized output of the Tesseract is passed to the post-processing module.

The classification and recognition module mainly deals with recognition of Urdu main bodies. In the training phase unique main bodies are classified into classes using the RASM class information. Two different classifiers are developed. Ligature-based classification and recognition using Tesseract and (2) Segmentation-based classification and recognition using HMMs.

In ligature based classification and recognition module, the main body as a whole is used for recognition. The modified Tesseract engine for the recognition of Nastalique main bodies is used for this purpose. The details can be found in [6]. The segmentation-based classification and recognition module deals with the segmentation of main body into constituent characters. Here the DCT coefficients as features and HMMs as classifier are used for recognition. Both classifiers outputs the recognized ranked options of main body identifiers against a single main body image.

After recognition, the recognized main body and diacritics of respective ligature are used to recognize the ligature string. For this purpose, a lookup table is used which contains the information pertaining to a recognized main body and its corresponding associated diacritics. In addition, the diacritics positional information is also maintained in the lookup table. The recognized diacritics and positional information are computed in pre-processing module. Therefore recognized diacritics, positional information of diacritics and recognized main body are used to recognize the respective ligature string using a lookup table. Against each recognized option in the ranked list of a main body, the respective ligature string is recognized from the lookup table. This module generates recognized ranked ligatures list against a ligature image.

3.3. Post-processing

The output of the classification and recognition module is the recognized sequence of ligatures. Each ligature has a ranked list. The word segmentation system converts the sequences of ligatures into best sequence of words using modified statistical model of [4]. All combinations of ligatures ranked list are formulated and given to the word segmentation system to generate the best sequence of words of a sentence.

The statistical word segmentation system is developed using language model of ligature N-grams and word N-grams reported in [23]. In post-processing, the recognized Latin text is also inserted between Urdu words according to the position computed in pre-processing module. The output of word segmentation system on sample input is given in Figure 8.

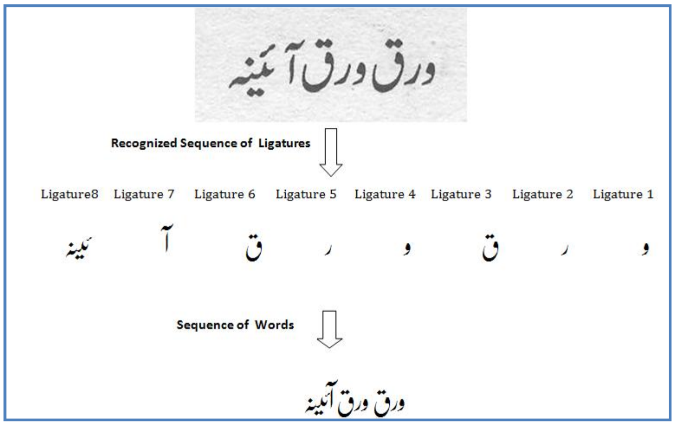


Figure 8. Output of word formation system

4. Testing and Results

For the testing of Urdu Nastalique OCR system, a test data of different page layouts having one column is prepared. The layouts have header, footer, mixture of font sizes, figure in text, etc. Some examples of layouts of test data are given in Figure 9.

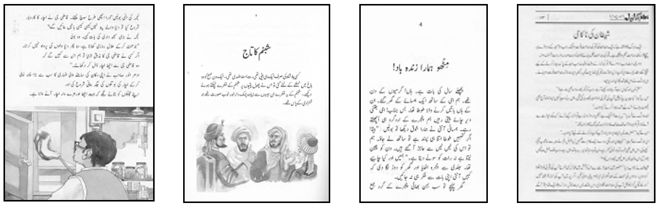


Figure 9. Examples of test data layouts

In addition, page having text in different font sizes from 14 to 44 font sizes has also been added in the test data. The synthesized document images at missing font size (does not appear in document images of books) are prepared to test the recognition accuracy of the OCR. Test data contains a total of 224 document images. The document images having normal font sizes contain approximately 700 ligatures per page. The number of ligatures of document images goes down for larger font sizes. The test data of 224 document images contains 371 on average ligatures per page. The desired ligature string against each ligature of document images is tagged. In addition, the broken main body and special ligature (written in different font style) are also marked. The font wise number of ligatures along with the count of broken and special ligatures in the test data are listed in Table 2.

The accuracy of each font size is computed at three levels; (1) classification and recognition (C&R) module, (2) post-processing (PP) module and (3) End-to-End system. In classification and recognition module accuracy, the ligature is marked as correct if the desired ligature string is found in the ligature's recognized ranked list. For post-processing (PP) module accuracy, the word segmentation accuracy is computed irrespective of C&R module errors. Therefore the PP module accuracy is computed as the total number of correct ligatures found in ranked lists versus total number of ligatures ranked at the top by the word segmentation system.

Table 2. Font wise ligature information of test data

Font Size	Total Urdu Ligatures	Broken Ligatures	Special Ligatures
14	30,631	506	652
16	14,647	51	344
18	12,108	151	220
20	9,445	23	1,066
22	6,837	13	16
24	698	6	74
26	2,350	2	6
28	113	1	0
32	208	0	23
34	1,455	0	0
36	357	1	222
38	1,248	3	0
40	126	6	13
42	1,029	0	0
44	899	0	0

In addition, End-to-End system accuracy is also computed which computes the accuracy in terms of total number of ligatures of input document image versus total number of correct ligatures ranked at top by the word segmentation system. The font wise ligature recognition accuracies are given in Table 3. Urdu OCR gives 90.10%, 95.78% and 86.15% per page ligature recognition accuracy for C&R module , PP module and End-to-End system respectively.

Table 3. Font wise ligature recognition accuracy

Font	Total Urdu Ligatures	C&R Module Accuracy %	PP Module Accuracy %	End-to-End System Accuracy %
14	30,631	91.87	94.26	86.60
16	14,647	93.31	95.80	89.39
18	12,108	91.85	94.97	87.23
20	9,445	86.37	94.03	81.22
22	6,837	96.90	96.44	93.45
24	698	85.24	96.30	82.09
26	2,350	95.53	97.02	92.68
28	113	88.50	96.00	84.96
32	208	89.42	93.01	83.17
34	1,455	97.46	98.24	95.74
36	357	63.59	97.36	61.90
38	1,248	98.16	98.78	96.96
40	126	50.79	100.00	50.79
42	1,029	97.67	98.81	96.50
44	899	94.10	97.99	92.21

5. Discussion

In this paper, the framework of Urdu Nastalique OCR to support the recognition of Urdu document images for 14 to 44 font sizes is reported. The system is tested on 224 document images and ligature recognition accuracies are reported for each desired font size. The complete process from pre-processing, classification and recognition, to post-processing is performed on each document image. The ligature recognition accuracy is computed using the tagged ligature string information. The module-wise accuracies are reported for each font size in Table 3. recognition accuracy of normal font sizes are stable. The main reason is the availability of sufficient data of normal text. As the larger font sizes appear in heading therefore the training and testing data is not sufficient for most of the font sizes. The drop in the recognition accuracy, especially for 36 and 40 is mainly due to unavailability of a large number of example images. In addition, currently Urdu OCR does not handle broken and special ligatures which also affect the ligature recognition accuracy. The results show that developed Urdu OCR can be used to port published Urdu content online with minimal editing effort.

6. Conclusion

In this paper, a framework of Urdu Nastalique OCR is discussed. Initially, the layout of the Urdu document images and complexities of Nastalique writing style is analyzed to finalize the framework. Each sub-module is tested and matured separately on data extracted from document images at each font size. After finalization of each sub-module separately, these are integrated in the OCR framework. The testing and maturation pass of the integrated OCR system is carried out to further improve the document recognition accuracy. The system has per page ligature recognition accuracy as 90.10% for C&R module, 95.78% for PP module and 86.15% for End-to-End system. The broken and special ligatures caused misrecognition which will be resolved in the future. The presented Urdu Nastalique OCR is online available at: www. UrduOCR.net

7. Acknowledgements

This work has been conducted through the project, Urdu Nastalique OCR supported through a research grant from ICTRnD Fund, Pakistan.

8. References

[1] S. Hussain, "Letter to Sound Rules for Urdu Text to Speech System," in Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Switzerland, 2004.

[2] M. Davis, "Unicode Text Segmentation," Addison-Wesley Professional, 2013.

[3] S. Hussain, "www.LICT4D.asia/Fonts/Nafees_ Nastalique," in 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore, 2003.

[4] M. Akram and S. Hussain, "Word Segmentation for Urdu OCR System," in 8th Workshop on Asian Language Resources, COLING2010, Beijing, China., 2010.

[5] A. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," in International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE), 2006.

[6] Q. Akram, S. Hussain, A. Niazi, U. Anjum and F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique," in 11th IAPR Workshop on Document Analysis Systems (DAS 14), Tours, France, 2014.

[7] M. Naz, Q. Akram and S. Hussain, "Binarization and its Evaluation for Urdu Nastalique Document Images," in The 16th International Multi Topic Conference (INMIC), Lahore, 2013.

[8] F. Shafait, D. Keysers and T. M. Breuel, "Layout Analysis of Urdu Document Images," in INMIC'06. IEEE, 2006.

[9] S. T. Javed and S. Hussain, "Improving Nastalique Specific Pre-Recognition Process for Urdu OCR," in 13th IEEE International Multitopic Conference 2009 (INMIC 2009), Islamabad, Pakistan, 2009.

[10] S. S. Bukhari, F. Shafait and T. M. Breuel, "High Performance Layout Analysis of Arabic and Urdu Document Images," in International Conference on Document Analysis and Recognition, 2011.

[11] Z. Shah and F. Saleem, "Ligature Based Optical Character Recognition of Urdu, Nastaleeq Font," in International Multi Topic Conference, Karachi, Pakistan, 2002.

[12] N. Sabbour and F. Shafait, "A Segmentaitotn Free Approach to Arabic and Urdu OCR," in SPIE, Volume 8658, 2013.

[13] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Mohsin, "Segmentation Free Nastalique Urdu OCR," World Academy of Science, 2010.

[14] G. S. Lehal and A. Rana, "Recognition of Nastalique Urdu Ligatures," in 4th International Workshop on Multilingual OCR (MOCR '13), New York, NY, USA, 2013.

[15] R. Safabakhsh and P. Abidi, "Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM," The Arabian Journal for Science and Engineering, 2005.

[16] S. T. Javed and S. Hussain, "Segmentation Based Urdu Nastalique OCR," in 18th Iberoamerican Congress on Pattern Recognition (CIARP 2013), Havana CUBA, 2013.

[17] A. Muaz, "Urdu Optical Character Recognition System," Unpublished, MS Thesis Report, National

University of Computer and Emerging Sciences , Lahore, 2010.

[18] A. Hasan, S. B. Ahmed, S. F. Rashid, F. Shafait and T. M. Breuel, "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," in International Conference on Document Analysis and Recognition, 2013.

[19] S. F. Rashid, F. Shafait and T. M. Breuel, "Discriminative learning for script recognition," in 17th IEEE International Conference on Image Processing, Hong Kong, 2010.

[20] S. Naz, K. Hayat, M. I. Razzak, M. W. Anwar, S. A. Madani and S. U. Khan, "The optical character recognition of Urdu-like cursive scripts," Pattern Recognition, pp. 1229-1248, 2014.

[21] N. Durrani and S. Hussain, "Urdu Word Segmentation," in 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Los Angeles, US, 2010.

[22] R. Smith, D. Antonova and D.-S. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in International Workshop on Multilingual OCR, Barcelona, Spain, 2009.

[23] F. Adeeba, Q. Akram, H. Khalid and S. Hussain, "CLE Urdu books N-Grams," in Conference on Language and Technology 2014(CLT14), Karachi, 2014.