Ligature-based font size independent OCR for Noori Nastalique writing style

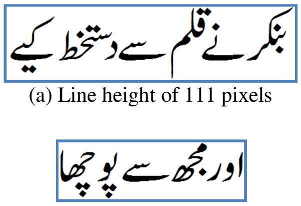
Qurat ul Ain Akram Sarmad Hussain
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
ainie.akram@kics.edu.pk, sarmad.hussain@kics.edu.pk

Abstract— **In this paper, a font size independent Optical Character Recognition (OCR) system for Urdu document images is presented. Urdu documents are written using Noori Nastalique writing style with different font sizes of normal text and headings. Most of current state of the art techniques of Urdu OCRs support recognition of text having single font size. The presented study deals with the recognition of Nastalique text having 14 to 28 font sizes. Three recognizers at three font sizes(called pivot) including 14, 16 and 22 are developed. Urdu document images having remaining font sizes such as 18, 20, 24, 26 and 28 are resized to the nearest pivot font size using Nearest Neighboring interpolation technique so that it can be recognized. The detailed analysis has been carried out to compute optimal scaling factor of each font size to improve recognition results. It has been observed that recognizers perform better at resized images by applying optimal scaling factors instead of simple computed scaling factors. The system is developed and matured on 1,965 main body classes covering 59,974 high frequent Urdu words. After maturation, system has 97.20%, 97.08%, 95.13%, 95.65%, 96.26%, 96.52%, 95.78%, 96.38%, 96.66% main body recognition accuracy for 14, 16, 18, 20, 24, 26, 28 font sizes respectively.**

Keywords- ***Urdu, Noori Nastalique, Font Size Independent, Image Resizing, Ligature, Main Body, Optical Character Recognition (OCR)***

I. INTRODUCTION

Tremendous progress in the development of Information Communication Technologies (ICTs) gives opportunities to address the need of porting published local content online. An OCR is an assistive technology which accelerates the process of porting local content online for users to access the desire information in their languages through laptops, tablets and mobile devices [1, 2]. Significant amount of Urdu content is published in the form of books, magazines and newspapers using Noori Nastalique writing style. Usually different font sizes are used to write normal text and headings therefore to port such content online, a font size independent OCR needs to develop. In Nastalique writing style, characters are joined diagonally to form the ligature [3]. A ligature image can be divided into two parts; (1) main stroke also called RASM and (2) secondary stroke(s) also called IJAM. Based on the shape similarity of Urdu ligature RASM, ligature are classified into different classes [4]. Development of OCR for Nastalique writing style is challenging due to complexities such as multiple contextual shapes of a character, character and ligature overlapping, thick thin transitions of character strokes, complex dots and marks placement rules etc. [4]. In addition, diagonal growing ligature length also adds complexity to find the line height of the text line for font size computation and size normalization. The text lines of the same font size can have varied line height due to length of ligatures in line as can be seen in Fig.1. In Latin script, x-height is normally used to find font size of text line and accordingly line normalization is applied to have all the text line images at the same size, but such techniques cannot be applied on Nastalique text due to the complexities discussed above.


بنکر نے قلم سے دستخط کیے
اور مجھ سے پوچھا

(a) Line height of 111 pixels

(b) Line height of 87 pixels

Fig. 1. Same font size text having varation in line height

In this paper, a framework to develop font size independent OCR for the recognition of 14 to 28 font sized Urdu Nastalique text is presented. Instead of developing recognizer at each font size, three recognizers at three font sizes (called pivot) including 14, 16 and 22 are developed. The document images having remaining font sizes such as 18, 20, 24, 26 and 28 are resized to the nearest pivot font size using Nearest Neighboring interpolation so that it can be recognized.

II. LITERATURE REVIEW

Current state of the art techniques for the classification and recognition of Urdu document images are divided into two categories; (1) character-based classification and recognition and (2) ligature-based classification and recognition. In character-based classification and recognition technique, RASM (main body) is segmented into primitives which can be characters and are classified based on shape similarity. Safabakhsh and Adibi [5] use structural and discrete features of segmented primitives to develop a HMMs based recognition system of Nastalique handwritten text. The system has 96.8% recognition accuracy. The segmentation technique based on branch point of thin stroke of Nastalique main bodies is used to develop HMMs based recognition model for Nastalique text [6, 7]. The reported accuracy of the system tested on 2,494 ligatures at 36 font size is 92.19%. A character-based classification and recognition technique of Urdu Nastalique RASMs is presented for 14 font size text [8]. By using consistent character traversal in a ligature, sequence of grapheme labels are classified using Discrete Cosine Transforms (DCTs) as features, extracted through local windowing, and HMMs as classifier. The system has 97.11% accuracy tested on 79,093 instance images of 5,249 main bodies and 87.44% main body recognition accuracy tested on document images of different books. Ul-Hasan et al. [9] use bidirectional LSTM networks for character recognition of Nastalique ligatures. The reported accuracy of the system is 94.85% tested on synthesized images of 2,003 text lines.

Ligature-based classification and recognition deals with extraction of features from the ligature. These extracted features along with labels of ligature types are used to train the classifier. Javed et al. [10] present HMMs based recognition of Nastalique text. The system has 92% recognition accuracy tested on synthesized data at 36 font size. The DCTs as features and SVM as classifier are used to recognize the main bodies and diacritics [11]. The shape context features are extracted from contours of main body and diacritics for the recognition of Urdu and Arabic text [12]. The system has 91% accuracy tested on synthesized data of Urdu. Tesseract is an open source multilingual and font size independent OCR engine which is used for the recognition of text of different languages [13-15]. Tesseract is modified to make a matured recognition system for Urdu like cursive script [16]. Two systems are developed for 14 and 16 font sizes separately. The reported accuracies of systems are 97.87% and 97.71% for 14 and 16 font size respectively tested on 22,125 instances of 1,475 main body classes.

Different approaches exist in literature to develop a font size independent OCR. Some studies propose to train the data at each font size to make it font size independent [17, 18]. whereas some techniques extract size invariants features [19, 20]. In addition, image normalization is applied on images to develop OCR at single font size [9, 13]. Majority of the above mentioned techniques for the recognition of Urdu text support single font size text. In this paper, the font size independent approach is presented which recognizes text having font size between 14 to 28.

III. METHODOLOGY

Recognition of Urdu text written using Nastalique writing style having multiple font sizes is a challenging task due to complexities of Nastalique, some of which are discussed above. The diagonally growing ligature height results in variation of line height of text lines having same font size, see Fig. 1, which adds additional complexity to determine font size of a line and to normalize the line height to single font size. In Latin script x-height is normally used to detect the font size of the input text and respective normalization techniques are applied to map input line image at standard font size at which OCR has been developed. In this paper, a framework for the development of font size independent OCR for Urdu like cursive script is presented, see Fig. 2. In pre-processing phase, Urdu document images are binarized using technique presented in [21]. The connected components of the document images are extracted and disambiguated as RASM or diacritics using the dimensional features. In addition, information of diacritics association with respective RASM is also maintained. The font size of the input image is computed using diacritics. The main bodies, diacritics and font size information are forward to the classification and recognition module. In this module, diacritics and main bodies are recognized. The ranked list of ligatures' Unicode is generated by processing recognized diacritics and ranked list of main bodies. In this paper, the main idea is to modify only classification and recognition module and other OCR phases will remain unchanged to make a font size independent OCR. The focus of this paper is to develop a font size independent system for the recognition of main bodies, highlighted with blue in Fig.2.

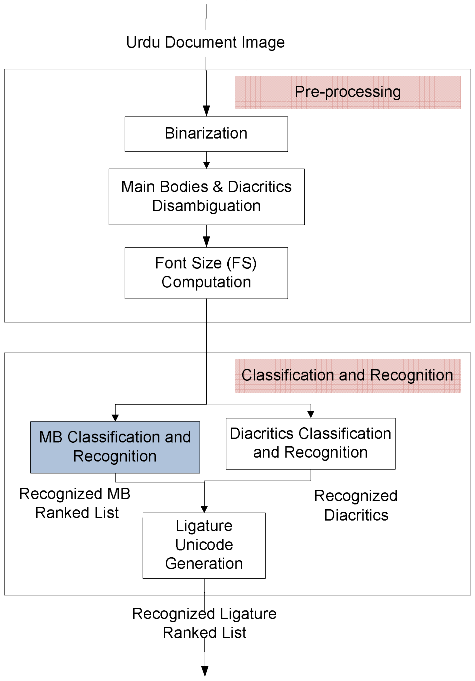

Urdu Document Image
Pre-processing
Binarization
Main Bodies & Diacritics Disambiguation
Font Size (FS) Computation
Classification and Recognition
MB Classification and Recognition
Diacritics Classification and Recognition
Recognized MB Ranked List
Recognized Diacritics
Ligature Unicode Generation
Recognized Ligature Ranked List

Fig. 2. Framework of OCR of font size independent system

The font size information and main bodies are fed to the main body (MB) classification and recognition module. To cover the range of font sizes from 14 to 28, three classifiers are developed. The main body images of remaining font sizes will be resized to nearest pivot font size so that it can be recognize from respective recognizer. The flow of MB classification and recognition module is given in Fig. 3. There are three sub-processes to develop this module as font size independent; (1) Development of recognizers, (2) Image resizing and (3) Maturation of recognizer using scaled data. The details are given below.

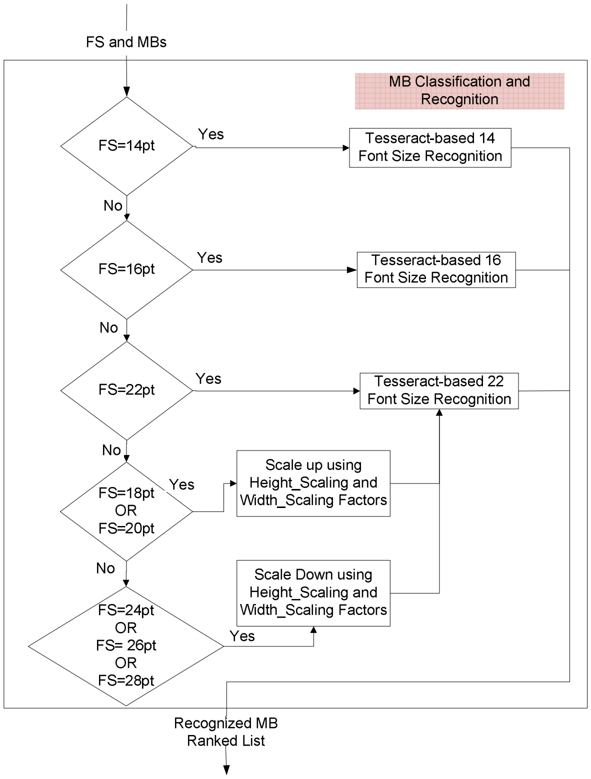

FS and MBs
MB Classification and Recognition
FS=14pt
Yes
Tesseract-based 14 Font Size Recognition
No
FS=16pt
Yes
Tesseract-based 16 Font Size Recognition
No
FS=22pt
Yes
Tesseract-based 22 Font Size Recognition
No
FS=18pt OR FS=20pt
Yes
Scale up using Height_Scaling and Width_Scaling Factors
No
Scale Down using Height_Scaling and Width_Scaling Factors
FS=24pt OR FS= 26pt OR FS=28pt
Yes
Recognized MB Ranked List

Fig. 3. Font Size independent Classification and Recognition of MBs

A. *Development of recognizers at pivot font sizes*

Based on the survey of font sizes appeared in Urdu books, it has been observed that majority of normal text is printed at 14 and 16 font sizes. Therefore two separate recognizers are developed at these font sizes to have reasonable recognition accuracy of document image. Majority of children books and headings are printed at 22 font size therefore third classifier is developed at 22 font size and text images of remaining font sizes are resized to 22 font size so that it can be recognized.

Modified Tesseract engine for the recognition of Nastalique main bodies [16] is used for development of recognizers. For each of pivot font size, four sub-recognizers are developed to improve the recognition accuracy of main bodies as discussed in [16]. Therefore, C4.5 algorithm is used to compute width thresholds of main body images which divide the main body training dataset into four sets. In addition using standard deviation of average width of main body types, the overlapping factor is also computed to handle main bodies which lie at the boundary of each set. The modified Tesseract is trained on each training set of a specific font size. The finalized width thresholds and recognition results of each pivot font size recognizer are discussed in Results and Discussion Section.

B. *Image resizing*

To make recognition system a size invariant, some techniques do image normalization, but to cover the range of the font sizes, image normalization without parameter tweaking does not work. Therefore, in this paper an image resizing technique along with tweaking of scaling factors to improve the recognition results of resized images is presented. Three interpolation techniques namely Nearest Neighbor, Bilinear and Bicubic [22, 23] are applied on sample data and results are analyzed. Based on efficiency and accuracy results, Nearest Neighbors is selected for this research study. The main body images of 18 and 20 font sizes are scaled up to 22 font size whereas main bodies of 24, 26 and 28 font sizes are scaled down to 22 font size so that these can be recognized from 22 font size recognizer. The scaling factor which is used to resize the source font size image to the target font size image is computed in two passes.

In first pass, the initial scaling factor is computed by analyzing height ratio of source and target main body images and width ratio of source and target main body images. The average width and average height are computed over all instances of a main body type of source font size. In the same way, average width and average height are computed over all instances of main body type of target font size. Then height and width ratios are computed which are used to resize the main body instances of source font size to the respective target font size. The same process is repeated to compute height and width ratios of 1,965 main body classes. The average values of all ratios of height and width are computed. The resultant values are termed as height scaling factor and width scaling factor.

In second pass, the optimal scaling factor is computed after analyzing the recognition results. Once the scaling factors of all font sizes have been computed, the Nearest Neighboring interpolation technique is used to resize the image. Optimal scaling factor is computed by analyzing the recognition results of rescaled images. Instead of four sub-recognizers, single recognizer of 22 font size is used to avoid the risk of misrecognition from wrong sub-classifier due to variation in width of resized MB image. The recognition accuracy is computed in terms of MB type recognition accuracy. For the testing and analysis, the dataset of each font size discussed in Dataset section is used. Different scaling factor configurations are used to optimize the scaling factors which includes computed scaling factor, computed scaling factor with ±3% of computed scaling factor, computed scaling factor with ±6% of computed scaling factor and computed scaling factor with ±9% of computed scaling factor. These configurations are applied to resize the images. The scaling factor configuration is selected for each font size which gives best recognition results. The configurations of scaling factors along with recognition results for each font size are given in Results and Discussion Section.

C. *Maturation of recognizer at scaled data*

To further improve the recognition results, 18, 20, 24, 26 and 28 font sizes data is rescaled using optimal scaling factors. The four sub-recognizers of 22 font size are matured by re-computing set division thresholds on 22 font size data and resized data. Four sub-recognizers at 22 font size are re-trained on modified data. To further improve the recognition accuracy, scaled data of all font sizes is tested from the respective sub-recognizer and recognition results are analyzed. The improvement pass is also carried out for the main body classes which have less than 80% recognition accuracy by adding resized images in training data.

IV. DATASET

To develop dataset for this research study, different text corpora are processed to extract 1,965 high frequent ligature classes. The selected ligature classes (main body types) covers 323,6154 ligatures instances. It also covers 59,974 unique Urdu words. The synthesized data at each of the selected font size i.e. 14, 16, 18, 20, 22, 24, 26, and 28 font sizes is developed by typing in Noori Nastalique writing style using Inpage software. For each of font size, a total of 35 tokens of each ligature class are typed, printed and then scanned at 300 DPI. A separate system is used to extract main bodies from each ligature class. In addition, image dataset [24] scanned from different Urdu books written in Noori Nastalique writing style at desired font sizes is selected. The main bodies are extracted automatically and distributed into classes. The real and synthesized data are used to prepare the training and testing data. For the development of recognizers at 14, 16 and 22 font sizes, ten (five synthesized and five real) instances are used for training and 15 non-overlapping instances (real and synthesized) are used for testing. The synthesized data is used for the main bodies classes which do not have sufficient real data for training or testing. For the development of resizing system, 25 instances (all real instances and remaining synthesized) of each main body class are used for each of 18, 20, 24, 26 and 28 font sizes.

V. RESULTS AND DISCUSSION

A. *Recognizers accuracy*

To develop an efficient and accurate recognition system using Tesseract, four sub-recognizers for each pivot font size are developed to reduce search space [16]. For dataset of each font size, C4.5 algorithm using width as feature is used to divide training dataset into four sub-datasets. The overlapping thresholds are computed to handle the main bodies which lie at boundaries. The width thresholds and overlapping constants for 14 font size, 16 font size, 22 font size (Pass-1) and 22 font size maturation pass on rescaled data (Pass-2) are given in Table 1. Separate Tesseract-based recognizers are developed using training data of each sub-dataset namely Recognizer-1 (R-1), Recognizer-2 (R-2), Recognizer-3 (R-3) and Recognizer-4 (R-4). After finalization of optimal scaling factors, a maturation pass of 22 font size recognizer is also carried out. The font wise accuracy results of 14, 16, 22 (Pass-1) and 22 (Pass-2) are given in Table 2. During maturation pass, the recognition accuracy of high frequent main body is tried to improve for all font sizes therefore in 22 (Pass-2) R-1 and R-2 have higher accuracy and R-3 and R-4 have lower accuracy.

TABLE 1. WIDTH THRESHOLDS AND OVERLAPPING CONSTANTS FOR 14 FONT SIZE, 16 FONT SIZE, 22 FONT SIZE (PASS-1) AND 22 FONT SIZE(PASS-2) FONT

Font size	**Width Threshold-1 (WT$_1$)**	**Width Threshold-2 (WT$_2$)**	**Width Threshold-3 (WT$_3$)**	**Overlapping Threshold 2*Sigma**
14	49	57	71	3.974
16	52	63	77	3.352
22 (Pass-1)	71	87	103	4.495
22 (Pass-2)	70	93	116	9.1

TABLE 2. FONT WISE ACCURACY OF SUB-RECOGNIZERS

Font size	**R-1 Accuracy (%)**	**R-2 Accuracy (%)**	**R-3 Accuracy (%)**	**R-4 Accuracy (%)**
14	99.30	98.73	97.03	93.73
16	98.83	99.00	97.74	92.24
22(Pass-1)	99.22	94.55	97.03	94.68
22(Pass-2)	99.52	95.52	95.44	94.14

B. *Image resizing accuracy*

Nearest Neighbors interpolation technique is used to resize the source image. Different scaling factor configurations are used to optimize scaling factors which include computed scaling factor, computed scaling factor with ±3% of computed scaling factor, computed scaling factor with ±6% of computed scaling and computed scaling factor with ±9% of computed scaling. A total of 25 tokens of each main body type are rescaled using respective scaling factor configuration and then recognized using 22 font size single recognizer. The configurations of scaling factors along with recognition results for each font size are given in Table 3. The configuration which gives highest accuracy (highlighted with bold in Table 3) is selected as optimal scaling factor configuration. The simple scaling factor does not give the optimal recognition results, see Table 3.

VI. CONCLUSION

In this paper, a framework for the development of font size independent OCR is presented. The presented technique is font size independent, but to improve the recognition results of frequently used font size text and to make a real time system, separate recognizers for 14 and 16 font sizes are developed. The Nearest Neighbors interpolation technique is used to resize the images at 22 font size. It has been observed that simple scaling factor used to resize the image does give optimal recognition accuracy. We need to tweak the scaling factors. Initially simple scaling factors are computed and then optimal scaling factors are finalized after analysis of recognition results of scaled data at 22 font size recognizer. Finally, a maturation pass is carried out to improve the recognition results of scaled data. The system is matured on training data of 1,965 main body types of each of the selected font size. The system has 97.20%, 97.08%, 95.13%, 95.65%, 96.26%, 96.52%, 95.78% and 96.66% main body recognition accuracy for 14, 16, 18, 20, 24, 26, 28 font sizes respectively. Due to unavailability of standard training and testing data for a range of font sizes, the comparison of existing Nastalique recognition systems with presented technique is not possible.

VII. ACKNOWLEDGEMENTS

This work has been supported by Urdu Nastalique OCR research project grant by ICTR&D Fund, Ministry of IT, Govt. of Pakistan. See www.UrduOCR.net for details.

TABLE 3. SCALING FACTORS CONFIGURATIONS AND TYPE WISE ACCURACY

Configurations	**F18ToF22 Type wise accuracy (%)**	**F20ToF22 Type wise accuracy (%)**	**F24ToF22 Type wise accuracy (%)**	**F26ToF22 Type wise accuracy (%)**	**F28ToF22 Type wise accuracy (%)**
(-9%,-9%)	94.54	95.02	95.52	94.72	96.04
(-6%,-6%)	94.70	94.82	95.70	95.31	95.24
(-3%,-3%)	**95.22**	94.82	96.41	95.62	95.89
(0,0)	94.99	95.33	96.18	95.64	96.03
(+3%,+3%)	94.77	95.55	95.82	95.65	95.38
(+6%,+6%)	94.67	95.56	96.51	**95.82**	**96.44**
(+9%,+9%)	94.55	**95.72**	**96.57**	95.68	96.31

REFERENCES

[1]. Indian Institute of Science. (2015, December, 03). Digital Library of India. Available: www.dli.ernet.in/

[2]. P. Gutenberg. (2014, December 03). Free ebooks - Project Gutenberg. Available: https://www.gutenberg.org/

[3]. M. Davis and L. Iancu. (2016, December 03). Unicode Text Segmentation. Unicode Text Segmentation 29. Available: http://unicode.org/reports/tr29/

[4]. A. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," in CISSE, 2006.

[5]. R. Safabakhsh and P. Adibi, "Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM," The Arabian Journal for Science and Engineering, vol. 30, pp. 95-118, 2005.

[6]. S. T. Javed and S. Hussain, "Segmentation Based Urdu Nastalique OCR," in CIARP 2013, Havana CUBA, 2013.

[7]. A. Muaz, "Urdu Optical Character Recognition System," Unpublished, MS Thesis Report, National University of Computer and Emerging Sciences , Lahore, 2010.

[8]. S. Hussain, S. Ali, and Q. u. A. Akram, "Nastalique segmentation-based approach for Urdu OCR," IJDAR, vol. 18, pp. 357-374, 2015.

[9]. A. Ul-Hasan, S. B. Ahmed, F. Rashid, F. Shafait, and T. M. Breuel, "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," in ICDAR 2013, 2013, pp. 1061-1065.

[10]. S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, "Segmentation Free Nastalique Urdu OCR," WASET, vol. 70, 2010.

[11]. G. S. Lehal and A. Rana, "Recognition of Nastalique Urdu ligatures," in Proceedings of the 4th International Workshop on Multilingual OCR, Washington, D.C., USA, 2013, pp. 1-5.

[12]. N. Sabbour and F. Shafait, "A segmentation-free approach to Arabic and Urdu OCR," in SPIE 8658, Document Recognition and Retrieval XX,, 2013, pp. 86580N-86580N-12.

[13]. R. Smith, "An Overview of the Tesseract OCR Engine," in ICDAR 2007- Volume 02, 2007, pp. 629-633.

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

[15]. M. A. Hasnat, M. R. Chowdhury, and M. Khan, "An Open Source Tesseract Based Optical Character Recognizer for Bangla Script," in ICDAR 2009, 2009, pp. 671-675.

[16]. Q. Akram, S. Hussain, A. Niazi, U. Anjum, and F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique," in DAS2014, 2014, pp. 191-195.

[17]. T. Kanungo, P. Resnik, S. Mao, D.-W. Kim, and Q. Zheng, "The Bible and multilingual optical character recognition," Commun. ACM, vol. 48, pp. 124-130, 2005.

[18]. R. Mehran, H. Pirsiavash, and F. Razzazi, "A Front-End OCR for Omni-Font Persian/Arabic Cursive Printed Documents," in DICTA'05, 2005, pp. 56-56.

[19]. G. S. Lehal and C. Singh, "A Gurmukhi script recognition system," in ICPR-2000, 2000, pp. 557-560 vol.2.

[20]. S. D. Zenzo, M. D. Buono, M. Meucci, and A. Spirito, "Optical recognition of hand-printed characters of any size, position, and orientation," IBM J. Res. Dev., vol. 36, pp. 487-501, 1992.

[21]. M. Naz, Q. Akram, and S. Hussain, "Binarization and its Evaluation for Urdu Nastalique Document Images," in INMIC 2013, Lahore, Pakistan, 2013.

[22]. R. C. Gonzalez and R. E. Woods, Digital Image Processing (3rd Edition): Prentice-Hall, Inc., 2006.

[23]. Z. Haifeng, Z. Yongfei, and H. Ziqiang, "Comparison of Image Amplifying Method," Modern Electronics Technique, vol. 24, pp. 33-36, 2010.

[24]. Q. Akram, A. Niazi, F. Adeeba, S. Urooj, S. Hussain, and S. Shams, "A Comprehensive Image Dataset of Urdu Nastalique Document Images," in CLT 16, Lahore, Pakistan, 2016.