Cluster Comput (2018) 21:503–514
https://doi.org/10.1007/s10586-017-0916-2

Urdu ligature recognition using multi-level agglomerative hierarchical clustering

Naila Habib Khan[1] · **Awais Adnan**[1] · **Sadia Basar**[2]

Received: 13 February 2017 / Revised: 4 May 2017 / Accepted: 9 May 2017 / Published online: 25 May 2017

Abstract Optical character recognition (OCR) system holds great significance in human-machine interaction. OCR has been the subject of intensive research especially for Latin, Chinese and Japanese script. Comparatively, little work has been done for Urdu OCR, due to the complexities and segmentation errors associated with its cursive script. This paper proposes an Urdu OCR system which aims at ligature-level recognition of Urdu text. This ligature based recognition approach overcomes the character-levelsegmentation problems associated with cursive scripts. A newly developed OCR algorithm is introduced that uses a semi-supervised multi-level clustering for categorization of the ligatures. Classification is performed using four machine learning techniques i.e. decision trees, linear discriminant analysis, naive Bayes and k-nearest neighbor (K-NN). The system was implemented and the results show 62, 61, 73 and 90% accuracy for decision tree, linear discriminant analysis, naive Bayes and K-NN respectively.

Keywords Agglomerative · Clustering · Classification · OCR · Urdu

✉ Naila Habib Khan
naila.khancs@yahoo.com

Awais Adnan
awais.adnan@imsciences.edu.pk

Sadia Basar
sadiaa.khancs@gmail.com

[1] Department of Computer Science, Institute of Management Sciences, Peshawar, Khyber Pakhtunkhwa, Pakistan

[2] Department of Information Technology, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan

1 Introduction

Script recognition system is considered to be a huge part of Natural Language Processing (NLP) [1]. It deals with the interaction between the computer and human languages. The main aim of NLP is to make a machine able to analyze, understand and generate human languages such as text translation, speech processing, voice controlled machines and automated scene understanding [1,2]. Communication improvement between man and machine has been the primary motivation for many researchers. One fundamental application of NLP is a text processing system. A text processing system, also known as optical character recognition (OCR), has served numerous benefits in this technology era, including conversion of century-old literature into computer understandable format.

OCR systems are "Systems which use a process by which text-based input patterns produce meaningful output" [3]. An OCR includes several phases, primarily its two major phases are feature extraction and classification. Currently, the OCR technology has developed and expanded to include several types of texts and fonts, as well as, support for handwritten text recognition [3]. Intelligent Character Recognition (ICR) and Intelligent Word Recognition (IWR) are both predecessors of OCR; they support handwritten character level recognition and word level recognition respectively. However, the advancements still haven't excelled for processing and recognition of cursive languages such as Urdu, Arabic, Persian, Pashto, Korean and Chinese. On the contrary, non-cursive languages such as English, German and French script are relatively easy to recognize [4].

Urdu is the national language of Pakistan and its Nastalique calligraphic script is extremely cursive and context sensitive in nature [5,6]. Urdu script also shares the same level of written complexity with Arabic, Pashto and Persian scripts

[4]. The abundant complexities associated with Urdu Script makes it a scarce language to be considered for OCR [6], hence limiting its research and development as compared to the Latin script [7]. In Urdu, a word is composed of sub-words (ligatures) and isolated characters. Ligatures itself is a unit block of text, composed of a combination of two or more characters. Due to extreme cursiveness and overlapping issues such as inter-ligature overlapping and intra-ligature overlapping, it is extremely challenging to perform segmentation [8]. Hence, Urdu character-level recognition systems that require excessive segmentation processes may lead to erroneous outcomes and disfigure the actual shape of characters. If working with ligature-level recognition systems, there are copious ligatures forms possible, each cannot be classified separately, and therefore, clustering them based on certain features seems to be a suitable solution. This research paper aims at developing a ligature based recognition system for Urdu script using multi-level clustering and classification.

The main goal of this research paper is threefold: (1) For Urdu script beginners, we want to highlight all the challenges associated with an Urdu OCR system. (2) For technical researchers, we want to emphasize on all algorithms that have been previously used for Urdu OCR and provide a distinctive clustering approach to successfully cluster Urdu script at ligature-level. (3) For NLP experts, we believe that this research will prove to be a great contribution to Urdu Informatics.

2 Related work

As discussed in Sect. 1, we can divide the existing literature into two broad categories, character-level recognition systems (analytical approach) and ligature-level recognition systems (holistic approach). Previously, several authors have opted ligature-level recognition systems [9–17]. Husain [9], recommended the use of a multi-holistic approach for recognition of Urdu script. Holistic approach has the capability of recognizing whole word or sub-parts. Similarly, Shah [10] opted to work on a ligature based OCR system for Urdu Nastalique font using an isolation algorithm, template matching technique was used for ligature classification. Khan et al. [12] used template matching technique along with correlation algorithm. The recognized characters were converted into the text based on dictionary searching algorithm. It was also observed that the higher are the number of ligatures, recognition rate increases. A 100% accuracy was achieved for 5-character and 4-character ligatures. On the contrary, Husain et al. [11] used Back Propagation Neural Network (BPNN) for recognition of Urdu ligatures. The main constraint of the system was that the secondary stroke must be written after the primary stroke.

Razzak et al. [13] proposed fuzzy-based preprocessing techniques to recognize and normalize the handwritten ligature stroke using both online and offline domain. The proposed system achieved an accuracy of 74.3% for Nastalique font and 60.7% for Naskh font. In another research paper, Razzak et al. [14] used segmentation free approach for recognition of online Urdu handwritten script. The handwritten script was recognized using three major algorithms (hybrid classifier, hidden Markov model and fuzzy logic) and robust features (structural and statistical). A total of 26 time-variant structural and statistical features were extracted from the base strokes. Contrarily, Akram et al. [15], targeted size and font-invariant text recognition, using the concept of text resizing using splines. In splines, the outline of the ligature was captured and scaled to an OCR trained size. Outline capturing was applied for font size normalization in situations when the input ligature was not standard. Lastly, the ligature outline was captured using a chain coding algorithm. Javed et al. [16] also stressed on features for ligature recognition. Global transformational features were extracted from the image and fed to Hidden Markov Model recognizer for identification purposes. After identification, diacritics and main body were further divided into frames, discrete cosine transform was calculated for each frame. Once the diacritics were recognized a global transformational feature vector was created.

Analytical approach based systems have also been the center of attention of many researchers [18–26]. Pal and Sarkar [18] developed a character recognition system to deal with printed Urdu script having Nastalique calligraphic style. Horizontal projection profile and vertical projection profile was used for line segmentation and character segmentation respectively. Similarly, Chanda and Pal [20] also used horizontal and vertical histogram for text segmentation. A binary tree classifier was used for classification of characters and achieved 97.51% accuracy. Another tree based dictionary searching method was suggested by Malik and Khan [19] for online handwriting recognition.

Moment invariant technique, primary and secondary component separation and support vector machine for classification was used by Pathan et al. [21]. In contrast, Shahzad et al. [23] formulated an OCR that had the capability of recognizing hand-sketched Urdu characters. For each character, the strokes were divided into two groups namely “primary strokes” and “secondary strokes”. These features were then trained and classified using a linear classifier. For Urdu Naskh, Nawaz et al. [24] presented an offline optical character recognition, the characters were considered in their isolated forms and pattern matching technique was used for classification. Another system was proposed using finite state model by Sattar et al. [17]. The Nastalique text recognizer included two components;

character shape recognizer and a next-state function. For recognition of characters within a ligature, cross-relation was used.

Neural Network is one of the most famous classifiers for an analytical approach based Urdu OCR. Zaman et al. [22] proposed a recognition system for segmented Arabic/Urdu characters, having four phases (input, preprocessing, classification and recognition) using Feed Forward Neural Network (FFNN). This FFNN classified a total of 53 classes, using 2500 input neurons, 20 hidden neurons, and 53 output neurons. Similarly, Ahmad et al. [25] developed an OCR system based on two modules; segmentation and classification. The pixel strength was used to detect words in a sentence and joints of characters in a compound word. In the second phase, the segmented characters were fed to a trained neural network for classification and recognition. The FFNN was trained on 56 different classes of characters, each having 100 samples. Shamsher et al. [26] also, used FFNN for printed isolated Urdu characters. The neural network was composed of three layers i.e. input layer, hidden layer, and an output layer; having 150, 250 and 16 neurons respectively. Input layer was responsible for receiving binary data. A prototype of the system was tested to achieve 98.3% of accuracy for Urdu characters.

3 Background and characteristics of Urdu Script

Urdu is the national language of Pakistan [6,27,28] and spoken by 70 million people [28]. It is also an official language of five Indian states [6]. Urdu, like Arabic and Persian, is written in Perso-Arabic script; therefore, these scripts share similarity at the written level. Arabic and Persian has great influence on Urdu, therefore it uses a modified and extended set of Arabic and Persian alphabets. Urdu alphabet has a total of 38 characters [23], 28 are similar to the Arabic alphabet [18] (see Table 1 and Fig. 1).

In Urdu, a word is a composition of ligatures while a ligature is a combination of characters [17,29]. In addition, blank spaces may or may not be regarded as separation/boundary between words. Urdu is written in the famous Nastalique calligraphic style, very complex and context sensitive in nature [30], whereas Arabic script is written using Naskh font. Urdu alphabet can be divided or grouped into 21 classes based on the similarity of primary strokes. Primary stroke is the when a character is considered without the accompanying dots or diacritics. These diacritic marks are also used in combination with characters for proper pronunciation.

Table 1 Comparison between Urdu and Arabic writing systems

Characteristics	Urdu	Arabic
No. of letters	38	28
Writing direction	Right to left	Right to left
Cursiveness	Yes	Yes
Vertical justification	Center	Base
Diacritics	Yes	Yes

3.1 Urdu Nastalique OCR challenges

The extremely cursive and calligraphic nature of Nastalique makes the development of an Urdu OCR system challenging [17]. To write Nastalique text a pen is used having a special nip called "qat" [8], which results in script with varying stroke width, making handwritten text recognition difficult. Even if developing an OCR for printed Urdu script, we have two approaches holistic and analytical. Regardless of the approach used, Nastalique font poses several challenges in developing a robust optical character recognition system. Below, we discuss key challenges that are faced at the time of an Urdu OCR system development.

3.1.1 Context sensitivity

Arabic script is cursive; characters are joined together to form ligatures/words. It's written in Naskh font and the characters can have two to four basic forms based on their location in the ligature/word; isolated, initial, middle and final [31]. On the contrary, Urdu is written in Nastalique font that is far more complex than this 4-shapes phenomenon. In Urdu, the shape of a character is not only affected by its position but also by its neighboring characters. This sensitivity is referred to as context sensitivity [17]. Figure 2 shows the context sensitive behavior of character 'te' from Urdu alphabet. It can be observed easily that the neighboring characters and its position (start, middle or final) in the ligature affect its shape.

3.1.2 Diagonality

One of the main characteristics of the Nastalique font is that it is written diagonally, from top right to the bottom left [32,33]. As new characters are joined to the former characters, a slant or slope is introduced in the ligature being written, this feature is known as diagonality. In comparison, English and Arabic are written horizontally along a single baseline. Figure 3a shows the diagonal nature of Urdu script, on the contrary, the horizontalness and straightness of Arabic script can be seen in Fig. 3b.

3.1.3 Overlapping problem

Urdu text takes less horizontal space as compared to the Arabic Naskh style of writing. Therefore, Urdu has extreme

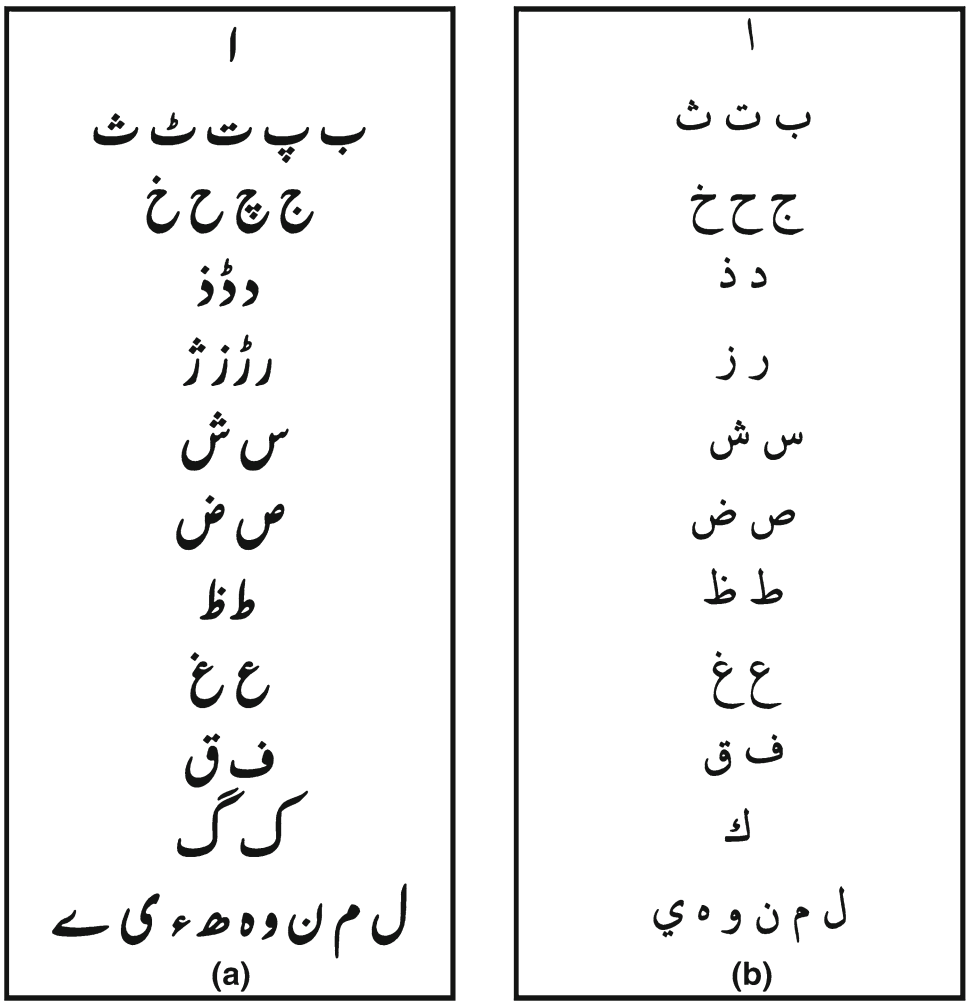

Fig. 1 **a** Urdu alphabet. **b** Arabic alphabet

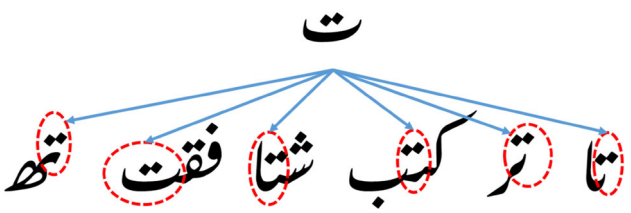

Fig. 2 Shape of character 'Te' affected by its neighboring characters

ligature overlapping issues [6,17]. Overall the overlapping can be divided into two types [34]. First, in intra-ligature overlapping, a character overlaps another character(s) within the same ligature. The second type of overlapping is known as inter-ligature overlapping, a character of one ligature overlaps a character(s) of another ligature. Both, inter-ligature overlapping and intra-ligature overlapping can be seen in Fig. 4. Inter-Ligature overlapping makes the process of ligature segmentation extremely difficult while intra-ligature overlapping makes the process of character segmentation challenging [35].

Fig. 3 **a** Diagonality in Urdu. **b** Horizontal baseline in Arabic

3.1.4 Diacritics

Urdu Characters are surrounded by special type of marks known as diacritics. The diacritics surrounds the characters

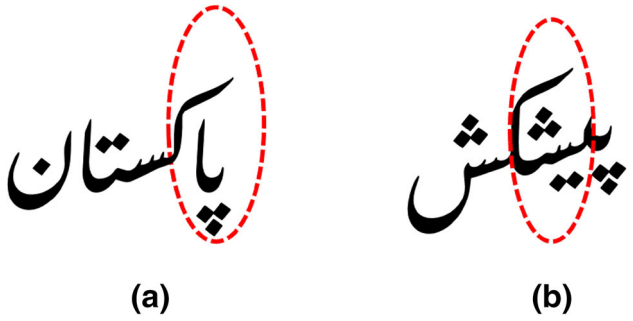

Fig. 4 **a** Inter-ligature overlapping. **b** intra-ligature overlapping

Fig. 5 Placement of *dots* at non-standard position

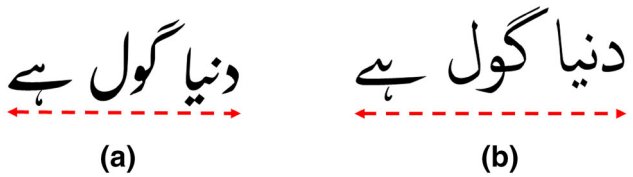

Fig. 6 **a** Spacing between Urdu ligatures. **b** Spacing between Arabic Ligatures

main body and lie above or below it [36]. There are three types of diacritics i.e. Nuqta (Dot), Aerab and "ط" Superscript. The context sensitive and sloping nature of Nastalique script makes the placement of diacritics difficult, standard positions cannot be followed (see Fig. 5). The diacritics for a character are shifted with the addition of every new character. The diacritics may be moved to a nearby position instead of its standard position [34]. This is mainly performed to avoid any clash or overlapping to occur between the nuqtas. Therefore, identifying the character to which the diacritics belong is difficult due to inter-ligature and intra-ligature overlapping.

3.1.5 Spacing

The Nastalique font consumes very less space and ligatures are written very tightly [35]. It is a difficult task to identify the full word [37]. This lack of space between ligatures makes the task of segmentation in Urdu OCR problematic [35]. Figure 6a shows Urdu written in its Nastalique calligraphic style, with little space between ligatures, as compared to Arabic Naskh font shown in Fig. 6b.

3.2 Application areas of Urdu OCR

Urdu script has a rich historical background. Pakistan, India, and Bangladesh are few of the countries where Urdu is spoken, understood and written widely. Due to the popularity of Urdu language at verbal and written level, and its massive hardcopy literature, abundant efforts have been directed towards Urdu OCR systems. A fully functional and efficient Urdu OCR system can have abundant applications in various fields. Some of the most renowned and emerging application areas of Urdu OCR are digital reformatting, automated text translation, text-to-speech conversion, Automated Number Plate Recognition (ANPR) and static-to-electronic media conversion. These application areas have been discussed in detail below.

- *Digital reformatting* In digital reformatting, original documents are converted into digital form. These digital documents act as surrogates, preserving and eliminating the need to use the original version. With an automated system, like OCR, it is possible to convert all the physical libraries into digital libraries. The Internet can then be used to transfer and spread the literature, making it available worldwide. Currently, the Internet is being used as a repository for making textual material online, it has been successful but with a few tradeoffs. Most of the literature on the internet now is in form of images containing text. These images consume a lot of storage space, also the time required to transfer the files from one place to another through the internet is slow. Hence, digital reformatting will allow the conversion of physical libraries to digital libraries with lesser time and space consumption.
- *Automated text translation* This is a famous application area of OCR, sometimes also referred to as "Machine Translation (MT)". Generally, automated text translation software translates text from a source language (for e.g. Urdu) to a target language (for e.g. English). Nowadays, Text translation software is being designed for personal, business as well as enterprise usage. These software's are extremely useful and lets you understand and convert a language script into the target language script in real time.
- *Text-to-speech conversion* OCR technology also provides handicapped accessibility to low-vision users. This is generally known as, "text-to-speech conversion" and more technically known as "speech synthesis". It involves converting the recognized text using OCR software into computer generated speech. This technology allows low-vision or blind people to read books, magazines or any other reading material after scanning it.

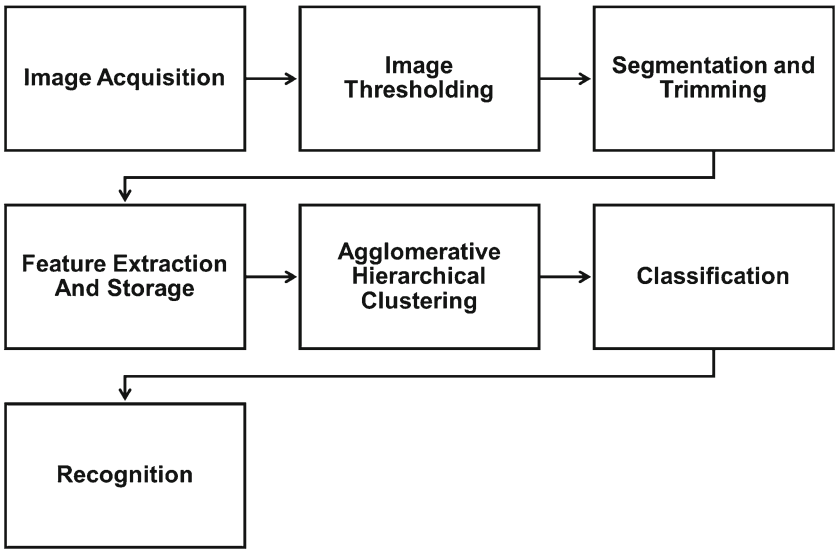

Fig. 7 Block diagram for recognition system

- *Automated number plate recognition* (*ANPR*) ANPR is a technology that reads vehicle registration plates using OCR technology. ANPR requires a fast video camera to capture the image. ANPR technology is being used worldwide by law enforcement agencies to keep track of vehicles such as vehicle license, vehicle registration and electronic toll collection on pay-per-use roads.
- *Static-to-electronic media conversion* E-media (Electronic media) is an emerging application area of OCR technology. Electronic media encompasses the use of electronics by the end user to access any content. On the contrary, the static media (print media) doesn't involve any use of the electronics. Static media such as newspaper can be converted to E-media using the OCR technology, by recognizing the newspaper headlines. Any handheld device, having a camera can be used to take a snap of the headlines or recognize the headlines in real-time. Once the headlines are recognized, the same news and its detailed content can be accessed online in a video form on the same handheld digital device.

4 Proposed methodology

In this study, we have proposed a unique clustering method, agglomerative hierarchical clustering, for ligature based Urdu OCR system. A corpus of 2430 ligatures was used for the clustering process. First, images containing the ligatures were fed to the system. Color image segmentation was carried out using global thresholding method, where the background was represented by black color and foreground (text) with white color. Next, segmentation was performed, horizontal projection profile and vertical projection profile was used for line and ligature segmentation respectively. In Urdu, the script is composed of characters of varying length, hence the segmented ligatures may be accompanied with unwanted pixels. Trimming was applied to remove unwanted pixels from top and bottom of the ligature images. Certain geometric features such as height, width, area, perimeter, aspect ratio, density function, area to perimeter ratio, horizontal histogram, vertical histogram, start-point, end-point and slope were extracted from each image. Agglomerative hierarchical clustering was applied and divided into two levels i.e. level-1 and level-2 clustering. In level-1 feature clustering, 32 clusters were generated for each of the 12 features. In level-2 feature clustering, an optimal number of sub-clusters (417) were generated for the already clustered features from level-1 for recognition feasibility. To verify the robustness of the clustering procedure, the classification was performed using four machine learning techniques; decision trees, linear discriminant analysis, naive Bayes and k-nearest neighbour. The research methodology described and proposed here can be broadly divided into 7 major phases (see Fig. 7).

4.1 Image acquisition, thresholding and segmentation

In the first phase of the recognition, ligature images were fed to the recognition system. These images were noise free and had no skew (see Fig. 8). Each of the images that were fed to the recognition system was converted to pure black and white colors using image thresholding. This image thresholding procedure resulted in the color image segmentation. The proposed system used the grayscale level of the image as the threshold level and applied global image thresholding.

In the final version of the bi-level thresholded image, Urdu ligatures were presented in a white color and the background

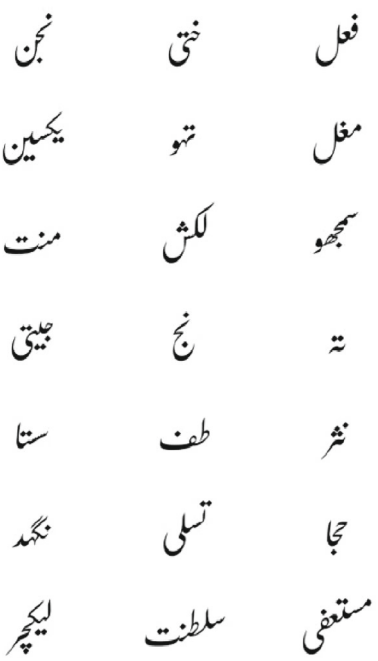

Fig. 8 Sample image fed to the OCR system

in black (see Fig. 9a). Thresholding procedure was followed by image segmentation. Image segmentation is a process of partitioning a digital image into multiple segments. In our proposed system, the segmentation module was divided into two sub-modules; line segmentation and ligature segmentation. For line segmentation, horizontal projection profile, the sum of pixel intensities over each row was used. Horizontal projection profile splits the text along horizontal text strips (see Fig. 9b). Next, ligature segmentation was performed using vertical projection profile, lines were analyzed one by one and individual ligatures were segmented (see Fig. 10). The resultant segmented images were used for feature extraction, classification and clustering purposes.

Next, the trimming process was carried out. Generally, trimming refers to cutting or clipping pieces of something. In our research, trimming is used to remove unwanted image segments. In Urdu, ligatures are formed from a combination of characters of varying heights, the height of each ligature may vary from its neighboring ligatures in a line of text (see Fig. 11).

4.2 Feature extraction and agglomerative hierarchical clustering

In this research, we have extracted a distinct set of geometrical features and then clustered the ligatures based on these features. The features proposed are easier to extract than morphological or topological features that require extreme knowledge overhead. The geometric features proposed for Urdu OCR were (see Fig. 12),

(1) *Width* Measurement from side to side of ligature image.
(2) *Height* Measurement for the tallness of ligature image.
(3) *Aspect ratio* Ratio of height divided by width of a ligature image.
(4) *Density function* Ratio of total number of pixels owned by the ligature to the area of image.
(5) *Perimeter* Sum of sides of a ligature image.
(6) *Area* Product of width and height.

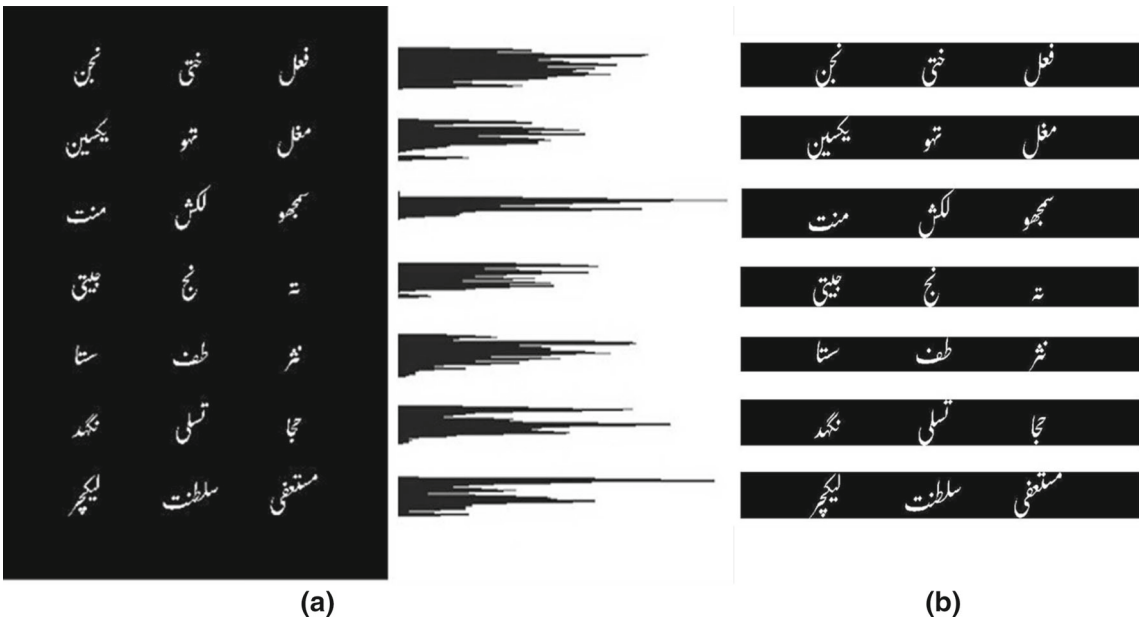

Fig. 9 **a** Horizontal project profile for line segmentation. **b** Result of segmented lines with horizontal projection profile

Fig. 10 Vertical projection profile for ligature segmentation

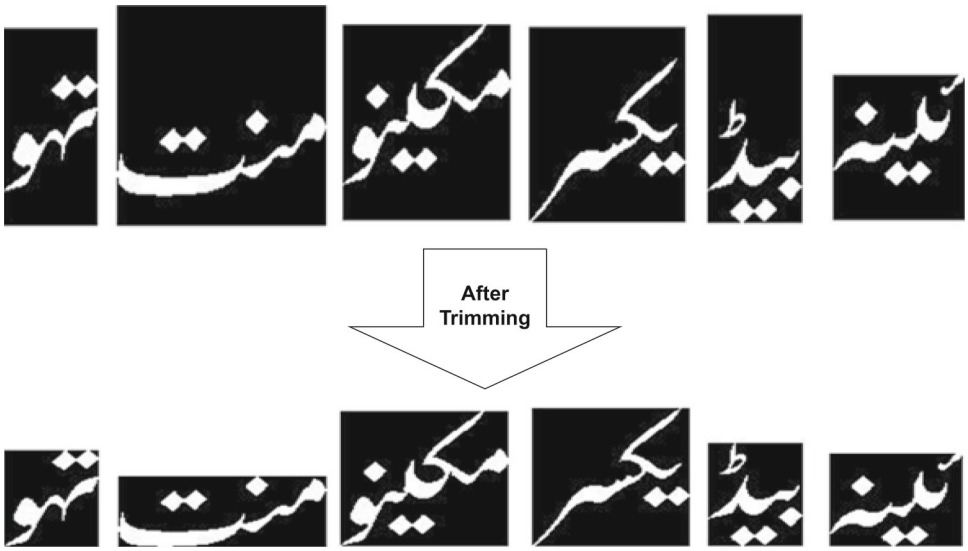

Fig. 11 Showing original ligatures (*first row*) followed by its trimmed version (*second row*)

(7) *Perimeter to area ratio* Division of perimeter by the area of a ligature image.
(8) *Horizontal projection profile* Sum of pixel intensities along each row in a ligature image. The row index for which we have the maximum horizontal histogram value was considered as a geometric feature.
(9) *Vertical projection profile* Sum of pixel intensities over each column in a ligature image. The column index for which we have the maximum vertical histogram value was considered as a geometric feature.
(10) *Start-point* The first white pixel scanned from top to bottom at the left-hand side border of the ligature image, where the ligature stroke touches the left-hand side border. Vertical-axis value was stored for the pixel.
(11) *End-point* The first white pixel scanned from top to bottom at the right-hand side border of the ligature image, where the ligature stroke touches the right-hand side border. Vertical-axis value was stored for the pixel.
(12) *Slope* The slope or diagonal line drawn from the start-point to end-point.

Once the geometric features were extracted, agglomerative hierarchical clustering was carried out. The agglomerative hierarchical clustering was divided into two levels i.e. level-1 and level-2. In level-1, the ligatures were divided into 32 clusters for each of its features. Next, in level-2, the ligatures were clustered again using the results of level-1 clustering. The total number of clusters in level-2 were generated using the equation,

$$L2_c = (C_{Min} - C_{Lig>2}) + C_A \tag{1}$$

Here $L2_c$ stands for clusters generated for level-2, C_{Min} refers to the minimum number of clusters generated by the clustering algorithm in default against the results of level-1. The total number of clusters generated, C_{Min} were 798 (see Fig. 13). But there was an issue, few of the clusters held only one or two ligature images, hence, overburdening the system with more number of clusters. Therefore, all those clusters that held only one or two ligatures images were collected and put into a single global cluster C_A. $C_{Lig>2}$ represents all the clusters having at least three ligature images. There were a total of 382 clusters that had more than 2 ligature images. Hence as result for level-2 ($L2_c$), a total of 417 clusters were generated (see Fig. 14).

4.3 Classification and recognition

To verify the robustness of this two-level clustering, four different classification algorithms were tested. Overall the Urdu ligature dataset was divided into two groups i.e. training and

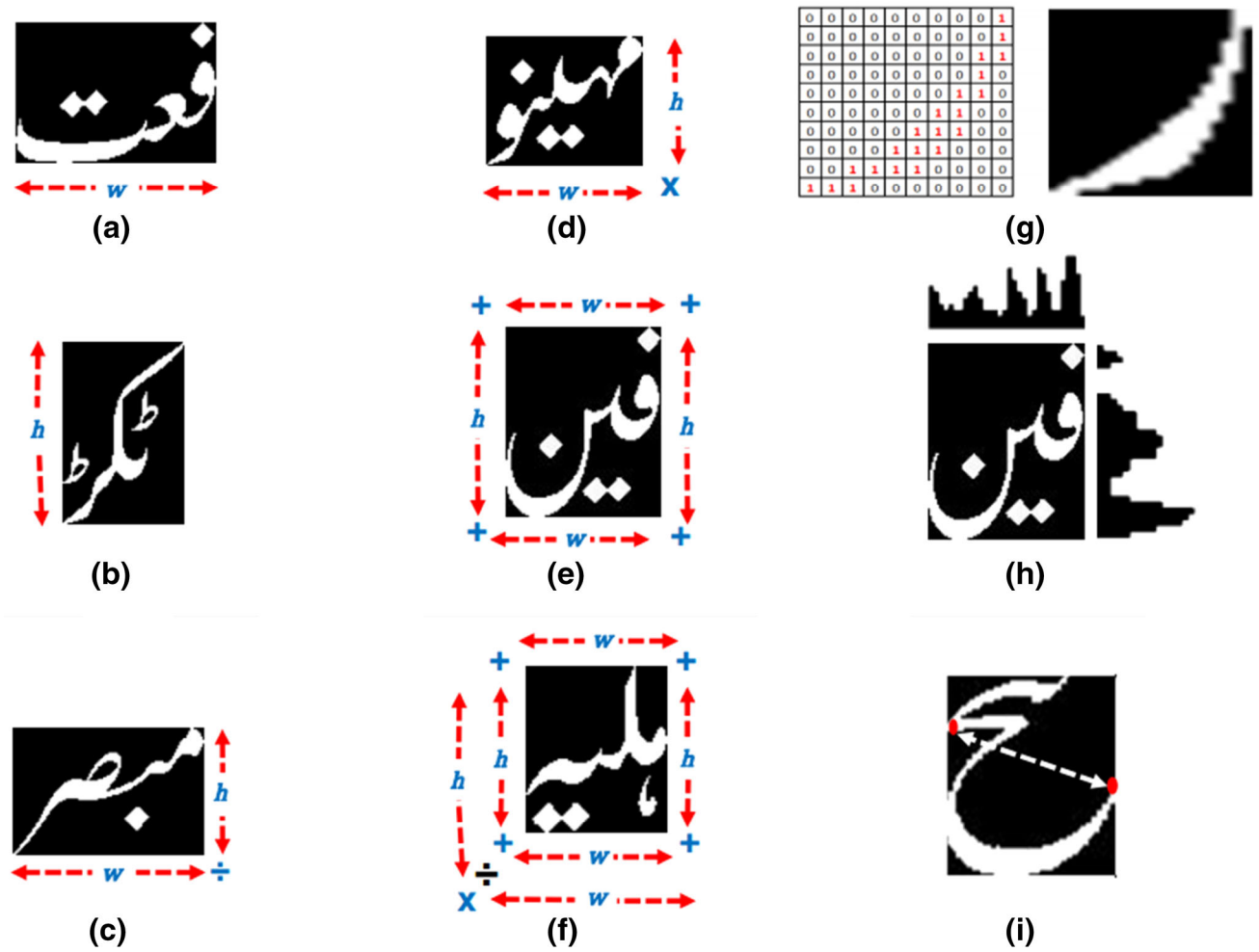

Fig. 12 Geometric features for the proposed Urdu OCR system. **a** Width. **b** Height. **c** Aspect ratio. **d** Area. **e** Perimeter. **f** Perimeter to area ratio. **g** Density function. **h** Horizontal and vertical projection profile. **i** Start-point, end-point and slope

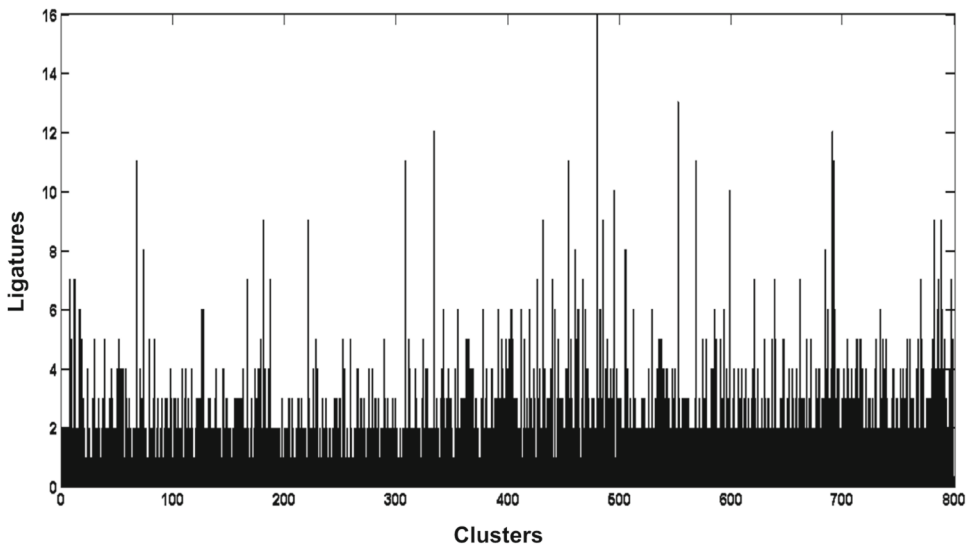

Fig. 13 Total number of clusters generated by default

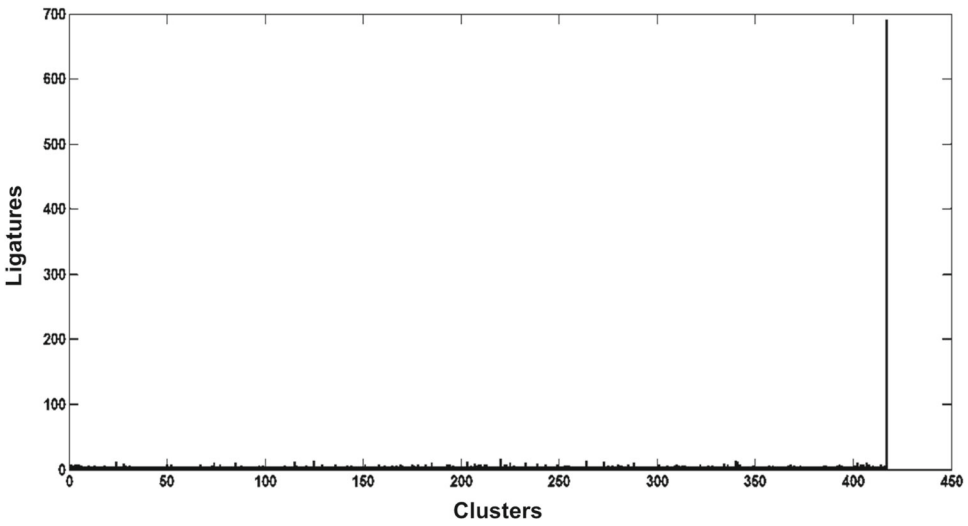

Fig. 14 Total number of clusters generated at level-2

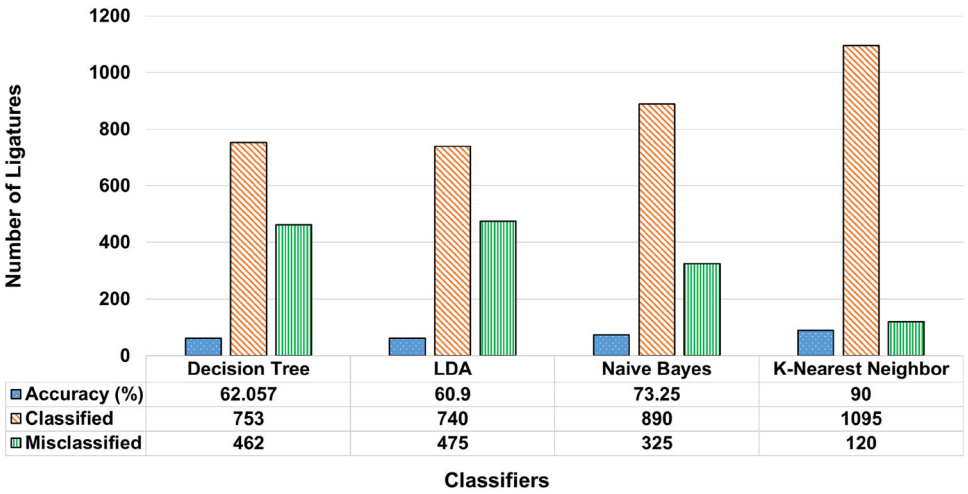

Fig. 15 Classification results

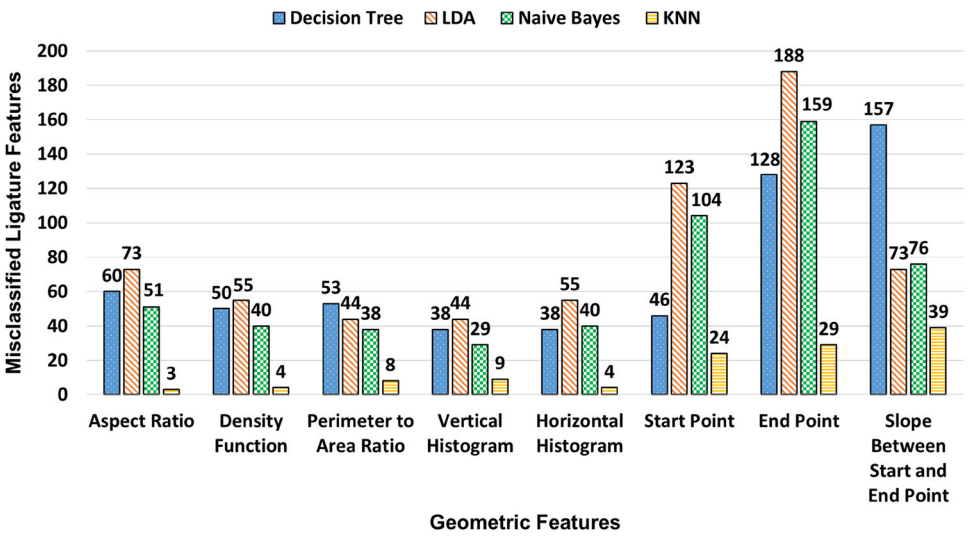

Fig. 16 Geometric Features Misclassification Rate

testing. Training data was used to train the classifier, the more is the data used for training the better are the results. Hence, for training, all 2430 ligatures were used. For testing, a total of 1215 ligatures were selected randomly. The same datasets (testing and training) were used for each of the classifiers i.e. decision tree, linear discriminant analysis, naive Bayes and

k-nearest neighbor. The accuracy, classification and misclassification rate for each of the classifiers was computed (see Fig. 15).

5 Conclusion and future directions

In this research study, a unique multi-level agglomerative hierarchical clustering was used to generate an ideal number of clusters. Classification and recognition were performed using four machine learning techniques i.e. decision trees, linear discriminant analysis, naive Bayes, and K-NN. The accuracy for decision tree, linear discriminant analysis, naive Bayes and K-NN were 62, 61, 73 and 100% respectively. Geometric features were also analyzed as to know which features were misclassified the most using each classifier (see Fig. 16).

To improve the overall accuracy of the proposed recognition technique for Urdu ligatures, we might need to apply preprocessing techniques such as noise removal on the ligature images. Similarly, in future research, the proposed ligature based clustering and recognition technique can be applied to real-world documents. In future, the proposed research robustness and effectiveness can be improved by using larger datasets as well as more robust geometric or morphological features. Since, deep learning is one of the best classifiers, applying it to this research will prove to be great contribution in the field.

References

1. Habash, N.Y.: Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies **3**(1), 1–187 (2010)
2. Olszewska, J.I.: Active contour based optical character recognition for automated scene understanding. Neurocomputing **161**, 65–71 (2015)
3. Kharma, N.N., Ward, R.K.: Character recognition systems for the non-expert. IEEE Can. Rev. **33**, 5–8 (1999)
4. Ahmad, R., Naz, S., Afzal, M.Z., Amin, S.H., Breuel, T.: Robust optical recognition of cursive Pashto script using scale, rotation and location invariant approach. PLoS ONE **10**(9), e0133648 (2015)
5. Choudhary, P., Nain, N.: A four-tier annotated urdu handwritten text image dataset for multidisciplinary research on Urdu Script. ACM Trans. Asian Low Resour. Lang. Inf. Process. **15**(4), 26 (2016)
6. Naz, S., Umar, A.I., Ahmad, R., Ahmed, S.B., Shirazi, S.H., Siddiqi, I., Razzak, M.I.: Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks. Neurocomputing **177**, 228–241 (2016)
7. Hakro, D.N., Talib, A.Z.: Printed text image database for Sindhi OCR. ACM Trans. Asian Low Resour. Lang. Inf. Process. **15**(4), 21 (2016)
8. Ahmad, Z., Orakzai, J.K., Shamsher, I., Adnan, A.: Urdu Nastaleeq Optical Character Recognition. In: Proceedings of World Academy of Science, Engineering and Technology, pp. 249–252 (2007)
9. Husain, S.A.: A multi-tier holistic approach for Urdu Nastaliq recognition. In: Proceedings of the 8th International Multi Topic Conference, Abstracts 2002, pp. 79–84 (2002)
10. Shah, Z.A.: Ligature based optical character recognition of Urdu-Nastaleeq font. In: Proceedings of 6th International Multitopic IEEE Conference (INMIC) (2002)
11. Husain, S.A., Sajjad, A., Anwar, F.: Online Urdu character recognition system. In: MVA2007 IAPR Conference on Machine Vision Applications (2007)
12. Khan, K., Siddique, M., Aamir, M., Khan, R.: An efficient method for Urdu language text search in image based Urdu text. IJCSI Int. J. Comput. Sci. Issues **9**(2), 523–527 (2012)
13. Razzak, M.I., Husain, S.A., Mirza, A.A., Belaid, A.: Fuzzy based preprocessing using fusion of online and offline trait for online Urdu script based languages character recognition. Int. J. Innov. Comput. Inf. Control **8**, 3149–3161 (2012)
14. Razzak, M.I., Anwar, F., Husain, S.A., Belaid, A., Sher, M.: HMM and fuzzy logic: a hybrid approach for online Urdu script-based languages' character recognition. Knowl Based Syst. **23**(8), 914–923 (2010). doi:10.1016/j.knosys.2010.06.007
15. Akram, Q.u.A., Hussain, S., Habib, Z.: Font size independent OCR for Noori Nastaleeq. In: Proceedings of Graduate Colloquium on Computer Sciences (GCCS), NUCES, Lahore (2010)
16. Javed, S.T., Hussain, S., Maqbool, A., Asloob, S., Jamil, S., Moin, H.: Segmentation Free Nastalique Urdu OCR. In: Proceedings of World Academy Of Science, Engineering and Technology, vol. 70 (2010)
17. Sattar, S.A., Haque, S., Pathan, M.K.: A finite state model for Urdu Nastalique optical character recognition. Int. J. Comput. Sci. Netw. Security **9**(9), 116 (2009)
18. Pal, U., Sarkar, A.: Recognition of Printed Urdu Script. Paper presented at the Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. 2 (2003)
19. Malik, S., Khan, S.A.: Urdu online handwriting recognition. In: Proceedings of the IEEE Symposium on Emerging Technologies, vol. 17(18), Islamabad (2005)
20. Chanda, S., Pal, U.: English, Devnagari and Urdu text identification. In: Proceedings of the International Conference on Cognition and Recognition, pp. 538–546 (2005)
21. Pathan, R.R.J.I.K., Ali, A.A.: Recognition of offline handwritten isolated Urdu character. Adv. Comput. Res. **4**(1), 117–121 (2012)
22. Zaman, S., Slany, W., Sahito, F.: Recognition of segmented Arabic/Urdu characters using pixel values as their features. In: ICCIT (2012)
23. Shahzad, N., Paulson, B., Hammond, T.: Urdu Qaeda: Recognition system for isolated Urdu characters. In: IUI 2009 Workshop on Sketch Recognition, Sanibel Island, Florida (2009)
24. Nawaz, T., Naqvi, S.A.H.S., ur Rehman, H.: Optical character recognition system for Urdu (Naskh Font) using pattern matching technique. Int. J. Image Process. **3**, 92–104 (2009)
25. Ahmad, Z., Orakzai, J.K., Shamsher, I.: Urdu compound character recognition using feed forward neural networks. In: ICCSIT 2009, pp. 457–462 (2009)
26. Shamsher, I., Ahmad, Z., Orakzai, J.K., Adnan, A.: OCR for printed Urdu Script using feed forward neural network. In: Proceedings of World Academy of Science, Engineering and Technology (2007)
27. Javed, S.T., Hussain, S., Maqbool, A., Asloob, S., Jamil, S., Moin, H.: Segmentation free nastalique urdu OCR. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 46, pp. 456–461 (2010)
28. Ahmed, S.B., Naz, S., Razzak, M.I., Rashid, S.F., Afzal, M.Z., Breuel, T.M.: Evaluation of cursive and non-cursive scripts using recurrent neural networks. Neural Comput. Appl. **27**(3), 603–613 (2016)

29. Javed, S.T., Hussain, S.: Segmentation based Urdu Nastalique OCR. In: Iberoamerican Congress on Pattern Recognition 2013, pp. 41–49. Springer, Heidelberg (2013)
30. Razzak, M.I., Husain, S.A., Mirza, A.A., Belaid, A.: Fuzzy based preprocessing using fusion of online and offline trait for online urdu script based languages character recognition. Int. J. Innov. Comput. Inf. Control **8**(5), 21 (2012)
31. Wali, A., Hussain, S.: Context sensitive shape-substitution in nastaliq writing system: analysis and formulation. In: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. pp. 53–58. Springer, Heidelberg (2007)
32. Hussain, S.: Complexity of Asian writing systems: a case study of Nafees Nasta'leeq for urdu. In: Proceedings of the 12th AMIC Annual Conference on e-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore 2003. Citeseer
33. Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., Madani, S.A., Khan, S.U.: The optical character recognition of Urdu-like cursive scripts. Pattern Recognit. **47**(3), 12291248 (2014)
34. Naz, S., Hayat, K., Razzak, M.I., Anwar, M.W., Akbar, H.: Arabic script based character segmentation: a review. In: 2013 IEEE World Congress on Computer and Information Technology (WCCIT), pp. 1–6 (2013)
35. Satti, D.A., Saleem, K.: Complexities and implementation challenges in offline Urdu Nastaliq OCR. In: Proceedings of the Conference on Language & Technology 2012, pp. 85–91 (2012)
36. Sabbour, N., Shafait, F.: A segmentation-free approach to Arabic and Urdu OCR. In: IS&T/SPIE Electronic Imaging 2013. International Society for Optics and Photonics, pp. 86580N-86580N-86512 (2013)
37. Akram, M., Hussain, S.: Word segmentation for Urdu OCR system. In: Proceedings of the 8th Workshop on Asian Language Resources, Beijing, China, pp. 88–94 (2010)

Naila Habib Khan is a Ph.D. scholar at Institute of Management Sciences, Peshawar, Pakistan. She received her BS (Computer Science) degree from Institute of Management Sciences, Peshawar, Pakistan in 2011 and MS (Information Technology) degree in 2014. She is a double Gold Medalist and has been awarded numerous merit scholarships during her academic career. Currently, she is working as a research assistant under the supervision of Dr. Awais Adnan at Institute of Management Sciences, Peshawar. Her areas of interest are Document Image Understanding, Pattern Recognition and Multimedia.

Awais Adnan is an Assistant Professor and Director ORIC (Office for Research Innovation and Commercialization) at Institute of Management Sciences, Peshawar. He teaches different courses at undergraduate, graduate and post-graduate level. Supervision of students at MS-IT, MS-CS and BS level are also part of his duties. Major areas of his research are multimedia, digital image processing and Network On Chip (NOC). He has also been working as a trainer in HRDC where he gives training on computer packages and data analysis tools to professionals from various government and public sector organizations.

Sadia Basar is a Ph.D. scholar at Hazara University, Mansehra, Pakistan. She completed her MS (Computer Science) degree from Institute of Management Sciences, Peshawar in the year 2014. She has five years teaching experience at university level. She has taught various courses at both graduate and undergraduate level. Currently, she is working as a research assistant under the supervision of Dr. Mushtaq Ali at Hazara University, Mansehra. Her areas of interest are Multimedia, Digital image processing, and Medical Image Segmentation.