International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Recognition of Nastaliq Urdu Text using Multi-SVM

Herleen Kour, Mehvish Yasin, Naveen Gondhi

Abstract: ***Optical Character Recognition has emerged as an attractive research field nowadays. Lot of work has been done in Urdu script based on various approaches and diverse methodologies have been put forward based on Nastaliq font style. Urdu is written diagonally from top to bottom, the style known as Nastaliq. This feature of Nastaliq makes Urdu highly cursive and more sensitive leading to a difficult recognition problem. Due to the peculiarities of Nastaliq Style of writing, we have chosen ligature as a basic unit of recognition in order to reduce the complexity of system. The accuracy rate of recognizing ligature in Urdu text corresponds to the efficiency with which the ligatures are segmented. In addition to extracting connected components, the ligature segmentation takes into consideration various factors like baseline information, height, width, and centroid. In this paper ligature Recognition is performed by using multi-SVM (Sup-port Vector Machine) approach which gives an accuracy of 97% when 903 text images are fed to it.***

Keywords : OCR, Nastaliq, Segmentation, Recognition, SVM

I. INTRODUCTION

In order to attain human like recognition ability in machines, the Pattern Recognition was introduced. Optical Character Recognition is the sub-branch of pattern recognition. In our day to day life, we come across situations where we want to reprint the modified text. But the editable document of text is not available. So, in such case the entire text needs to be typed which is quite exhaustive. This problem can be rectified with the help of OCR, which deals with the recognition of characters of text. It involves the photo scanning of text documents, their analysis and lastly their translation from text documents into text editable documents. The software that performs the Recognition is known as Optical Character Reader. In modern era, there is requirement of storing data for longer period of time, so as to make it searchable and accessible for future and it is achieved by making the information available in digital form. With the help of OCR, the human ability can be imitated in order to extract text from image. There are various tools and techniques on which the OCR is based on. Therefore, the accuracy level is obtained and the result is generated. Urdu language is one of the popular languages of South Asia. It is one of the 23 social languages of India It is spoken by millions of people all over the world. There are two basic styles of writing Urdu, Naskh and Nastaliq. Urdu is written in Nastaliq style conventionally The Urdu Script contains 38 letters which when joined together make words of the language. Diacritics are known as dots/marks that combine with characters to distinguish them from one another.

Manuscript published on January 30, 2020.
* Correspondence Author
Herleen Kour*, Computer science and engineering,Shri Mata Vaishno Devi University, Katra, India. Harleenkour700@gmail.com
Mehvish Yasin, Computer science and engineering, Shri Mata Vaishno Devi University, Katra, India. 17mms013@smvdu.ac.in
Dr. Naveen Gondhi, Computer science and engineering,Shri Mata Vaishno Devi University, Katra, India.Naveen.gondhi@smvdu.ac.in

II. PECULARITIES AND CHALLENGES

Table- I: Peculiarities and challenges faced by Urdu Script

Challenge	Property	Example
Directionality	For reading and writing text: right to left.	حامدکاجنم دن۱۲ کوہے-
	for reading and writing the number: left to right	۱۱،۱۲،۱۳
Non-Monotonicity	Linking of characters is done.	بج
Cursive	Characters connected to form ligatures.	ک+ا+م=کام
Character Overlapping	Characters are seen overlapped vertically	آج بارش ہے-
Uppercase/ Lowercase	No uppercase or lowercase	ا،ب،پ،ت
Stretching	Characters change their default shape	ش،س
Positioning	ligature are placed on top of the previous ligature	محسنا خاموش پاکستا
Spacing	The spaces may vary in size.	میرٗانام ہرٗلین ہے

III. IITERATURE SURVEY

Urdu text is widely printed in Nastaliq Style of writing. A detailed survey carried out by CRULP showed that almost all of the books are published in Noori Nastaliq font.

The complexity of Urdu such as context sensitivity and cursive style of writing makes the segmentation of characters di cult as compared to Latin Script.

Therefore, rather than using character segmentation we have used the higher unit also known as ligature segmentation. A segmentation free approach was put forward by Hussain in 2002 [1] in which isolated characters were used as units of recognition. The classification was done based on the centroid to centroid distance and labeling of connected components. The recognition yielded an accuracy of 100% however; it could not recognize compound characters. In 2007, I Shamasher, Z. Ahmad, J.K Orakzai, and A. Adnan developed OCR classifying characters with the help of Neural Network thus acquiring an accuracy of 98.30%, \OCR for Printed Urdu Script using Feed Forward Neural Network" [4]. Dataset is fed to the network at input and after passing through the network, an output is generated. Supervised learning is used to train feed forward neural networks. .

In another study [8], the authors binarize gray scale document images and extract text-lines using horizontal projections. White spaces are removed using a set of heuristics and the dots are associated with their corresponding primary ligatures. An accuracy of 97% is reported by this system. In 2009, Ahmed Muaz [5] published segmentation based thesis "Urdu Op-tical Character Recognition System" as an

enhancement in each module of [6]. It can be analyzed that this technique showed a great success with 97% accuracy in segmentation,92% in Recognition and 88% in post processing but after analysis and investigation, lots of segmentation errors are seen for large test samples and longer ligatures. Some particular classes are difficult to be recognized by the recognition.Some of the main errors encountered in this technique are summarized as:

Errors caused due to Thinning process are:

- Absence of junction point: When characters undergo thinning pro-cess they are reduced to a single pixel stroke, hence they cannot be recovered.
- Change in alphabet: Sometimes, characters in a ligature are thinned to some other characters.
  Errors caused due to association of diacritics with their corresponding seg-ments.
- Multiple Diacritic characters: In Ligatures containing two or more dot holding characters, the Diacritics don't get associated with their corresponding segments due to accommodation of all diacritics of Ligature.
- Overlapping segments: The Ligatures containing overlapping char-acters makes recognition di cult.Thus Character Segmentation seems to be a di cult task for researchers. Due to the challenges in character segmentation, most of the researchers have taken the next higher unit, ligature, as recognition unit.

In 2010, Holistic Urdu Handwritten Word Recognition using Support Vector Machine (SVM) was put forward [3]. SVM is a supervised Machine Learning Algorithm that can be used in order to perform classification and regression

while it is mostly used for regression purposes. In this algorithm, each data item is plotted with the value of each feature being the value of each coordinate. Then we perform classification by finding the hyper plane that differentiates two classes very well. In Optical Character Recognition, SVM can be used to extract two different feature sets by performing classification. The extraction of structural and gradient features which constitute the compound feature set is done on each Urdu ligature/word. Gradient feature extraction is made which is a mathematical way of finding direction in any dimensional space. The strength and direction of each pixel is calculated. The other processing steps can be followed like quantization, Gaussian filters etc. Structural Feature Extraction provides the physical attributes of an entity or images which include topological features, projection pro les. Upper and lower features can be captured to provide the framework of the top and bottom parts of word. The image is converted into a 2-dimensional array and the distance from top to the rst ink pixel is calculated and similarly the distance from bottom to the last ink pixel is calculated to extract topology, projection. In this way di erent characters were classi ed.Israr Ud Din, Imran Siddiqi, Shehzad Khalid and Tahir Azam presented a Holistic approach for printed Urdu Nastaliq font in 2017. About 1525 unique statistical features were extracted and recognized with the help of HMM [10].

The OCR system either takes into consideration the segmentation techniques or individually characterizes the character primary and secondary characters [1] [2]. In this technique, compound characters cannot be recognized. The recognition of compound characters was made by Bansal and Sinha [3] while working on Devanagari Script with an accuracy rate of 93%. From table 1, it can be seen that recognition of characters and ligatures yield an accuracy of more than 90%. However, less efforts are made to associate the primary and secondary components. The comparative analysis of various techniques show that the maximum accuracy can be attained by using segmentation approach followed by recognition of characters by neural networks. The BLSTM provides a better result than HMM. With the help of unique ligatures, an accuracy rate of about 92% can be attained by using structural and statistical feature extraction techniques [6]. However, this recognition technique works for fixed ligatureizes only. The handwritten characters and therecognition provides an accuracy rate of (94-97%) and 96.80% respectively. While using a holistic approach [9], the units of recognition are primary and secondary components with an accuracy rate of 88.87%. In analytical approach, the CNNs provide a better accuracy rate (98.12%) [12] than RNNs (94.97%).The combined effect of both [10] produce an accuracy of 96.40%. We can infer from the above table that in spite of having some loopholes in the various approaches of OCR, the researchers are still devising novel ideas and algorithms which can yield better results

Approach	Feature set	Classification	Unit of recognition	Accuracy	Remarks
Segmentation [1]	Implicit, Explicit And Mixed Strategies	Centroid To Centroid Distance And Labeling Of Connected Component	Isolated Characters	100%	Do not recognize compound characters
Compound Character, Recognition [3]	Statistical	Hybrid	Characters	93%	Isolated characters can be recognized only
Segmentation [2]	Implicit, Explicit	Component Labeling	Isolated Characters	96.90%	Syntactic, semantic constraints need to be incorporated
Analytical [4]	Custom	Neural Networks	Isolated Characters	98.30%	Ignores diacritics
Real World Images[5]	Corpora	Feed Forward Neural Network	Isolated Characters	93.4%	Fixed size ligatures only
Unique Ligatures Recognition [6]	, Transformational, Structural, Statistical	HMMS, Template Matching	Primary And Secondary	92%	MLSTM provide better
Analytical [7]	UPTI	Bidirectional LSTM	Characters	95%	Font size invariant
High Frequency Ligature	Projection, Concavity, Curvature	HMMS	Ligatures	97.93%	No association of primary and secondary components
Analytical[10]	Statistical	CNNS	Characters	98.12%	Do not recognize compound character
Holistic[9]	Structural, Statistical	HMMS And DWT	Unique Primary And Secondary Component	88.87%	No work to show performance on manual features
Analytical [11]	Multi-Dimensional LSTM And Statistical	CNNS, RNNS	Characters	94.97%	Better performance as compared to cnns

IV. PROPOSED METHODOLOGY

The process of OCR is followed in three steps:

1. Segmentation
2. Feature Extraction
3. Recognition

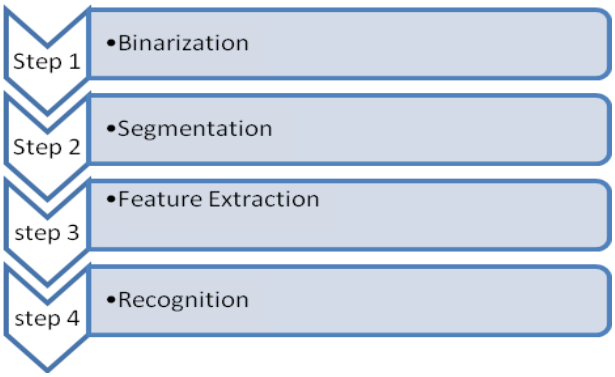

Fig:1 Generic Flow of OCR process

We perform the segmentation carried out at two levels, line and ligature. We first apply Binarization using global threshold and text line segmentation. Then we extract the ligatures for each segmented text line.

In Nastaliq Style of Urdu Script, the characters are combined completely which results in more challenging segmentation of characters. There are various segmentation approaches for ligatures, which can be classified into top-down, bottom-up and hybrid. In top-down approach, the image is divided into text lines and words/characters assuming them to be straight lines. In bottom-up approach, a clustering process is followed, hence the observation starts with small units of pixels, characters, words, text lines and pairs of components are merged as one moves up the hierarchy. The hybrid approach combines the top-down and bottom- up approach in various ways. The positions of the piece- wise separating lines can be obtained by using the horizontal projection. In Urdu Script, the global horizontal projection method includes the problems of over segmentation and the under segmentation. Ligatures are classified into primary and secondary ligatures.

The main body is represented by primary ligature while, the secondary ligature is represented by diacritics/dots corresponding to the primary ligatures. In addition to this, we perform character segmentation and recognize them with the help of multi SVM.

1V.1 Binarization

The first step in Optical Character Recognition is Binarization. A threshold value is selected locally or globally for each pixel or an image respectively. Thus, the image is classified into foreground and background pixels. If any errors are introduced during this step, the recognition will not be accurate. Incorrect Binarization can result in loss of information or joining of characters of a text. In Nastaliq font, we have a large numbers of ligatures and diacritics. Therefore two important points should be kept in mind.

- Threshold values should not be too high because having high level of threshold tends to break the continuous elements in the text leading to over-segmentation.
- Threshold values should not be too low because having low level of threshold tends to merge or join isolated components in text leading to under-segmentation.

1V.11 Segmentation

Segmentation is the process of division of content of paragraph into lines which are further segmented into words or sub-words. The sub-words/ligatures are made such that they can be processed easily.

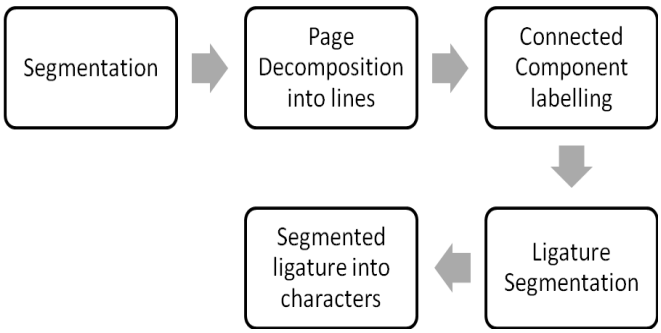

Figure: Segmentation Process

A. Page Decomposition into Lines: Firstly, the segmentation points need to be detected. For this, we analyze the row with maximum pixels (local peaks) and row with minimum pixels (local valleys). Hence, by counting the number of pixels in each row of image (horizontal projection pro le); we get the number of text lines of an image.

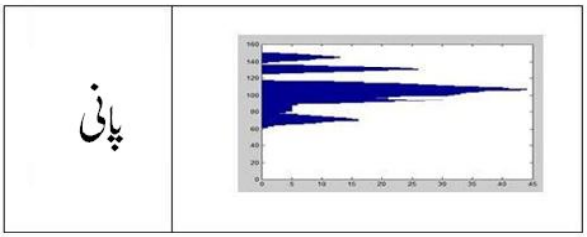

Figure 1: Horizontal projection profile

To overcome the challenge of over segmentation and under segmentation, we put forward steps of Modified Horizontal Projection for line in order to segment the text line. The Steps are discussed as:

- For image “pani” we Segment the image into horizontal zones with the help of horizontal projections and calculate the height of each row.
- Segment the zones by using estimated row height.

Once the row height has been calculated, we use

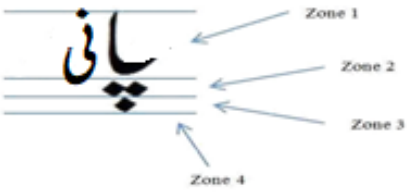

Figure 2: Multiple zones of a text

these row heights to label zones. We can either have Type 1 or Type 2. Zones depending upon the placement of diacritic marks that is, zones which have dots either above or below the ligatures and zones having underlines, etc. The zones with multiple text lines can be labeled as zone 3. Next, we merge the zones, depending upon whether they satisfy the criteria for Type 1 or Type 2.

Zone	Height	Width	Type
1	35	62	1
2	9	18	2
3	7	9	2
4	1	62	2
1	35	62	1

Table 3: Information of zones for image in Fig.2

Table 3 shows the different zones for figure 2 where the corresponding heights and widths are calculated. The zones are either of type 1 or Type 2 depending upon their heights. In other words, if a zone has a height lesser than half row height, it is concluded that the zone belongs to type1. The zone containing diacritics is considered to be Type 2, irrespective of whether the diacritics lie above or below the baseline. Zones having heights 1.5 times as compared to row height are considered as Type 3.

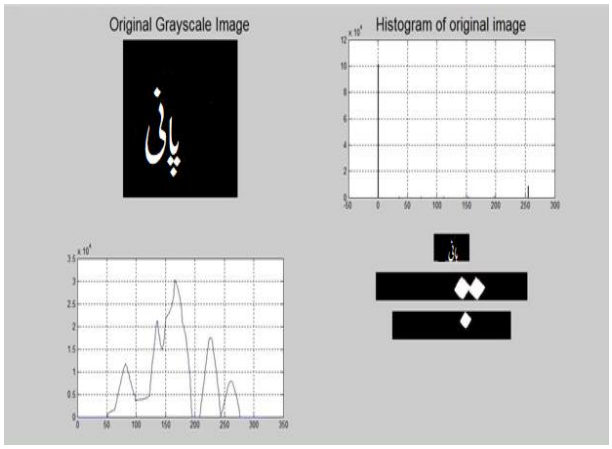

Figure 3: Horizontal Segmentation

Thus, in this way we can perform the segmentation using estimated row height. But in some cases, if an image possesses noisy component, it results in under- segmentation. Therefore, we apply traditional horizontal projection method and undergo rough segmentation of binarized image. There might be a problem with associated dots and diacritics and missegmentation. In order to overcome such problems, we apply morphological dilation to the document image such that the primary and secondary ligatures seem to be joined. The text line boundary which exists between the sequential local peaks (valley index) can be found from local peaks in dilated version of an image which in turn can be detected with the help of median zone height that acts as threshold for finding

B. **Connected Components Segmentation:** We have chosen ligature which is a higher unit of recognition as basis for segmentation. Ligature/connected component is an isolated character or combination of characters when joined together. One of the important challenges in Nastaliq Urdu inter-ligature overlapping, as a result of which we cannot use vertical histogram projection in order to segment ligatures. Thus, we have used the connected component labeling .The components that are extracted are represented by a different colour and a rectangle around it. We can either have a 4-connected neighbour or 8-connected neighbour based on its position. Any neighbour pixel is said to be 4-connected if it is located to immediate left, right, top, bottom positions of that pixel. A neighbour pixel is said to be 8-*connectedif* it is located to immediate top, top-right, bottom, bottom-left, left or top-left positions of that pixel. If two pixels are connected, they will always belong to the same component and will be assigned same label. On the other hand, if there is no connected path between any two pixels, it means they belong to different components as a result of which we can travel between them without crossing the background pixels.

C. Ligature Segmentation: Mostly, ligature is used as a fundamental unit of recognition. This is because of the reason that the segmentation of word into characters is challenging due to word spacing. There are 3 steps in ligature segmentation:

Division of Line into Ligature- Most of the systems use projection profile method in order to calculate vertical histo gram of text lines. Since the ligatures overlap in Nastaliq style, therefore this method cannot be applied. A better method is to split text lines into connected components by bounding box and extract the information.

Figure 6: Input image with noise

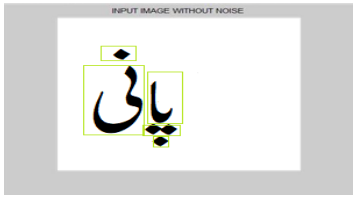

Figure7: Input image without noise

Figure 8: Segmented ligature

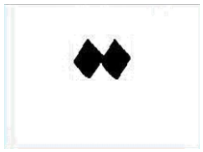

Figure 9: segmented Diacritic

Figure 10: Segmented Diacritc

Figure 11: Segmented Ligature

Figure 12: Segmented Diacritic

D. Identification of Base Forms- The various characteristics like height, width, centroid, overlapping, co-ordinates and baseline information are considered as per the connected components which in turn can be categorized into primary and secondary ligatures. The horizontal projection of pixels can be used in order to end the distinction between the ligature base and the diacritics thereby, providing the baseline measures. The row that contain maximum amount of ink pixels is considered to be the baseline. Sometimes, incorrect baseline may occur. Such errors can be rectified by checking a couple of heuristics. The primary rule in case of Nastaliq Style is that every ligature should be in close vicinity of the baseline, thereby touching it. Also, the baseline must be in the lower half of the line/ligature. Therefore, the false segmentation and identification can be reduced.

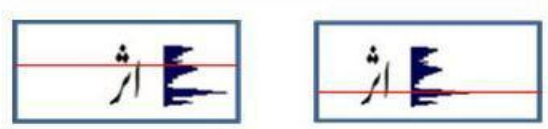

Figure 13: false baseline and corrected baseline

E. Base and Mark Association-Formation of ligatures by association of secondary components is the major challenge. The components are concluded to be primary depending upon height. A threshold value is defined and if the height of component is greater than threshold, then the component is primary, else the component is secondary. Another way is by using centroid-to-centroid distance after calculating the centroid for each shape and then projecting them vertically in order to form association Although this method is simple, but in some cases, it does not work reasonably well and does not provide accurate results. The reason for this is that the centroid of diacritics do not associate with the right base forms because of shifting of letters to left or right due to the context sensitivity characteristic of Nastaliq style of Urdu script. Therefore, instead of current ligature, the diacritics project to the previous letter/ligature.

Figure 14: Diacritics of character tay projecting on previous ligature

In order to address such issues, the complete horizontal span is made with respect to the diacritics, which is then associated with the base form. Following steps are followed:

- The secondary are joined with primary components with the help of vertical structuring elements by using morphological dilation.
- The extraction of connected components from dilated image is performed.
- If secondary components (dots) are combined with only one primary component, then they are associated with the particular primary component

There might be a situation where a dot is associated with more than one base forms (overlapping), in such a case, it is associated with the left side of the ligature. Similarly, if the diacritic forms a complete overlap with respect to multiple base forms, then the distance of the diacritics with each of the ligatures is calculated and is associated with the one having the lesser distances.

Figure 15: Text line

Figure 16: Extracted ligature

Once we have segmented text into words/ligatures, next we will segment ligatures into sub-words/characters. Segmenting text into characters is quite challenging task because of the cursive nature of Urdu Script. In order to segment ligature into characters, segmentation is performed by two ways:

- **Vertical Segmentation**-Corresponding to each column, a vertical histogram is plotted for each ligature which further corresponds to minimum intensity pixel values. A vertical line is drawn from these segmentation points which in turn segments ligature into characters

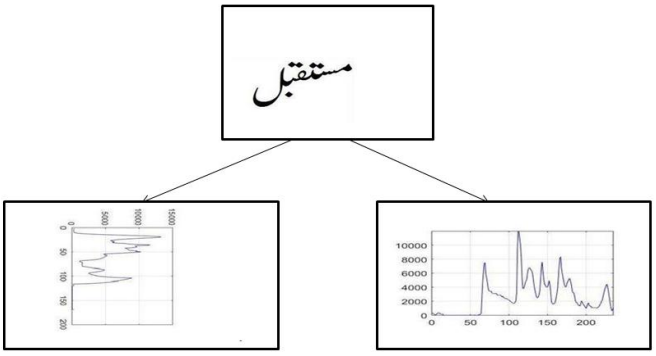

Figure 17: Horizontal and Vertical projection of word "Mustaqbil"

- **Horizontal Segmentation**- Corresponding to each row, a horizontal histogram is plotted for each ligature which further corresponds to minimum intensity pixel values. A vertical line is drawn from these segmentation points which in turn segments ligature into characters. The segments are further taken as input to vertical projection and each horizontal segments of ligature are fed as image to vertical projection pro le and thus, we get the characters.

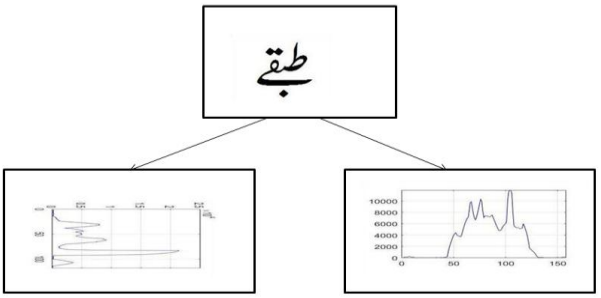

Figure 18: Horizontal and Vertical projection of word "tabke"

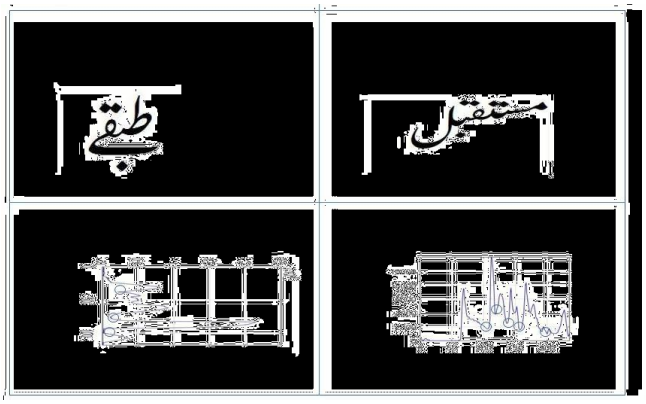

Figure 19: Segmentation points

We can infer from the projection profile that there are some minima points. These points can be used in order to detect the segmentation points

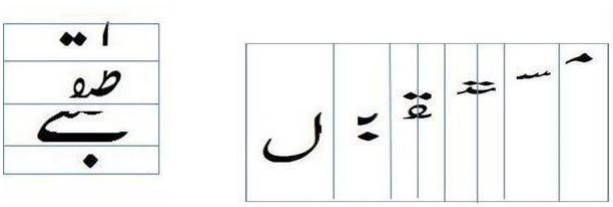

Figure 20: Segments obtained

- **Diacritic Separation**-The diacritic separation is performed by connected component labeling. While we plot horizontal or vertical projection profile, it is an easy task to separate diacritics from their respective components because of the clear demarcation of white spaces between the two. In order to identify diacritics later in post processing, a sequence number is associated with each of them corresponding to the components. Here diacritic separation is done with the help of connected component labeling.

Figure 21: Original image

Figure 22: Diacritics separation

1V.111 Feature Extraction:
From each image two different type of features are extracted "Gradient features" and "Structural features" with different sets various experiments are conducted and found that the combination of these two sets produces good results.

1.Gradient feature extraction :Gradient features are directional features and can be extracted from the gradient of the gray scale image. A gradient vector of discrete direction is decomposed into two components by specifying number of standard directions. In our study after preprocess scale Roberts filter mask were applied on images.

Let i "x ,y" be an input image, the horizontal and the vertical gradient components. After applying the Robert filter mask were calculated as:

$G(x)=i(x+1),(y+1)-i(x, y)$
$G(y)=i(x+1),y-i(x,y+1)$

The strength is given as $F(x, y)= \sqrt{G(x)^2+G(y)^2}$

Direction: $\theta(x , y)=\tan^{-1}(Gy/ Gx)$

The direction of the vector is returned in the range of -Π to Π.

The Gradient image was divided into 81 blocks with a vertical and a horizontal blocks.

For each blocks, the gradient strength was accumulated in 32 directions. The total size of the feature set in the feature vector 9*9*32= 2592.

By down sampling the number of blocks from 9*9 to 5*5 Gaussian filter was applied to reduce the spatial resolution and the direction was reduced to 16.

2. Structural feature extraction : The physical attributes of an image is called structural feature. It includes projection profile upper and lower profile and so on. The shape of the printed text is captured by upper and lower profile features. Each word is converted into a two dimensional array. Each column was examined from right to left. The distance from top of an image to the first ink pixel was calculated. The value of the distance in the range of 0 to 64 was returned.

In order to make all the features of the same type and in the same scale a variable transformation was applied on all the features.

$$Y=x^{0.4}$$

Advance in static learning theory and increase in computer processing in recent years, lead the development techniques such as SVM.

Over years, SVMs were successfully used in many applications like image classification, handwriting recognition SVM are a set of supervised learning methods that are used for classification and regression. It has better empirical performance. In addition to performing linear classification SVM can efficiently perform non-linear classification using Kernel trick.

The main objective of SVM is to find the best separating hyper plane that provides the highest margin distance between the nearest points of two classes called Functional margin.

This approach guarantees that larger the margin lower is the generalization error of the classifier. The multiclass SVM approach aims to assign labels to a finite set of several elements based on set of linear or non-linear SVMs.

1V.1V Recognition: Advance in static learning theory and increase in computer processing in recent years, lead the development techniques such as SVM.

Over years, SVMs were successfully used in many applications like image classification, handwriting recognition SVM are a set of supervised learning methods that are used for classification and regression. It has better empirical performance. In addition to performing linear classification SVM can efficiently perform non-linear classification using Kernel trick.

The main objective of SVM is to find the best separating hyper plane that provides the highest margin distance between the nearest points of two classes called Functional margin.

This approach grantees that the larger the margin lower is the generalization error of the classifier. The multiclass SVM approach aims to assign labels to a finite set of several elements based on set of linear or non-linear svm.

The Multiclass SVM approach aims to assign labels to a finite set of several elements based on a set of linear or non-linear basic SVMS. The most popular approaches in the literature are to reduce the single multiclass problem into multiple binary problems. By doing so, each problem can be seen as a binary classification which is assumed to produce an output function that gives relatively large values for examples that belong to the positive class and relatively small values for the examples that belongs to the negative class. The two common approaches to build such binary classifiers which are trained to distinguish:

i. One of the labels against all the remaining labels (known as one-versus-all)

ii. Every pair of classes (Known as one-versus-one) Classification of new instances for one-versus-all case is done by a winner takes-all strategy, in which the classifier with the highest output function assigns the class. The classification of one-versus- one case is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally, the class with more votes determines the instance classification.

Classification and recognition is the last step in OCR. It is used in order to classify the input into one of the output classes .Recognition involves 2 phases:

Training Phase: The first phase of recognition in which the feature vectors are fed to the system along with their respective labels. In this way, the system is made to learn the relations between the input data and the unique classes. These relations are then used to decide the classes for objects with missing label. Multi-SVMs are used in order to recognize the ligatures in our study. We train multi-SVMs to learn the Unicode corresponding to ligatures.

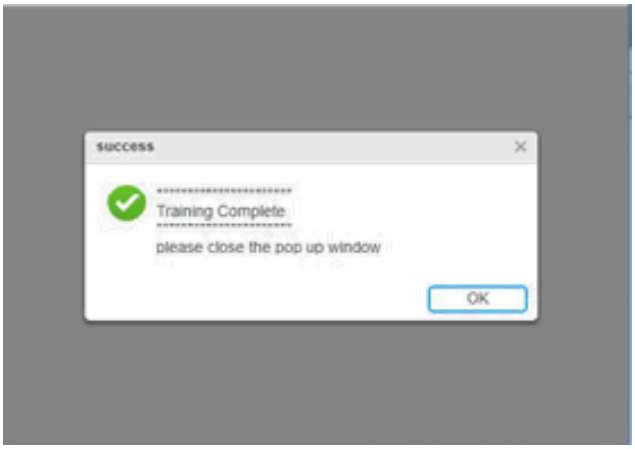

Figure 23: Training

Finally, the Unicode binary output is converted to the corresponding character. The first phase of recognition in which the feature vectors are fed to the system along with their respective labels. In this way, the system is made to learn the relations between the input data and the unique classes

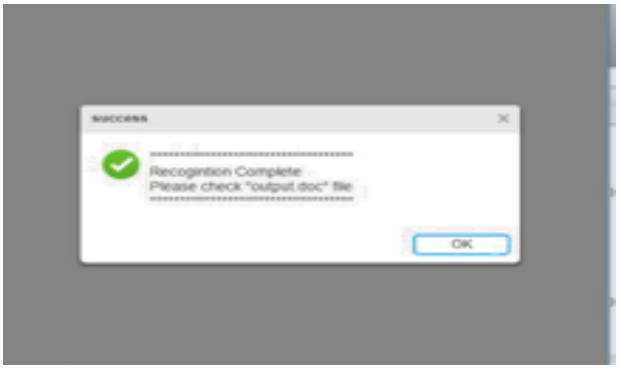

Figure 24: Recognition

- **Diacritics association**-After ordering of segments, the diacritics are associated with each segment. This can be done by mapping the Unicode to the different diacritics thereby, recognizing them

V. RESULTS

We tested ligature segmentation approach on 315 ligatures and it successfully segmented 89% connected components. It is not able to segment 36 ligatures thus, error percentage is 11%. The recognition rate is 97%. The results are illustrated.

	Segmentation	Unpredicted	Total
Ligature Tested	300	15	315
Segmentation Failure	31	5	36
Error	10%	33%	11%
Accuracy	90%	67%	89%

Fig 25: Segmentation results

Samples Trained	**637**
Samples Tested	**266**
Recognition Failure	**11%**
Error	**3%**
Accuracy	**97%**

Fig 26: Recognition results

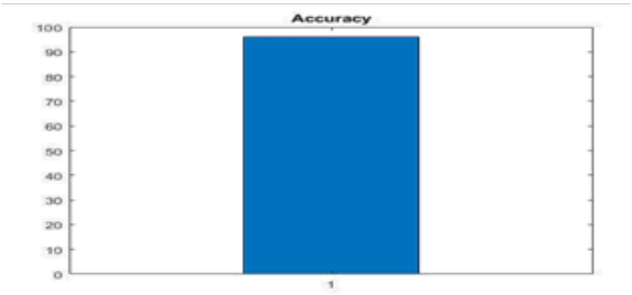

Figure 27: Accuracy

V1: CONCLUSION

Segmentation technique is explored thoroughly in this study and its pros and cons have been revealed. We presented both holistic and analytical approaches for Segmentation of Nastaliq Urdu text that constitutes an important step in Urdu Character Recognition. We have also discussed the placement of dots and diacritics in a more accurate fashion. In addition to this, we have taken into consideration height, width, baseline, forligature segmentation. The peculiar nature of Nastaliq style makes character recognition more difficult. More e orts are put forward by researchers for images of printed texts rather than hand-written text whether online or offline. Till now, there no multilingual algorithms that include unlimited database as there is high resemblance between Arabic content languages.

FUTURE SCOPE

1. One of the important tasks in OCR is preprocessing steps like noise removal, skew detection and correction.
2. The accuracy of recognizing ligatures can be increased by developing some preprocessing engine.
3. In order to accommodate a variety of fonts and sizes, a module can be developed which is independent of size.

This technique should be further refined such that the accuracy of segmentation increases.

REFERENCES

1. S. A. Hussain, International Multitopic Conference on Abstracts, Karachi, pp. 528-532, 2002.
2. U.Pal and A. Sarkar, International Conference on Document Analysis and Recognition. vol.2, p. 1183, 2003
3. S. Javed, S. Hussain, A. Maqbool, S. Asloob, S.Jamil, and H. Moin, World Acad.Sci. Eng.Technol. vol. 46, pp. 456-461, 2010
4. Shamsher, Z. Ahmad, J.K Orakzai, and A. Adnan, World Acad.Sci. Eng.Technol. vol. 23, pp.172-175, 2007 .
5. A. Muaz, "Urdu Optical Character Recognition System," MS Thesis Report, National University of Computer and Emerging Sciences, 2010
6. Z. Ahmed, J.K.Orakzai and I. Shamsher ,International Conference on Computer Science and Information Technology, pp. 457-462, 2009
7. Adnan Ul-Hasan, Saad Bin Ahmed, Faisal Rashid, Faisal Shafait, and Thomas, M Breuel, International Conference on Document Analysis and Recognition, pp.1061-1065, 2013
8. Israr Uddin Khatak, Imran Siddiqui, Shehzad Khalid, ad Chawki Djeddi, International Conference on document analysis and recognition, pp.71-75, Tunisia, 2015.
9. Bansal V and R.M. K Sinha, Syst. Man Cybern Syst. Humans.vol. 30, pp. 500-505, 2000.
10. Israr Uddin Khatak, Imran Siddiqui, Shehzad Khalid, International Conference on Frontiers of Information Technology, pp.155-160, 2017.
11. Saeeda Naz, Arif I Umer, Riaz Ahmad, Saad B Ahmed, Syed H Shirazi, Umer Siddiqui and Muhammad I Razzak, Neurocomputing. vol. 177, p. 228, 2016.
12. Saeeda Naz, Arif I Umer, Riaz Ahmad, Saad B Ahmed, Imran Siddiqui, Muhammad I Razzak and Faisal Shafait, Neurocomputing. vol. 243, pp. 228-241 , 2017.
13. A. Daud, W. Khan, and D. Che, \Urdu language processing: a survey", Arti cial Intelligence Review, vol. 47, pp. 1{33, June 2016.
14. U.Pal and A. Sarkar, International Conference on Document Analysis and Recognition, vol. 2, pp. 1183{1187, 2003
15. A. Ul-Hasan, S. B. Ahmed, F. Rashid, F. Shafait, and T. M. Breuel,, in Document Analysis and Recognition (ICDAR), pp. 1061{1065, IEEE, 2013
16. I. U. Din, Z. Malik, I. Siddiqi, and S. Khalid, J. Appl. Environ. Biol. Sci, vol. 6, pp. 114{120, 2016.
17. Ibrar Amad, Xiaojie Wang, Ruifan Li, Manzoor Ahmed, Rahat Ullah, IEEE Access, vol. 5, pp. 10924-10940, 2017
18. H. Malik and M. A. Fahiem, \Segmentation of printed Urdu scripts using structural features," in Visualisation, 2009. VIZ'09. Second International Conference in, pp. 191{195, IEEE, 2009.
19. K. S. Kumar, S. Kumar, and C. Jawahar, in Document Analysis and Recognition, vol. 9, p. 141, 2007.
20. T. M. Breuel, in International workshop on document analysis systems, pp. 188{ 199, Springer,2002.
21. S. S. Bukhari, F. Shafait, and T. M. Breuel, in Document Analysis and Recognition (ICDAR), pp. 748{752, IEEE 2013.
22. S. T. Javed and S. Hussain, in Multitopic Conference INMIC, pp. 1{6. IEEE, 2009.
23. Gurpreet Singh Lehal in Document Analysis and Recognition (ICDAR), pages 1130-1134. IEEE, 2013
24. S. Hussain, S. Ali, et al., International Journal on Document Analysis and Recognition (IJDAR), vol. 18, no. 4, pp. 357{374, 2015.
25. Hande Adiguzel, Emre Sahin, and Pinar Dugulu. A hybrid approach for line segmentation in handwritten documents. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, pages 503-508.IEEE, 2012.
26. Shuwair Sardar and Abdul Wahab. Optical character recognition system for Urdu. In Information and Emerging Technologies (ICIET), 2010 International Conference on, pages 1-5. IEEE, 2010
27. Sankaran, N., Jawahar, C.V.: Recognition of printed Devanagari text using BLSTM Neural Network. In: 21st international conference on pattern recognition, 2012.
28. Saad Bin Ahmed, Saeeda Naz, Muhammad Imran Razzak, Sheikh Faisal Rashid, Muhammad Zeeshan Afzal, Thomas M. Bruel, Evaluation of cursive and non cursive scripts using recurrent neural networks, vol. 27, pp.603-613, 2015.
29. Sa a Shabbir and Imran Siddiqi, Optical Character Recognition System for Urdu Words in Nastaliq Font in International Journal of Computer Science and Applications. Vol.7 no. 5, 2016

AUTHORS PROFILE

Herleen Kour, has done B.tech from Baba Banda Singh Bahadur Engineering college, Punjab. She is pursuing M.Tech from Shri Mata Vaishno Devi University ,Katra(J&K).She has published research papers in Springer and IEEE conferences. She has a research paper in Scopus Indexed Journal. Her area of interest includes digital image processing, machine learning, deep learning.

Mehvish Yasin, has done B.tech from Baba Ghulam shah Badshah University, Rajouri (J&K). She has done M.Tech from Shri Mata Vaishno Devi University, Katra (J&K). Her area of interest includes digital image processing, machine learning, deep learning. She has several research papers in conferences and journals.

Dr. Naveen Kumar Gondhi, received the Ph.D. degree in computer science. He is currently an Assistant Professor with Shri Mata Vaishno Devi University, Katra, India. He has several research papers in national/international journals to his credit. His research areas include computer network management, expert systems, mobile computing, and cluster and grid computing.