IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 6, 2018, accepted November 13, 2018, date of publication January 7, 2019, date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2887103

Improving Urdu Recognition Using Character-Based Artistic Features of Nastalique Calligraphy

QURAT UL AIN AKRAM AND SARMAD HUSSAIN
Department of Computer Science and Engineering, University of Engineering and Technology at Lahore, Lahore 54890, Pakistan
Corresponding author: Qurat ul Ain Akram (ainie.akram@gmail.com)

ABSTRACT The state-of-the-art Urdu recognition approaches for Nastalique use features along with the sequence of characters' labels for classification and recognition. In Arabic-like cursive script, the characters are joined together to form a ligature. The conventional methods process the connected stroke of ligatures as a sequence of characters. However, connected stroke of a ligature image has a sequence of pairs of characters and their joiners, instead of a sequence of characters. The character has a distinctive shape that clearly distinguishes it from other characters. The joiner preserves the connecting stroke shape of a character with the next character. In this paper, an implicit Urdu character recognition technique is presented for the Nastalique writing style that is based on recognition of characters and joiners. The detailed analysis of the Nastalique calligraphy is carried out to extract the artistic features of characters and their joiners. The presented technique is tested on Dataset-1 of 1446 ligature classes covering 3 309 762 ligature instances and 91 129 unique Urdu words. In addition, the system is also tested on 1600 text lines of UPTI dataset called Dataset-2. The character recognition accuracies are 95.58% and 98.37% on Dataset-1 and Dataset-2, respectively. The results reveal that the system outperforms the state-of-the-art hidden Markov models and deep learning-based Urdu recognition techniques.

INDEX TERMS OCR, sequence learning, Nastalique, implicit recognition, character based recognition, character and joiner sequence.

I. INTRODUCTION

The tremendous advancement of Information and Communication Technologies (ICT) in the local content availability highlights the significant need of porting Urdu published content online in the form of editable text so that it can be searched and retrieved by the local users. Most of the Urdu content is in the form of published books, magazines and other documents. The digitization of these documents is important to make it available online. To convert this published material into digital content efficiently, the development of Optical Character Recognition(OCR) system is important. Researchers are actively working on analysis of script characteristics and developing the recognition systems of Urdu document images for more than a decade. However, most of published research focuses on the development of a recognition module of OCR.

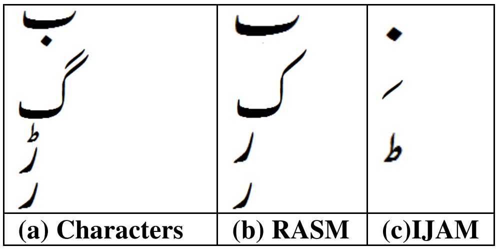


FIGURE 1. Urdu Characters and corresponding RASM and IJAM. (a) Characters. (b) RASM. (c)IJAM.

Urdu belongs to the Arabic script which is cursive in nature. The development of OCR for document images of Arabic script languages is a challenging task. In cursive script, one or more characters are joined together to form ligature, and ligatures are grouped to form words. Each ligature has two parts; (1) RASM and (2) IJAM [1]. RASM also called main body, is the main stroke without the associated diacritic, see Figure 1(b). IJAM are mandatory diacritics used to disambiguate the same RASM having different consonant behavior as can be seen in Figure 1(c). A RASM can or cannot contain IJAM which is dependent on consonantal behavior.

Alif Class	ا		FEH Class	ڡ	ف
BEH Class	ٮ	بپتٹث	QAF Class	ٯ	ق
JEEM Class	ح	جچحخ	KAF Class	ک	گک
DAL Class	د	دڈذ	LAAM Class	ل	ل
REH Class	ر	رڑزژ	MEEM Class	م	م
SEEN Class	س	سش	NOON Class	ں	ن
SUAD Class	ص	صض	WAO Class	و	و
TOAY Class	ط	طظ	HEY Class	ہ	ہ
AIEN Class	ع	عغ	HEY DO-CHASHMY Class	ھ	ھ
YEY Class	ی	ی	HAMZA Class	ء	ء
YEY-BARI Class	ے	ے			
(a) RASM Class Name	(b)RASM Shape	(c) Characters having same RASM	(a) RASM Class Name	(b)RASM Shape	(c) Characters having same RASM

FIGURE 2. Urdu RASM classes.

The RASM of character ر (REH) does not have any IJAM whereas RASM of character ڑ (Rreh) has an IJAM which is indicated in Figure 1(c). The Urdu character set has a total of 21 unique RASMs which are given in Figure 2. Figure 2(a) specifies name of RASM, Figure 2(b) indicates shape of RASM of Urdu character, and all characters with possible variations of diacritics of same character class are shown in Figure 2(c).

Most of published Urdu content i.e. Urdu documents, newspapers and books is written using Nastalique writing style. This writing style was developed in the 14th century CE in Iran by combining the rules of two writing styles, Naskh and Ta'līq [2]. Due to its beauty and efficient adjustment of text on paper, this writing style is extensively used to write the content of Arabic script languages including Balochi, Kashmiri, Pashto, Persian, Punjabi, Saraiki, Turkic and Urdu in Afghanistan, India, Iran, Pakistan and other south Asian countries. A special pen called QALAM is used to write text in Nastalique. Nastalique is a complex writing style as compared to Naskh. In Nastalique writing style, the QALAM is diagonally moved from top right to bottom left to write the characters in a ligature.

The diagonality of the writing style causes characters as well as ligatures to overlap as can be seen in Figure 5. In Nastalique, characters have contextual shapes based on the preceding and succeeding characters, and have complex diacritic placement rules. The detailed analysis of each character shape based on its position in a ligature is reported in [3] and [4]. The different shapes of a character are identified along with its joining rules with other characters. In Nastalique, the character shape inventory contains a total of 998 unique shapes of characters. This inventory is reduced to 488 unique shapes of character RASM classes. Another detailed character recognition based study of Nastalique character shapes reports 603 unique characters' contextual shapes which are reduced to 250 shapes of character's RASM classes [5]. The images of two letter ligatures showing the examples of initial shapes of SUAD and FAY [4], highlighted with black color, are illustrated in Figure 3.

The recognition of Urdu document images is a challenging task due to cursiveness, contextual character shaping and complex diacritics placement rules [3]. The focus of this paper is to develop a technique which extracts Nastalique artistic features of character and joiner sequence of a ligature. These artistic features of Nastalique play a very significant role for recognition of a character's core shape and improve the overall character recognition accuracy. The extracted features along with the labeled classes are classified using a sequence learning approach i.e. Hidden Markov Models(HMMs). To evaluate the technique, two datasets; (1) Dataset-1 containing 1,446 ligature classes of 3,309,762 ligature instances, and (2) Dataset-2 which is subset of UPTI dataset [6] having 1,600 text lines images are used. Dataset-2 is used to evaluate the performance of presented technique using artistic features and to compare it with state-of-the-art Urdu recognition techniques.

The rest of the paper is organized as follows. The current state-of-the-art techniques for character recognition are presented in Section 2. Section 3 discuses the complete methodology opted for the analysis and recognition of Urdu characters using character and joiner sequence. The dataset is explained in Section 4. The results and discussion are presented in Section 5. At the end Section 6 concludes the paper.

II. LITERATURE REVIEW

The state-of-the-art recognition techniques for Urdu like cursive script focus on the recognition of text using

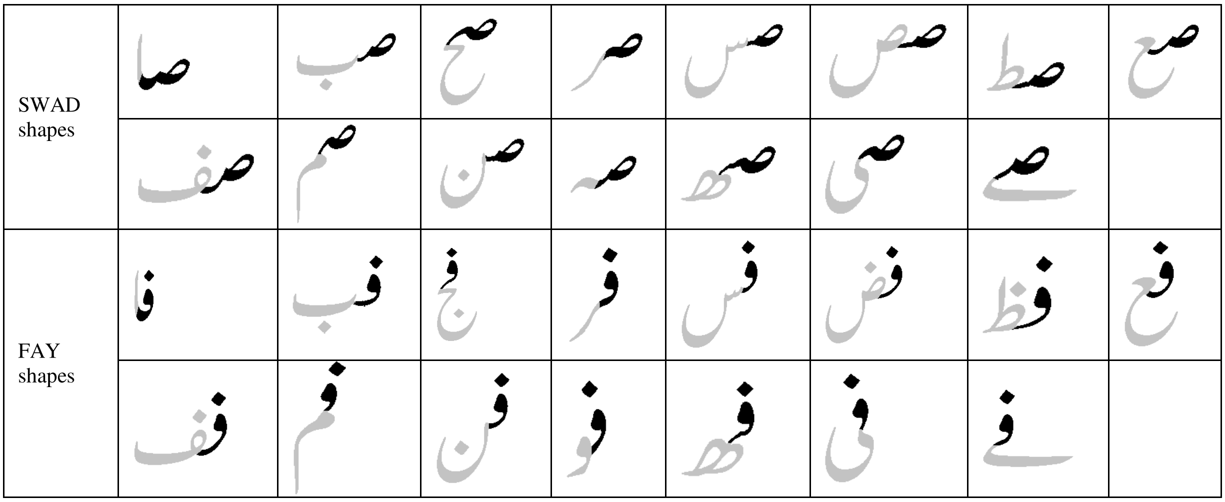


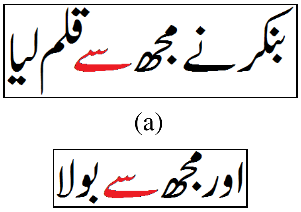
FIGURE 3. Examples of fifteen initial shapes of letter SWAD and FAY in two letter ligatures.

two approaches; (1) ligature based recognition and (2) character based recognition. The ligature based recognition techniques extract features from image. The features along with the sequence of ligatures transcription are used to train the system using different classification techniques [7], [8]. Sabbour and Shafait [6] report a dataset called UPTI having 10,000 synthetically generated Urdu text lines images written in Nastalique writing style. They develop ligature recognition system using shape context based features of contours of main body and diacritics. The reported accuracy of the system is 91% on UPTI test data. Akram *et al.* [9] also develop a ligature based recognition system by modifying state-of-the-art open source multilingual OCR engine. The system has 97.87% and 97.71% RASM class recognition accuracies tested on real and synthesized dataset of 14 and 16 font sizes respectively. Ahmad *et al.* [10] use a Gated Bidirectional Long-Short Term Memory (GBLSTM) networks to recognize the text line by recognizing the sequence of ligatures. The pixel values computed from ligature images along with ligature based labels are fed to the GBLSTM for training and recognition. The GBLSTM based ligature recognition system has 96.71% accuracy tested on 29,935 ligatures of 1600 text lines of UPTI dataset.

The character based recognition techniques extract characters' features from the images. The features along with the character transcription are used to train the classification system. There are two categories of character based recognition technique; (1) Explicit segmentation based recognition and (2) Implicit segmentation based recognition.

Explicit segmentation based recognition techniques segment the document image into characters by using image processing techniques and writing style characteristics. The features from the segments are extracted, and along with characters labeling are used to train the system. The study by Shaw *et al.* [11] use sliding window of size 1 × 27 horizontally on text line images of Devanagari script to detect and then remove the headline using a morphological operator. Headline removal results in the segmentation of a text line into characters. The HMMs are used to recognize 118 characters with 81.63% character recognition accuracy. The segmentation of Arabic text written in Naskh writing style is usually done by analyzing the characteristics of characters joining along the baseline. Lorigo and Govindaraju [12] devise a way to mark segmentation points by computing horizontal and vertical gradients on the baseline. These points are refined using character shape characteristics and some heuristics. The system has 92.3% segmentation accuracy tested on 200 Arabic handwritten images of IFN/ENIT. Another segmentation technique uses vertical projection profile, first derivative of the upper contour and distance between the baseline and pen tip, to mark the junction points which are later filtered by a trained Neural Network (NN) [13]. The structural and statistical features are used to train a Support Vector Machine (SVM) using character labels. The reported character recognition accuracy of the system is 98.3%. The explicit character based recognition techniques are also available for Urdu text written using Nastalique writing style [14], [15]. Due to Nastalique complexities of diagonality, cursiveness, contextual character shaping and overlapping, the segmentation of text into characters using image processing techniques is a challenging task. The thinned ligature images are segmented into character segments at the junction points. The discrete cosine transforms (DCTs) features are computed by sliding a window on the strokes of character segments. The features and characters' labels sequences are used to train an HMM. The recognized character segments are arranged in sequence to recognize the ligature class. The system is trained and tested on a dataset of 2,494 high-frequency ligatures' images which are synthetically generated at 36 font size. The reported RASM class recognition accuracy is 92.19%.

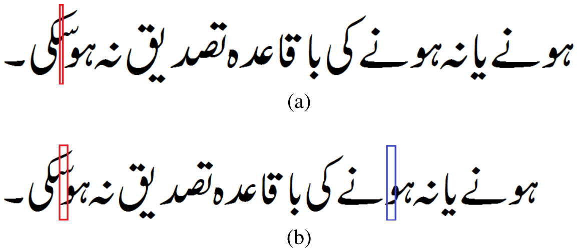
The implicit segmentation based recognition techniques use the concept of sequence learning using state-of-the-art sequence learning techniques. The text images with corresponding characters/ligatures/words labels are fed to sequence learning approaches including HMMs [5], [16]–[19], Recurrent Neural Networks (RNN) and variants of RNN [20], [21]. These sequence learning techniques are algorithmically strong enough to learn long context by extracting similar shape patterns and assigning labels to these patterns accordingly. The extensive research is available on use of HMMs and RNNs with variants to develop robust character recognition systems for different languages such as Arabic, Chinese, Devanagari and Urdu etc., by tweaking the feature extraction approach, devising efficient labeling and tweaking parameters of the learning system. The sliding window based contextual features and Bidirectional Long-Short Term Memory (BLSTM) networks are used to recognize the sequence of characters of Devanagari script [22] with 94.35% character recognition accuracy. The sixteen different pixels based features are extracted by sliding a window of size 80x3 on Arabic text line images. These features along with character level transcriptions are used to train HMM [23]. The dataset is synthetically generated by printing 2,766 text lines of 46,062 words using eight different font styles separately. The reported accuracy of the system is above 95% for text lines of each font. AlKhateeb *et al.* [24] use horizontal sliding window to extract 30 intensity features from Arabic handwritten text line images. These features along with sequence of labels are used for HMMs training. The system has 82.32% accuracy on standard IFN/ENIT dataset.



(a)

(b)

FIGURE 4. Same font size text having variation in line height. (a) Line height of 138 pixelsv, (b) Line height of 84 pixels.


ہونے یا نہ ہونے کی باقاعدہ تصدیق نہ ہو سکی۔
ہونے یا نہ ہونے کی باقاعدہ تصدیق نہ ہو سکی۔

(a)

(b)

FIGURE 5. Character and Ligature overlapping in horizontal sliding window. (a) Characters overlapping within ligature. (b) Characters overlapping of two ligatures, different feature sets of character و (WAO), highlighted with red and blue windows (overlapping characters strokes vs. actual character stroke).

The implicit segmentation based recognition techniques for Urdu characters are also available in the literature. An HMM based character sequence recognition system is presented by Hussain *et al.* [5] for the recognition of document images of Urdu books. The consistent character based RASM stroke traversal technique is presented to avoid confusing features computation of different overlapped characters which are captured in the same window. The DCTs features are extracted by a sliding window using the consistent traversal. Features along with sequence of characters' labels are used to train HMMs. The system is tested on 79,093 instance images of 5,249 RASM classes containing synthesized and real (extracted and cleaned from Urdu books) data. The reported accuracy of the system is 97.11%. Ul-Hasan *et al.* [25] extract pixels based features by sliding a window of size 30 × 1 on normalized line height image. The UPTI dataset having synthesized images of 10,063 Urdu text lines are used to train and test the system. The character transcription and extracted features are used to train a BLSTM network. The reported character recognition accuracy is 94.85% tested on 2003 text line images. The statistical features extracted from sliding windows on normalized line height images are used to recognize Urdu characters sequence using multi-dimensional long short-term memory (MLSTM) network [26]. The character recognition accuracy of the system is 94.97% tested on 1600 text lines images of UPTI dataset. The multi-dimensional long short term memory recurrent neural network (MDLSTM RNN) with connectionist temporal classification (CTC) as output layer gives 96.40% Urdu character recognition accuracy tested on 1600 text line images UPTI [27]. The character recognition accuracy is further improved by using a MDLSTM RNN with a matured output layer for sequence labeling giving 98% character recognition accuracy tested on UPTI dataset [28]. The hand crafted features are extracted using Convolutional Neural Networks (CNN) which are fed to MDLSTM for Urdu characters training and recognition. The system has 98.12% character recognition accuracy tested on 1,600 text lines images of UPTI dataset.

A sliding window on text line image requires line height normalization. The text lines containing different diagonally growing ligatures have different line heights, as can be seen in Figure 4. To normalize varying line height images to a standard line height e.g. 48 pixels as used in [26], different resizing factors are applied for height normalization. The sizes of the same ligatures in different line height images will be different resulting in different feature set of the same ligature, e.g. the ligature سے, highlighted with red color, will eventually have two different sizes. The feature set for characters س (SEEN) and ے (YEY-BARI) in both lines will be different.

The horizontal sliding window on Urdu Nastalique text line images requires intelligent feature extraction and character labeling techniques. Due to characters and ligatures overlapping in Nastalique writing style, as can be seen in Figure 5, the horizontal sliding window may contain overlapping strokes of characters and ligatures. These overlapping

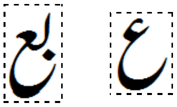


FIGURE 6. Different shapes of AIEN at final and isolated positions.

TABLE 1. Character and ligature trigram context.

	37 M Corpus	UPTI Corpus
Ligature trigram context	9,752,694	130,736
Character trigram context	43,191	9,852
Ligature RASM class trigram context	7,222,021	117,601
Character RASM class trigram context	3,095	1,121
Total ligatures classes	73,902	6,626
Total ligature RASM classes	34,786	3,883

characters' strokes in a window will results into the extraction of a confusing feature set of a character.

In addition, the character based transcription of Urdu Nastalique text line images requires an intelligent labeling technique which will generate different character labels based on different contextual shapes of a character. In Figure 6, AIEN character has different shapes at isolated and final positions. The same character labels of different shapes of a character as presented in [26]–[28] may generate confusions during recognition. However, it is the strength of the sequence learning approach which performs well for the recognition of Nastalique text lines having same labels of multiple contextual character shapes. The sequence learning approaches such as HMMs and RNN require significant dataset of text line images having contextual variations of each character and ligature to cover shape variations and overlapping of characters and ligatures. The characters' and ligatures' context coverage is extracted by processing 37 millions Urdu words corpus [29]. The publically available dataset i.e. UPTI is also processed to see ligatures and characters trigram coverage. Total number of ligatures and characters trigram contexts are given in Table 1. The character and ligature trigram context coverage requires a significantly larger dataset of Urdu text lines for sequence learning techniques. However, this context can be drastically reduced when ligature and character RASM classes (see Figure 2) will be considered. For such approach, the recognition process would be to recognize ligature/character RASM classes and diacritics sequence separately [30], and later the respective ligature string will be generated by using the recognized characters RASM sequence and diacritics.

III. METHODOLOGY

The character based recognition techniques for cursive script require consistent stroke traversal to avoid false feature set computation due to character overlapping and line height normalization. In addition, training dataset development covering all shape variations of a character is a tedious task. It becomes more challenging when trying to cover all the contextual shapes for sequence based learning approaches such as HMMs and deep learning based systems when applied on text line images. The motivation of this research study is to devise a technique which requires minimum training data and covers maximum context. Instead of characters and ligatures context, ligature RASM classes are used to devise an artistic features based character RASM sequence recognition technique. This drastically reduces the dataset size and gives maximum coverage. Later, the recognized diacritics will be used to generate the ligature string. The character based recognition technique extracts Nastalique calligraphy based artistic features which simplifies the challenge of multiple shapes of a character. To write the ligature in Arabic, first character's core shape is written, then the joiner is written (used to join character with next character) and then next character's core shape and joiner are written. The visualization of ligature as character and joiner pair sequence in Naskh and Nastalique is given in Figure 7. Hence instead of describing an image of a ligature L as sequence of characters i.e. $\boldsymbol{L = C_1\ C_2\ C_4\ \ldots C_n}$, the ligature image L is described as sequence of Character and Joiner pairs as $\boldsymbol{L = C_1\ J_1\ C_2\ J_2\ C_3\ J_3 \ldots C_n\ J_n}$. This formulation extracts and recognizes the character's core and the joiner shapes separately using contextual information. The character's core shape is the primary shape having distinctive shape features which are used to disambiguate it from different characters whereas the joiner gives joining information of characters. The detailed analysis of Nastalique calligraphy reveals that multiple shapes of a character are based on their contextual position in a ligature. If contextual shapes of a character are analyzed in terms of **Character (C)** and **Joiner (J)** then it can be better visualized that most of the characters have the same core shape in all the contexts whereas joiner's shapes vary and preserve the contextual joining information. The joiner does not give any information about the character's core shape but is used to give the positional information of a character. The shapes of the joiner of a character vary based on the next character in the ligature. A detailed analysis is carried out to extract the artistic features of Nastalique i.e. Character and Joiner shapes. The modeling of these artistic features is also presented which will be used to develop a recognition system of character sequence of a ligature image.

A. ARTISTIC FEATURES IDENTIFICATION

A ligature is composed of a paired sequence of characters and joiners. An investigation of character shapes in terms analyzing character core shapes and joiner shapes is carried out. This analysis on Nastalique text reveals that different contextual shapes of a character can be simplified by focusing on the core shape of a character. Hence, a character written in Nastalique can be artistically defined as a character having distinctive shape features, and a joiner giving information of contextual connection with next character which may have many contextual shapes. This reduces the complexity of the development of recognition system for Nastalique writing style. A detailed analysis of each ligature is carried out to extract character and joiner shapes in different contexts.

(a)Ligature طفیلے is written as ے+ل+ی+ف+ط in Nastalique Writing Style

(b) Ligature طفیلے is written as ے+ل+ی+ف+ط in Naskh Writing Style

Final Position Joiner ی	ی	Joiner ی with ل	ل	Joiner ل with ی	ی	Joiner ی with ف	ف	Joiner ف with ط	ط
$J_{\text{Yey\_Final}}$	C_{YEY}	$J_{\text{Laam\_Yey}}$	C_{LAAM}	$J_{\text{Yey\_Laam}}$	C_{YEY}	$J_{\text{Fay\_Yey}}$	C_{FAY}	$J_{\text{Toay\_Fay}}$	C_{TOAY}

(c) Transcription of ligature in terms of sequence of Character ($C_{\text{character}}$) highlighted black color, and Joiner ($J_{\text{Chaacter1\_character2}}$) highlighted with gray color

FIGURE 7. The ligature as character and joiner paired sequence in Naskh and Nastalique, characters shapes are highlighted with black color and joiner shapes are highlighted with gray color. (a) Ligature طفیلے is written as ے+ل+ی+ف+ط in Nastalique Writing Style. (b) Ligature طفیلے is written as ے+ل+ی+ف+ط in naskh writing style. (c) Transcription of ligature in terms of sequence of Character ($C_{\text{charector}}$) highlighted black color, and Joiner ($J_{Chaacters1\_character2}$) highlighted with gray color.

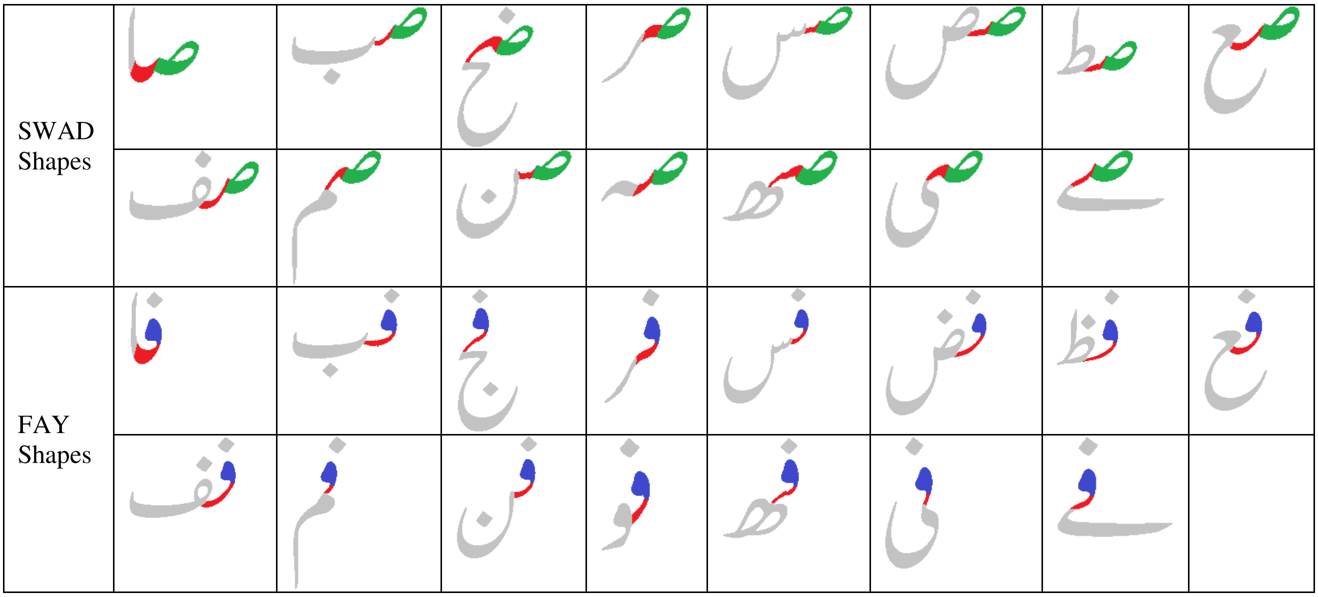


FIGURE 8. Character core and Joiners of initial shapes of SWAD and FAY, core highlighted with green and blue and Joiners highlighted with red color.

Based on its findings, the unique shapes of characters and joiners are extracted and transcribed. Some examples of Nastalique character and Joiner shapes are given in Figure 8, highlighting the unique shapes of SWAD and FAY in green and blue colors, and their joiners' shapes are highlighted with red color. A total of fifteen initial shapes of SWAD and FAY discussed in [4] are reduced to one SWAD and one FAY core shapes. The remaining contextual shaping information is maintained by joiners which is highlighted with red color. This analysis of Nastalique artistic features in terms of character and joiner sequence is carried out in an intelligent manner. First, single-letter ligatures are analyzed. The shapes of character and joiner are observed and sequence of labels is transcribed. Then the ligatures of length two are investigated efficiently in such a way that all two letters ligatures are classified separately which have specific analyzed letter as postfix characters of ligature. To further illustrate, ligatures e.g. با،سا،صا،جا are classified separately as already analyzed postfix letter ا (ALEF) and بب،سب،صب،جب ligatures are separately listed as already analyzed letter ب (BEH) as postfix. As the ***n-1*** length ligature contexts have been analyzed i.e. ا (ALEF) and ب (BEH), therefore only

TABLE 2. Examples of ligature transcription in terms of analyzed characters and joiners sequences.

Ligature	Transcription
ا	C_{ALEF}
ب	$J_{\text{Beh\_Final}}\ C_{\text{BEH1}}$
ص	$J_{\text{SWAD\_Final}}\ C_{\text{SWAD1}}$
ف	$J_{\text{Fay\_Final}}\ C_{\text{FAY}}$
با	$C_{\text{ALEF}}\ J_{\text{Beh\_Alef}}\ C_{\text{BEH2}}$
صا	$C_{\text{ALEF}}\ J_{\text{Swad\_Alef}}\ C_{\text{SWAD}}$
فا	$C_{\text{ALEF}}\ J_{\text{Fay\_Alef}}\ C_{\text{FAY}}$
صب	$J_{\text{Beh\_Final}}\ C_{\text{BEH1}}\ J_{\text{Swad\_Beh}}\ C_{\text{SWAD}}$
فب	$J_{\text{Beh\_Final}}\ C_{\text{BEH1}}\ J_{\text{Fay\_Beh}}\ C_{\text{FAY}}$
صج	$J_{\text{Jeem\_Final}}\ C_{\text{JEEM}}\ J_{\text{Swad\_Jeem}}\ C_{\text{SWAD}}$
صد	$C_{\text{DAL}}\ J_{\text{Swad\_Dal}}\ C_{\text{SWAD}}$
فد	$C_{\text{DAL}}\ J_{\text{Fay\_Dal}}\ C_{\text{FAY}}$
صر	$C_{\text{REH}}\ J_{\text{Swad\_Reh}}\ C_{\text{SWAD}}$
فر	$C_{\text{REH}}\ J_{\text{Fay\_Reh}}\ C_{\text{FAY}}$
صع	$J_{\text{Aien\_Final}}\ C_{\text{AIEN}}\ J_{\text{Swad\_Aien}}\ C_{\text{SWAD}}$
فع	$J_{\text{Aien\_Final}}\ C_{\text{AIEN}}\ J_{\text{Fay\_Aien}}\ C_{\text{FAY}}$

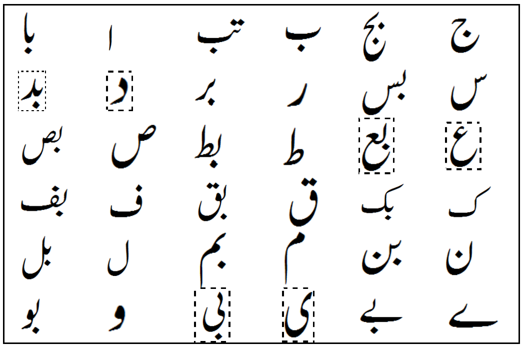


FIGURE 9. Final and isolated shapes of urdu letters.

remaining context i.e. first character and joiner shapes need to analyze. The analyzed context of ligatures of length ***n-1*** is used to analyze the context of ligatures of length ***n***. In the same way, the complete ligature RASM classes inventory is analyzed and the sequence of character and joiner labels' transcription is defined, some examples are given in Table 2.

The final and isolated shapes of character RASM are special cases of character core and joiner shapes. For such positional cases, the joiner shapes give information whether the character is at isolated or final position. The non-joiner characters i.e. ا (ALEF), د (DAL), ر (REH), و (WAO) and ے (YEY-BARI), do not join with any character therefore, these characters do not have joiner. Most of the characters have same joiner shape at isolated and final positions. The core character shape is also same at isolated and final positions, except some characters i.e. د (DAL), ع (AIEN) and ی(YEY) which have different character shapes, highlighted with a dotted border in Figure 9.

B. CHARACTER RECOGNITION USING ARTISTIC FEATURES MODELING

The motivation of this technique is to recognize segments i.e. the character and joiner from the input ligature image. The HMM based sequence learning technique is modeled to recognize the character and joiner segments sequence of the main body of a ligature. The HMMs based recognition formulation takes the extracted features as input and outputs the best sequence of segments in terms of character and joiner pairs of ligature images. The segment is generic term used to represent either character core shape or joiner. Therefore, a ligature can also be thought of as the sequence of character and joiner segments, mathematically S = $(\mathrm{s}_1\mathrm{s}_2\mathrm{s}_3 \ldots \mathrm{s}_\mathrm{m})$ segments. The features from the input ligature image are extracted i.e. $F = (f\,_1 f_2 f_3 \ldots f_n)$. The sequence of segments i.e. $S = (s_1 s_2 s_3 \ldots s_m)$ where s_i can be character or joiner which is recognized using the feature vector F. The recognized label of s_i determines whether it is character or joiner. The objective is to recognize the sequence of segments S using feature vector F, therefore Equation 1 will be

$$P\,(S \mid F) = P(s_1 s_2 s_3 \ldots s_m | f_1 f f_3 \ldots f_n) \tag{1}$$

By applying Bayes' rule, the Equation 1 is simplified as

$$P\,(S \mid F) = \frac{P\,(F \mid S)\,.P(S)}{P(F)} \tag{2}$$

The feature vector F is used to recognize the maximal probable sequence of segments using Equation 3 as

$$\hat{S} = \frac{P\,(F \mid S)\,.P(S)}{P(F)} \tag{3}$$

Equation 3 is further simplified by ignoring $P(F)$ as it is common in all possible sequences of segments, contributing equally in all comparisons to select the maximal probable sequence.

$$\hat{S} = P\,(F \mid S)\,.P(S) \tag{4}$$

The computation of $\hat{S}$ is simplified by taking the Markov assumption that the segment i.e. character or joiner is only dependent on its previous segment (bigram probability), i.e. $P\,(S)$ is simplified to $P\,(s_i \mid s_{i-1})$. $P\,(F \mid S)$ is also simplified by assuming that each feature f_i in the feature vector F is dependent only on its corresponding segment s_i i.e. $P\,(f_i \mid s_i)$. Therefore, Equation 4 is simplified as

$$\hat{S} = \prod\nolimits_1^n P\,(f_i \mid s_i)\,.P\,(s_i \mid s_{i-1}) \tag{5}$$

Finally top k segments' sequences $\varphi = \left\{\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4, \ldots, \hat{s}_k\right\}$ will be selected to generate the best character sequence from recognized segments sequence. The selected segments sequences will have probabilities $\left\{\mathrm{P}_{\hat{s}_1}, \mathrm{P}_{\hat{s}_2}, \mathrm{P}_{\hat{s}_3}, \mathrm{P}_{\hat{s}_4}, \ldots, \mathrm{P}_{\hat{s}_k}\right\}$ such that $\mathrm{P}_{\hat{s}_1} \geq \mathrm{P}_{\hat{s}_2} \geq \mathrm{P}_{\hat{s}_3} \geq, \mathrm{P}_{\hat{s}_4} \geq \ldots \geq \mathrm{P}_{\hat{s}_k}$. In this paper, the artistic features of character and joiners are extracted using a sliding window with consistent stroke traversal of characters and joiners in a sequence. These features along with the character and joiner labels are fed to HMMs for character based recognition. The complete feature extraction,

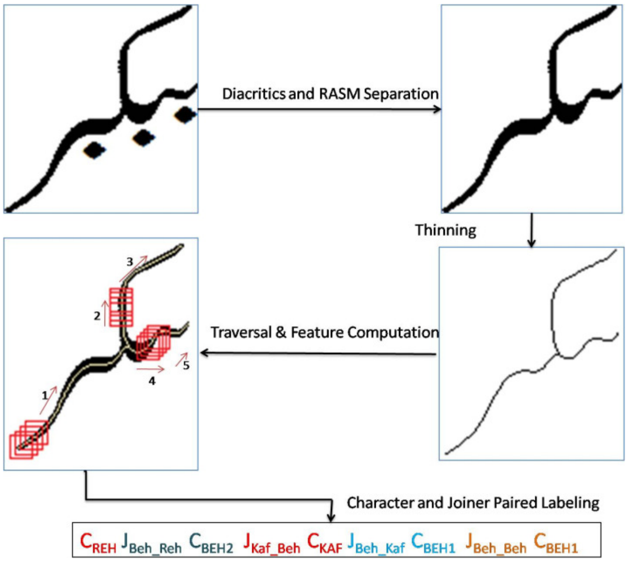

Diacritics and RASM Separation
Thinning
Traversal & Feature Computation
3
2
1
4
5
Character and Joiner Paired Labeling
C_{REH} J_{Beh_Reh} C_{BEH2} J_{Kaf_Beh} C_{KAF} J_{Beh_Kaf} C_{BEH1} J_{Beh_Beh} C_{BEH1}

FIGURE 10. **Process flow of feature extraction of nastalique ligature.**

and classification and recognition framework is discussed below.

C. FEATURE EXTRACTION

To develop the classification and recognition system of ligature RASM classes using Nastalique artistic features, the input ligature image is binarized [31] and diacritics and RASM (main body) are separated using dimensional features. The horizontal traversal of window is not applicable to compute the character and joiner features in a sequence. Therefore, the feature computation method discussed in [5] is used. The thinning algorithm is applied to find the thinned contour of the RASM which is used for traversal. A window of defined size is traversed on the thick RASM by moving the center point of the window on thinned contour. The traversal starts from the last character stroke and in sequence all the character strokes of a ligature are traversed in reverse order. Therefore sequence of labels (s_i) of character and joiner is also defined in reverse order. The feature set (f_i) of Equation 5 is computed for HMMs. The process flow of feature extraction from Nastalique ligatures for the classification of character and joiner pairs is given in Figure 10.

1) THINNING

A stroke based local window is used instead of a global window to avoid computation of confusing feature set due to overlapping of characters. This local window is traversed on the stroke of the RASM to capture the pen movement and to extract the features of the respective character and joiner. To move the window along the stroke of the characters, the thinning algorithm is applied to convert the thick stroke into the single pixel contour. To do this, first the RASM image is processed to remove salt and pepper noise which causes the thinning algorithm to generate an incorrect thinned contour of the RASM. After that, a thinning algorithm [32] is applied on the RASM image to generate the thinned contour. This thinned contour will be used to traverse the window along the characters' strokes using consistent traversal.

2) TRAVERSAL

The horizontal sliding window which is traditionally used to extract the features does not handle the character and ligature overlapping strokes appear in same window frame which causes extraction of confusing features against a single character see Figure 5. Therefore in this paper an intelligent stroke based sliding window technique is used which intelligently handles the erroneous features computation due to the overlapping of strokes of characters in a ligature. Using the thinned contour, a consistent traversal algorithm is applied to capture the Nastalique artistic calligraphic features of character and joiner paired sequence. The traversal algorithm traverses the ligature RASM in reverse order of its characters sequence starting from the last character [5]. Hence, the character and joiner labeling is also defined in reverse order so that HMMs correctly classify the character and joiner. The start point of the traversal is computed which is the last contour point of the last character. Thinned contours of most of RASM images have junction points. The priority rules are defined for each of the eight directions. These priority rules ensure the sequence of the strokes of a character is traversed in the same order as they are written using the pen movement. The thinned contour is overlaid on the thick stroke of the image, as can be seen in Figure 10. During traversal, the window is moved on the thick stroke by placing the center point of the window on the thinned contour. The complete RASM stroke is traversed and the corresponding portion of the original image captured by window is used to extract the feature vector. This traversal ensures the strokes traversal is following the same rules of moving the window as Urdu writer moves the pen to write Nastalique text. The next step is to compute the features.

3) COMPUTATION OF FEATURES

The feature set F of Equation 5 is used for classification of the characters and joiners using HMMs. Based on the literature review, the discrete cosine transform (DCT) as features perform well as compared to the structural and dimensional features [33]. The low frequency components of DCT extract meaningful information from the windowed image. The DCTs features are computed from $w \times w$ size window. To ensure that the center point of the window is on the contours of the RASM, the window size is selected as an odd number which is $w = 13$. This size is finalized by experiments to ensure complete information of the stroke is computed even at extreme thick portion when complete pen nip is used to write the segment of the stroke. The window is placed on the RASM stroke by placing the center of the window on the thinned contour overlaid on the RASM stroke. The DCT features of the window are computed and top-left three DCT coefficients are used as features f_i. The window is moved along the next contour point and DCT features from

$$C_{REH}\ J_{Beh\_Reh}\ C_{BEH3}\ J_{\ Kaf\_Beh}\ C_{KAF}\ C_{\ BEH1}\ J_{\ Beh\_Beh}\ C_{BEH1}$$
$$C_{REH}\ J_{Beh\_Reh}\ C_{BEH3}\ J_{\ Kaf\_Beh}\ C_{KAF}\ J_{Beh\_Kaf}\ C_{BEH1}\ C_{BEH1}$$
$$C_{REH}\ J_{Beh\_Reh}\ C_{BEH3}\ J_{\ Kaf\_Beh}\ C_{KAF}\ J_{Beh\_Kaf}\ C_{BEH1}\ J_{\ Beh\_Beh}\ C_{BEH1}$$
$$C_{REH}\ J_{Beh\_Reh}\ C_{BEH3}\ J_{\ Kaf\_Beh}\ C_{KAF}\ J_{Beh\_Kaf}\ C_{SEEN}$$

FIGURE 11. Ranked recognized segments sequence of characters and joiners of ligature RASM ببکبر.

the respective window placed on RASM stroke are computed and stored as f_{i+1}. The complete RASM stroke is traversed and the feature set $F = (f_1 f_2 f_3 \ldots f_n)$ is computed.

D. CLASSIFICATION AND RECOGNITION

The feature set $F = (f_1 f_2 f_3 \ldots f_n)$ and the sequence of segments labels i.e. $S = (s_1 s_2 s_3 \ldots s_m)$ are used to train the HMM model. The Sphinx toolkit [36] is used for training which generates a trained HMM model i.e. $\lambda = (A,B,\pi)$. In the HMM model, A,B and π correspond to transition probabilities, emission probabilities and initial state probabilities, respectively. During training, the HMM uses the Baum Welch algorithm to maximize the feature sequence probabilities $P(F|\lambda)$ incrementally.

After training, the next step is to recognize the sequence of segments from the input feature set of the RASM image. For this, the Viterbi-type beam search algorithm is used. This algorithm takes the feature set as input, uses the optimized HMM $\lambda = (A,B,\pi)$, and computes trigram based a ranked list of highly probable sequence of segments. A corpus of the character and joiner segments is generated so that HMMs can compute the probabilities $P(s_i \mid s_{i-1})$. This corpus is generated by computing the ligatures' probabilities from the text corpus [29]. The sequence of segments of characters and joiners in a ligature is repeated ***n*** times, where ***n*** is the frequency of the respective ligature in the text corpus. HMMs compute the prior probability i.e. $P(s_i \mid s_{i-1})$ of Equation 5 from this corpus. The weighted prior probability is combined with the likelihood model of the features sequence. The weight threshold of prior probability is set to 20 which is also computed experimentally. The feature set computed from the input ligature image i.e. F is fed to the trained HMM which recognizes and generates Top k Segments' Sequences $\varphi = \{\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4, \ldots, \hat{s}_k\}$. The recognized segments sequence which is the character and joiner labels sequence of the input image of ببکبر are given in Figure 11.

The recognized sequence of labels of character and joiner are further processed to recognize the respective character Unicode sequence using Equation 6.

$$CU = R\_CHAR\_Label\ R\_JOINER\_Label \tag{6}$$

where CU is Character Unicode, *R_CHAR_Label* is recognized character label and *R_JOINER_Label* is recognized joiner label. A lexicon is generated containing the character Unicode against respective character and joiner sequence so that formula given in Equation 6 can be used to recognize the character Unicode.

TABLE 3. Sample entries of lexicon having character Unicode against recognized segments of character and joiner.

Character and Joiner Labels	Character Unicode	Character and Joiner Labels	Character Unicode
C_{ALEF}	ا	$J_{Beh\_Reh}\ C_{BEH3}$	بـ
$J_{Beh\_Final}\ C_{BEH1}$	ب	$J_{\ Kaf\_Beh}\ C_{KAF}$	کـ
$J_{SWAD\_Final}\ C_{SWAD1}$	ص	$J_{Beh\_Kaf}\ C_{BEH1}$	بـ
$J_{Fay\_Final}\ C_{FAY}$	ف	$J_{\ Beh\_Beh}\ C_{BEH1}$	بـ
$J_{Beh\_Alef}\ C_{BEH2}$	ب	$J_{Swad\_Dal}\ C_{SWAD}$	ص
$J_{Swad\_Alef}\ C_{SWAD}$	ص	$J_{Fay\_Dal}\ C_{FAY}$	ف
$J_{Fay\_Alef}\ C_{FAY}$	ف	C_{REH}	ر
$J_{Swad\_Beh}\ C_{SWAD}$	ص	$J_{Swad\_Reh}\ C_{SWAD}$	ص
$J_{Fay\_Beh}\ C_{FAY}$	ف	$J_{Fay\_Reh}\ C_{FAY}$	ف
$J_{Jeem\_Final}\ C_{JEEM}$	ج	$J_{Aien\_Final}\ C_{AIEN}$	ع
$J_{Swad\_Jeem}\ C_{SWAD}$	ص	$J_{Swad\_Aien}\ C_{SWAD}$	ص
C_{DAL}	د	$J_{Fay\_Aien}\ C_{FAY}$	ف

The sample entries of lexicon are given in Table 3. Each pair of the character and joiner is searched in the lexicon and respective character Unicode is recognized. The sequence of recognized characters' Unicode is used to recognize the ligature RASM. The recognized ligature RASM using the recognized sequence of character and joiner are given in Table 4. For better understanding, the recognized character and joiner pairs which correctly matched in the lexicon are enclosed in the opening and closing curly braces, e.g. $\{J_{Beh\_Beh}\ C_{BEH1}\}$. The recognized character's Unicode is highlighted with underlined characters in Table 3.

The joiners have contextual shape variations. The analysis of recognition results also shows that joiners have confusions with other joiners causing misrecognition of character Unicode. Therefore another formula given in Equation 7 is also used to recognize the character Unicode sequence.

$$CU = R\_CHAR\_Label \tag{7}$$

In this approach, the character Unicode is recognized using only recognized labels of characters. The recognition output of HMM is filtered and all the joiners are ignored. The lexicon containing character Unicode against the recognized character label is used to generate characters Unicode sequence for the ligature RASM. The sample entries of the lexicon are given in Table 5. The recognition of character Unicode sequence using recognized character labels is given in Table 6.

The presented technique is integrated as RASM classification and recognition module with the Urdu Nastalique OCR framework [30], highlighted with blue color in Figure 12, to test the text line images.

The text lines of UPTI dataset are also used to test the performance of the presented technique. The images are binarized using algorithm defined in [31] to convert the gray scale image into binarized format. The RASM and diacritics are disambiguated using dimensional features. The diacritics association information with their respective RASM is also maintained. The RASM classes as sequence of character classes Unicode are recognized using the presented system.

TABLE 4. Recognition of Character Unicode sequence using recognized sequence of character and joiner information highlighted with gray.

HMMs Recognized Top K Sequence	Recognized Char Seq.
$\{C_{REH}\}\ \{J_{Beh\_Reh}\ C_{BEH3}\}\ \{J_{Kaf\_Beh}\ C_{KAF}\}\ \{J_{Beh\_Kaf}\ C_{BEH1}\}\ \{J_{Beh\_Beh}\ C_{BEH1}\}$	ببکبر
$\{C_{REH}\}\ \{J_{Beh\_Reh}\ C_{BEH3}\}\ \{J_{Kaf\_Beh}\ C_{KAF}\}\ C_{BEH1}\ \{J_{Beh\_Beh}\ C_{BEH1}\}$	بکبر
$\{C_{REH}\}\ \{J_{Beh\_Reh}\ C_{BEH3}\}\ \{J_{Kaf\_Beh}\ C_{KAF}\}\ \{J_{Beh\_Kaf}\ C_{BEH1}\}\ C_{BEH1}$	بکبر
$\{C_{REH}\}\ \{J_{Beh\_Reh}\ C_{BEH3}\}\ \{J_{Kaf\_Beh}\ C_{KAF}\}\ J_{Beh\_Kaf}\ C_{SEEN}$	کبر

TABLE 5. Sample entries of lexicon having character Unicode against recognized segments of character.

Character Labels	Character Unicode	Character Labels	Character Unicode
C_{ALEF}	ا	C_{BEH3}	ب
C_{BEH1}	ب	C_{KAF}	ک
C_{SWAD1}	ص	C_{REH}	ر
C_{FAY}	ف	C_{AIEN}	ع
C_{BEH2}	ب	C_{DAL}	د
C_{SWAD}	ص	C_{JEEM}	ج

TABLE 6. Recognition of Character Unicode sequence using recognized sequence of character information.

HMMs Recognized Top K Sequence	Recognized Char Seq
$C_{REH}\ C_{BEH3}\ C_{KAF}\ C_{BEH1}\ C_{BEH1}$	ببکبر
$C_{REH}\ C_{BEH3}\ C_{KAF}\ C_{BEH1}\ C_{BEH1}$	ببکبر
$C_{REH}\ C_{BEH3}\ C_{KAF}\ C_{BEH1}\ C_{BEH1}$	ببکبر
$C_{REH}\ C_{BEH3}\ C_{KAF}\ C_{SEEN}$	سکبر

The diacritics are recognized separately using different features based intensity and dimensional information. The text line images of UPTI dataset are tested using the integrated OCR framework system.

IV. DATA SET

Two different datasets are used to evaluate the presented technique. Dataset-1 containing cleaned images of ligature RASM is used for training and testing of the system. Dataset-2 is used to compare the performance of proposed system with the current state-of-the-art Urdu Nastalique recognition systems. To develop Dataset-1, 1,446 high frequent ligature classes are extracted by processing different text corpora. The selected ligature classes cover 3,309,762 ligature instances and 91,129 unique Urdu words. The majority of published Urdu content is written in Noori Nastalique, therefore 35 tokens of each of the selected ligature class are typed in Noori Nastalique writing style using InPage software. These are printed and scanned at 300 DPI to generate the document images. These images are processed to extract connected components. These are classified into ligature classes. In addition, real samples of the selected ligature classes are extracted from document images scanned from Urdu books [34]. The document images are binarized using algorithm defined in [31] to convert gray scale images into binarized format. The connected components of the image are extracted and classified into main body classes (RASM) using Urdu OCR [30] output and ground truth information. The instances of RASM classes are finalized after doing a manual pass. The Dataset-1 is divided into training and testing datasets. The training dataset has 30 instances of each selected RASM class by selecting 15 synthesized and 15 real instances. In the same way, a total of 15 instances are selected for the testing dataset with eight real instances and seven synthesized tokens. The synthesized instances are used when the number of real tokens are less than the required number of instances. Dataset-2 is a subset of UPTI dataset having 1600 randomly selected text lines. The available dataset is typed using Alvi Nastalique, which is used to type Jang Newspapers. These randomly selected text lines are also typed in the Noori Nastalique writing style, printed and scanned at 300 DPI. These lines have a total of 31,831 ligature instances. Dataset-2 is used to compare the performance of the proposed system with state-of-the-art Urdu recognition systems.

TABLE 7. Accuracy on Dataset-1 of RASM character sequence recognition using character (C) and Joiner(J) Vs. Character (C) only.

Total Ligatures	21,690
Total RASM Classes	1,446
Accuracy with Char and Joiner	74.19%
Accuracy with Char	95.58%

V. RESULTS AND DISCUSSION

Testing dataset of Dataset-1 containing 15 samples of each of 1,446 RASM classes is used to test the proposed system. Each image of the ligature RASM is thinned and traversed to compute DCTs as features. The features are fed to a trained HMMs and recognized top sequences of characters and joiners are generated. The lexicon is used to recognize the character Unicode against the sequence of character and joiner. Two approaches are used to recognize a character Unicode; (1) Character Unicode recognition using recognized character and joiner segments, see Equation 6, and (2) Character Unicode recognition using only recognized character segment, see Equation 7. The accuracy of both approaches are computed and given in Table 7. The recognition of a RASM class image is correct if the desired character class sequence of the RASM exists in the recognized ranked character sequences.

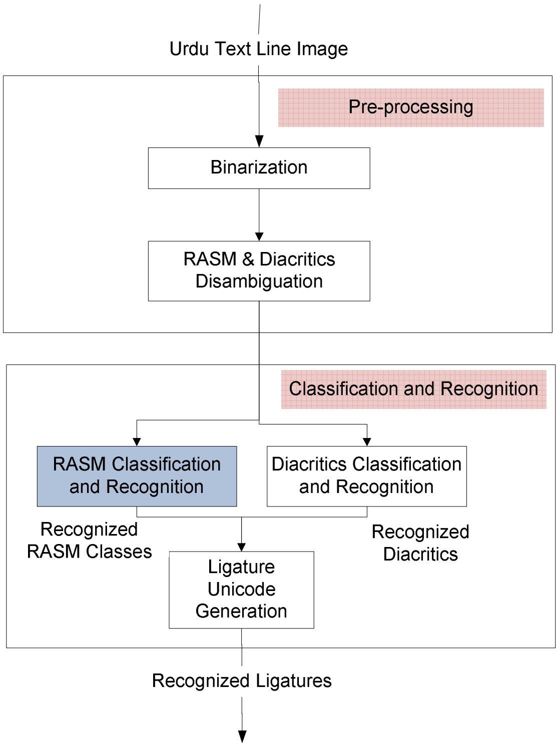
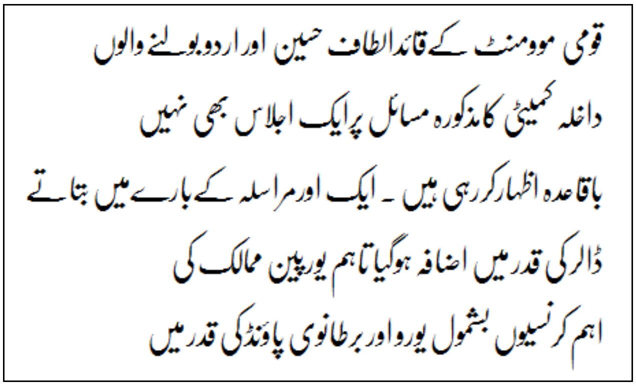
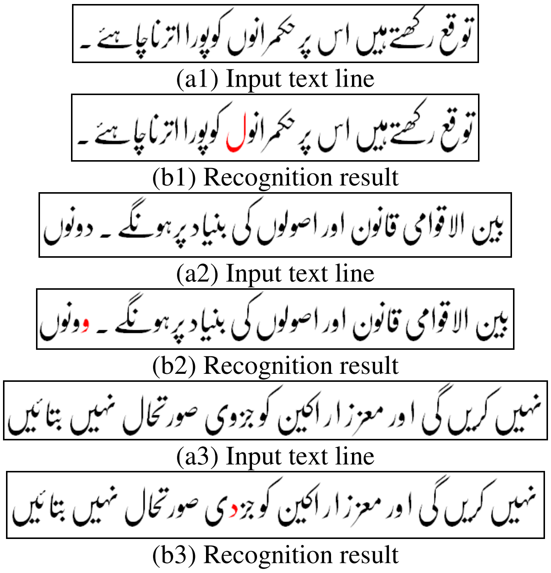

Urdu Text Line Image
Pre-processing
Binarization
RASM & Diacritics Disambiguation
Classification and Recognition
RASM Classification and Recognition
Diacritics Classification and Recognition
Recognized RASM Classes
Recognized Diacritics
Ligature Unicode Generation
Recognized Ligatures

FIGURE 12. Process flow of integrated Urdu OCR framework.



قومی موومنٹ کے قائدالطاف حسین اور اردو بولنے والوں
داخلہ کمیٹی کا مذکورہ مسائل پر ایک اجلاس بھی نہیں
باقاعدہ اظہار کررہی ہیں ۔ ایک اور مراسلہ کے بارے میں بتاتے
ڈالر کی قدر میں اضافہ ہوگیا تاہم یورپین ممالک کی
اہم کرنسیوں بشمول یورو اور برطانوی پاؤنڈ کی قدر میں

FIGURE 13. Examples of correct line recognition.



توقع رکھتے ہیں اس پر حکمرانوں کو پورا اترنا چاہیے ۔

(a1) Input text line

توقع رکھتے ہیں اس پر حکمرانول کو پورا اترنا چاہیے ۔

(b1) Recognition result

بین الاقوامی قانون اور اصولوں کی بنیاد پر ہونگے ۔ دونوں

(a2) Input text line

بین الاقوامی قانون اور اصولوں کی بنیاد پر ہونگے ۔ وونوں

(b2) Recognition result

نہیں کریں گی اور معززاراکین کو جزوی صورتحال نہیں بتائیں

(a3) Input text line

نہیں کریں گی اور معززاراکین کو جزدی صورتحال نہیں بتائیں

(b3) Recognition result

FIGURE 14. Input and recognized text lines represented by a1-a3 and b1-b3, respectively, Incorrectly recognized characters are highlighted with red color.

The character Unicode recognition using the recognized labels of character core shapes, drastically improve the results, giving 95.58% character recognition accuracy. The Nastalique character recognition is challenging due to complexity of shapes. This complexity is reduced as transcribing the character core shape and joiner shape separately. The character core shape is consistent and is clearly different from other characters. Whereas the joiner shapes preserve the joining information of a character with next character in a ligature and may confuse with the joiner shapes of other character, see Figure 8. Due to this strategy, HMMs learn the core shapes without any ambiguity and correctly recognize it. However, the confusing part i.e. joiner which have confusing shapes with joiners of other characters, have low recognition accuracies. The HMMs learns a ligature as sequence of characters and joiners separately rather it learns the ligature as sequence of characters. The detailed analysis shows that the majority of the misrecognition are caused due to misrecognition of joiners. According to Equation 6, the character Unicode is misrecognized, even if a character is correctly recognized and joiner is misrecognized by the HMMs. The main focus is to recognize the character Unicode if the character core shape is recognized correctly. Therefore by using Equation 7, the character Unicode is recognized by using recognized character core shape which drastically improves the recognition results.

The UPTI dataset is also used to do the comparative study. The proposed system of RASM recognition is integrated with the existed Urdu OCR system [30]. The text line image is processed and the connected components are disambiguated as diacritics and main bodies of the ligatures. The diacritics' association information of the ligature is also computed. The diacritics are recognized separately. The main bodies are recognized using the proposed recognition system. The character sequences are processed to form the ligatures by using the recognized diacritics and diacritics' association information. The recognition results of the ligature image is considered correct if the desired character Unicode sequences are found. The character recognition results of the proposed system along the comparisons with other techniques are given in Table 8. Recognition output of some lines are given in Figure 13. The majority of characters recognized incorrectly have confusions with other characters core shape such as character و (WAO) and character د (DAL) are confused with each other. In the same way, due to the variation in the image quality, the character ں (NOON) is sometimes confuses with character ل (LAAM). Some examples of text

TABLE 8. Comparison with state-of-the-art Urdu recognition techniques.

Systems	Segmentation	Features	No. of Lines	Classifier	Acc (%)
Ul-Hassan et al. [25]	Implicit	Pixels	2003	BLSTM	94.85
Naz et al. [26]	Implicit	Statistical	1600	MDLSTM	94.97
Naz et al. [27]	Implicit	Statistical	1600	MDLSTM	96.4
Sabbour and Shafait [6]	Holistic	Contour	----	BLSTM	91
Naz et al. [35]	Implicit	Convolutional	1600	MDLSTM	98.12
Ahmad et al. [10]	Holistic	Pixels	1600	GBLSTM	96.71
Proposed	Implicit	Nastalique Artistic Features	1600	HMMs	**98.37**

lines having incorrectly recognized characters are given in Figure 14, the incorrect characters are highlighted with red color.

VI. CONCLUSION

In this paper, an implicit Urdu character recognition technique is presented for the recognition of sequence of characters and joiners. In Arabic script, each character image has a core shape which distinguishes it from the other characters, and joiner is used to join the character with next characters having multiple shapes based on the context. The artistic features of Nastalique calligraphy are analyzed to identify the shapes of characters and joiners. By using the robust stroke based traversal of Urdu ligature, the features are extracted which are used to train and recognize the characters and joiners sequence using HMMs. To see the significance of the recognition of characters Unicode using recognized labels of the core character shape, two approaches are used. In the first approach character Unicode is recognized using the recognized labels of character and joiner, giving 74.19% character recognition accuracy. In second approach, the character Unicode is recognized using the recognized labels of character core shapes only, giving 95.58% on Dataset-1. As second approach drastically improves the recognition results, therefore this approach is used to test the performance on UPTI dataset which gives 98.37% character recognition accuracy. The results show that the presented system outperforms the state-of-the-art Urdu character recognition systems. This recognition model can be easily applied for recognition of text of other Arabic script languages such as Arabic, Balochi, Pashto, Punjabi, Saraiki and Sindhi. In addition, recognition of handwritten text can be significantly improved by using the presented technique.

REFERENCES

[1] Wikipedia. (2014). *Arabic Diacritics*. Accessed: Mar. 30, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Arabic_diacritics

[2] P. Hamed. *Famous Calligraphers–Persian Calligraphy–All About Persian Calligraphy*. Accessed: Mar. 30, 2018. [Online]. Available: http://www.persiancalligraphy.org/Famous-Calligraphers.html

[3] A. Wali and S. Hussain, "Context sensitive shape-substitution in Nastaliq writing system: Analysis and formulation," in *Proc. Int. Joint Conf. Comput., Inf., Syst. Sci., Eng. (CISSE)*, 2006, pp. 53–58.

[4] S. Hussain, S. Rahman, A. Wali, A. Gulzar, and S. J. Rahman, "Grammatical analysis of Nastalique writing style of Urdu," Center Res. Urdu Lang. Process., FAST-nu, Lahore, Pakistan, 2002.

[5] S. Hussain, S. S. Ali, and Q.-U.-A. Akram, "Nastalique segmentation-based approach for Urdu OCR," *Document Anal. Recognit.*, vol. 18, pp. 357–374, Dec. 2015.

[6] N. Sabbour and F. Shafait, "A segmentation-free approach to arabic and urdu OCR," *Proc. SPIE, Document Recognit. Retr. XX. Int. Soc. Opt. Photon.*, vol. 8658, pp. 86, 580N-86, and 580N-12, 2013.

[7] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, "Segmentation free Nastalique Urdu OCR," *World Acad. Sci., Eng. Technol.*, vol. 46, pp. 456–461, Oct. 2010.

[8] N. H. Khan, A. Adnan, and S. Basar, "Urdu ligature recognition using multi-level agglomerative hierarchical clustering," *Cluster Comput.*, 2017, doi: 10.1007/s10586-017-0916-2.

[9] Q.-U.-A. Akram, S. Hussain, A. Niazi, U. Anjum, and F. Irfan, "Adapting tesseract for complex scripts: An example for Urdu Nastalique," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2014, pp. 191–195.

[10] I. Ahmad, X. Wang, Y. H. Mao, G. Liu, H. Ahmad, and R. Ullah, "Ligature based Urdu Nastaleeq sentence recognition using gated bidirectional long short term memory," *Cluster Comput.*, 2017, doi: 10.1007/s10586-017-0990-5.

[11] B. Shaw, S. K. Parui, and M. Shridhar, "Offline handwritten Devanagari word recognition: A segmentation based approach," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[12] L. Lorigo and V. Govindaraju, "Segmentation and pre-recognition of Arabic handwriting," in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Aug./Sep. 2005, pp. 605–609.

[13] R. Mehran, H. Pirsiavash, and F. Razzazi, "A front-end OCR for Omni-font Persian/Arabic cursive printed documents," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2005, p. 56.

[14] S. T. Javed and S. Hussain, "Segmentation based Urdu Nastalique OCR," in *Proc. 18th Iberoamerican Congr. Pattern Recognit. (CIARP)*, Havana, Cuba, 2013, pp. 41–49.

[15] A. Muaz, "Urdu optical character recognition system," M.S. thesis, Nat. Univ. Comput. Emerg. Sci., Islamabad, Pakistan, 2010.

[16] S. Abirami, V. Essakiammal, and R. Baskaran, "Statistical features based character recognition for offline handwritten Tamil document images using HMM," *Int. J. Comput. Vis. Robot.*, vol. 5, pp. 422–440, Jan. 2015.

[17] J. V. Monaco and C. C. Tappert, "The partially observable hidden Markov model and its application to keystroke dynamics," *Pattern Recognit.*, vol. 76, pp. 449–462, Apr. 2018.

[18] P. P. Roy, A. K. Bhunia, A. Das, P. Dey, and U. Pal, "HMM-based Indic handwritten word recognition using zone segmentation," *Pattern Recognit.*, vol. 60, pp. 1057–1075, Dec. 2016.

[19] O. Samanta, U. Bhattacharya, and S. Parui, "Smoothing of HMM parameters for efficient recognition of online handwriting," *Pattern Recognit.*, vol. 47, no. 11, pp. 3614–3629, Nov. 2014.

[20] A. Graves, "Supervised sequence labelling," in *Supervised Sequence Labelling With Recurrent Neural Network*. Berlin, Germany: Springer, 2012, pp. 5–13.

[21] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[22] N. Sankaran and C. V. Jawahar, "Recognition of printed Devanagari text using BLSTM neural network," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 322–325.

[23] H. A. Al-Muhtaseb, S. A. Mahmoud, and R. S. Qahwaji, "Recognition of off-line printed Arabic text using hidden Markov models," *Signal Process.*, vol. 88, pp. 2902–2912, Dec. 2008.

[24] J. H. AlKhateeb, J. Ren, J. Jiang, and H. Al-Muhtaseb, "Offline handwritten Arabic cursive text recognition using hidden Markov models and re-ranking," *Pattern Recognit. Lett.*, vol. 32, pp. 1081–1088, Jun. 2011.

[25] 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 *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2013, pp. 1061–1065.

[26] S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi, and M. I. Razzak, "Urdu Nasta'liq text recognition system based on multi-dimensional recurrent neural network and statistical features," *Neural Comput. Appl.*, vol. 28, pp. 219–231, Feb. 2017.

[27] S. Naz *et al.*, "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks," *Neurocomputing*, vol. 177, pp. 228–241, Feb. 2016.

[28] S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, and F. Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks," *SpringerPlus*, vol. 5, no. 1, p. 2010, 2016.

[29] F. Adeeba, Q. Akram, H. Khalid, and S. Hussain, "CLE Urdu Books N-gram," in *Proc. Conf. Lang. Technol.*, Karachi, Pakistan, 2014, pp. 87–92.

[30] Q. Akram, S. Hussain, F. Adeeba, S. Rehman, and M. Saeed, "Framework of Urdu Nastalique optical character recognition system," in *Proc. n Lang. Technol. (CLT)*, Karachi, Pakistan, 2014, pp. 23–30.

[31] M. Naz, Q.-U.-A. Akram, and S. Hussain, "Binarization and its evaluation for Urdu Nastalique document images," in *Proc. 16th Int. Multi Topic Conf. (INMIC)*, Lahore, Pakistan, Dec. 2013, pp. 213–218.

[32] H. Lei, W. Genxun, and L. Changping, "An improved parallel thinning algorithm," in *Proc. 7th Int. Conf. Document Anal. Recognit.*, Aug. 2003, pp. 780–783.

[33] G. Lehal and A. Rana, "Recognition of Nastalique Urdu ligatures," in *Proc. 4th Int. Workshop Multilingual OCR*, Washington, DC, USA, 2013, pp. 1–5.

[34] Q.-U.-A. Akram, A. Niazi, F. Adeeba, S. Urooj, S. Hussain, and S. Shams, "A comprehensive image dataset of Urdu Nastalique document images," in *Proc. Conf. Lang. Technol. (CLT)*, Lahore, Pakistan, 2016, pp. 81–88.

[35] S. Naz *et al.*, "Urdu Nastaliq recognition using convolutional–recursive deep learning," *Neurocomputing*, vol. 243, pp. 80–87, Jun. 2017.

QURAT UL AIN AKRAM is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, UET Lahore, Lahore, Pakistan. She is also an Assistant Manager Research with the Center for Language Engineering (CLE), KICS, UET Lahore, where she leads the document image processing and pattern recognition research domain. She has managed the Urdu Nastalique Optical Character Recognition Project at CLE. Her areas of interest are natural language processing, image processing, and computer vision.

SARMAD HUSSAIN is currently a Professor of computer science and heads the Center for Language Engineering, KICS, UET Lahore. His research is focused on developing computing solutions for Pakistani languages, including research in character recognition, linguistics, localization, language computing standards, speech processing, and computational linguistics. He has been serving on many national and international committees. Some of his current international memberships include the Security and Stability Advisor Committee of ICANN, the Executive Committee of Asian Federation of Natural Language Processing, and the Pakistan representative on the International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques. At national level, he currently serves on IDN ccTLD Committee of the Ministry of IT, National Standards Committee of National Language Authority, and is the Chairperson of Society of Natural Language Processing of Pakistan.