Recognition of Urdu Ligatures - A Holistic Approach

Israr Uddin Khattak*, Imran Siddiqi*, Shehzad Khalid* and Chawki Djeddi†
* Bahria University, Islamabad, Pakistan
Emails: israr.uddin@yahoo.com, imran.siddiqi@bahria.edu.pk, shehzad@bahria.edu.pk
†LAMIS Laboratory, Larbi Tebessi University, Tebessa, Algeria.
Email:c.djeddi@mail.univ-tebessa.dz

***Abstract*—This paper presents an effective segmentation-free and scale-invariant technique for recognition of Urdu ligatures in Nastaliq font. The proposed technique relies on separating the main body of ligatures from the secondary components and training a separate hidden Markov model for each. Features capturing projection, concavity and curvature information of ligatures are extracted using right-to-left sliding windows and are fed to the models for training. The system trained and evaluated on a total of more than 2,000 frequently occurring Urdu ligatures from a standard database realized a recognition rate of 97.93%.**

I. Introduction

Optical Character Recognition (OCR) is one of the most classical pattern recognition problems that has received more than three decades of extensive research attention. Today, commercial OCR systems reporting near to 100% recognition rates are available for texts in different scripts like Roman and Chinese etc. Research on Arabic OCR has also matured over the recent years reporting acceptably good recognition rates. Despite these tremendous developments, OCRs for many languages around the globe are either non-existent or are witnessing early days of research, Urdu being one of them and makes the subject of our study. Research on Urdu and similar cursive scripts like Pashto, Farsi etc. is still in its infancy with limited literature available till date.

Urdu is the national language of Pakistan and a popular language of the Indian sub-continent that is spoken and written by millions of people around the globe. The alphabet of Urdu is a superset of the Arabic alphabet but as opposed to Arabic which follows the Naskh writing style, Urdu employs the Nastaliq style. Though both Naskh and Nastaliq are written right to left, Naskh is written horizontally while Nastaliq is written diagonally from top right to bottom left making it highly cursive and more sensitive to context eventually leading to a more challenging recognition problem [1]. An example illustrating the Naskh and Nastaliq writing styles can be seen in Figure 1.

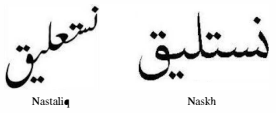

Fig. 1. Nastaliq and Naskh writing styles

Traditionally, OCRs for cursive text are categorized into segmentation-based and segmentation-free methods. Segmentation-based approaches segment the text into characters which are then recognized. Segmentation-free approaches, on the other hand, employ complete words or ligatures as units of recognition. The main advantage of segmentation-based approaches [2], [3], [4], [5] is that the number of distinct classes to be recognized is the same as the number of letters in the alphabet and their various (context dependent) shapes. This number is much smaller as compared to the total number of words or ligatures in the vocabulary which are units of recognition in segmentation-free approaches. The segmentation of cursive scripts like Nastaliq into characters, however, is a challenging task itself hence most of the approaches reported for recognition of Urdu text rely on segmentation-free techniques [6], [7], [8], [9], [10], [11], [12]. Figure 2 shows an example Urdu word and its segmentation into ligatures and subsequently into characters.

A comprehensive survey on segmentation-based and segmentation-free Urdu OCR systems can be found in [19]. An analysis of the existing recognition methods for Urdu reveals that most of the proposed techniques either work on isolated characters or consider only a limited number of ligatures. One of the most comprehensive studies has been presented in [16] where the authors adapt Tesseract for recognition of 1,475 unique (primary) ligatures of Urdu achieving 97.87% and 97.71% accuracies on font sizes of 14 and 16 respectively. The main drawback of this technique is its dependency on font size and a fresh training is to be carried out for every font size.

Fig. 2. (a) A example Urdu word and its segmentation into (b) ligatures and (c) characters

This paper presents a segmentation-free and font size invariant approach for recognition of Urdu ligatures. Recognition is carried out using Hidden Markov Models (HMM) where a separate HMM is trained for each ligature. The main body ligatures and dots/diacritics are separately recognized and are later associated. The system trained on more than 2,000 high frequency ligatures realized an overall recognition rate of 97.93%. The details of the proposed recognition methodology are presented in the next section.

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

II. Proposed Methodology

As discussed earlier, the proposed technique relies on a segmentation-free approach that works on ligatures. Among well-known segmentation-free approaches proposed in the literature, ligatures have been the most popular unit of recognition considered in a number of studies [13], [14], [12], [15]. There are more than 26,000 unique ligatures in Urdu and a study on the occurrence frequencies of these ligatures indicate that most of these ligatures occur very rarely. More than 99% of entire Urdu corpus can be constituted using only 10% (around 2600) of all Urdu ligatures [13]. It is therefore an attractive choice to consider high frequency ligatures (HFL) only which results in a manageable number of classes realizing acceptable recognition rates for commonly occurring Urdu text. Ligatures are further distinguished into primary components and secondary components. The primary components represent the main body of the ligature which may have zero or more secondary components which include dots and diacritics.

Like any pattern classification problem, the proposed ligature recognition methodology is divided into two main phases, training and recognition, as discussed in the following.

A. Training

Training involves making the model(s) learn to discriminate between different ligature classes. We have chosen to employ Hidden Markov Models (HMMs) which have been successfully applied to a wide variety of recognition problems [20]. In our study, we have worked on the high frequency ligature (HFL) database developed by the Center of Language Engineering (CLE), Pakistan (www.cle.org.pk). The database contains 2,017 frequently occurring Urdu ligatures while the number of characters in each ligature varies from 1 to 8. The main body of each ligature is separated from secondary components by extracting the largest connected component from each ligature (Figure 3). In addition to the main body of each ligature, we also consider 11 frequently occurring secondary components (Figure 4) in our study making a total of 2,028 unique components. A separate HMM was trained using 30 instances of each ligature to cater for noise and intra-ligature variations caused by the image acquisition process.

Fig. 3. Decomposition of a ligature into primary and secondary components

Features are extracted by normalizing the height of each ligature image to a predefined size of 60 pixels and sliding a window (frame) of 60×7 pixels with an overlap of 4 pixels as illustrated in Figure 5-a. For each frame, we extract a combination of projection, concavity and curvature features which have been effectively applied to a number of problems including handwriting recognition [17], word spotting [18] and writer recognition [23]. The computational details of these features are discussed in the following.

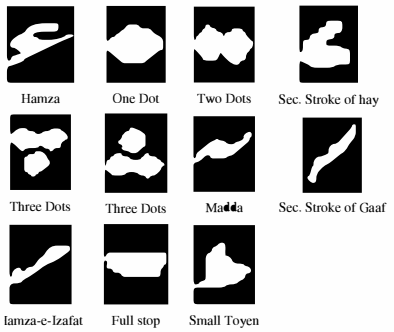

Fig. 4. Secondary components considered in the study

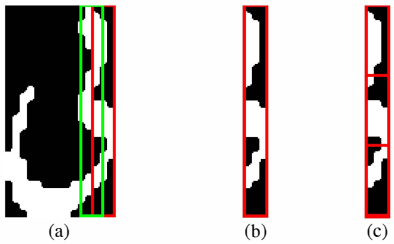

Fig. 5. Sliding window for extraction of features (a) Overlapped frames (b) A single frame (c) Sub-frames in a frame

- ***Projection Features:*** Within each frame, we compute the horizontal and vertical projection of the text pixels. Projections are computed by summing the total number of text pixels in each row (column) of the frame and normalizing by the width (height) of the frame. In our case, for a frame size of 60×7 we get a 60-dimensional horizontal and a 7-dimensional vertical projection.
- ***Concavity Features:*** Concavity features originally proposed in [22], [21] for Arabic handwriting recognition and later extended by Azeem & Ahmed [17], aim to capture the local concavity and direction information of strokes. These features are computed by using eight masks of size 3×3 each. These masks compute the horizontal edges in the upper and lower contours, vertical edges in the left and right contours and upper and lower edges in the left and right diagonal contours of the (ligature) image as illustrated in Figure 6. The number of 'on' pixels in the 8 edge images is counted in a 8-bin histogram which is used as feature. While projection features are computed from the entire frame, the concavity features are computed by first dividing the frame into 3 equal parts (Figure 5-c) and counting the edge pixels in each of the sub-frames (20×7). This gives a total of 24 (8×3) values per frame.
- ***Curvature Features:*** While concavity captures the orientation information in the contours of a ligature, the curvature information is captured by computing a histogram of the differential chain codes also known as

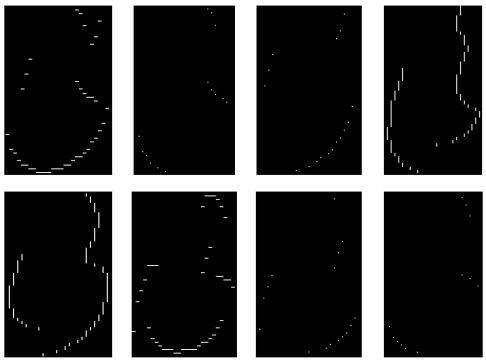

Fig. 6. Edge images obtained using the concavity features

the curvature density function. The ligature contour is first reprsented by a sequence of Freeman chain codes and the differential chain codes are computed by subtracting each value in the sequence from the next [24]. The differential chain code at pixel p_i corresponds to the angle θ_i between the vectors $p_{i-1}p_i$ and p_ip_{i+1} as indicated in Figure 7 and the histogram of these codes is computed within each frame generating a 7 dimensional feature vector.

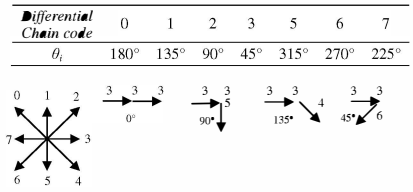

Fig. 7. Differential chain codes and the corresponding angles

Once the features are computed, each frame is represented by a single feature vector comprising vertical projection (7 values/frame), horizontal projection (60 values/frame), edge counts (24 values/frame) and the curvature distribution (7 values/frame). This combination results in a 98 dimensional feature vector for each frame. Features extracted by sliding the frames in a right-to-left fashion on each ligature image are used to train the HMMs. Since the HMMs are discrete, the feature space is quantized to a 75 symbol codebook and a 15 state right-to-left HMM (Figure 8) is trained using 30 instances of each ligature considered in our study. The training is carried out using the standard Baum-Welch algorithm. Once the models are trained, the Unicode of each ligature is associated with its respective model.

B. Recognition

During recognition, the query ligatures presented to the system are first split into primary and secondary components in a similar fashion as in the training phase. Features extracted from the query ligature are fed to each of the trained models and the model that reports the maximum probability is picked, the respective Unicode being the output. As discussed earlier, the main body and the secondary components of each ligature are recognized separately and are later associated. The association of secondary components with the primary

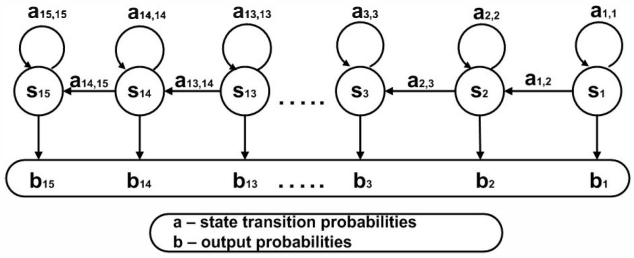

Fig. 8. Structure of HMMs employed in our study

component, however, is a complete problem in itself and its discussion is beyond the scope of this paper.

In the next section, we describe the experimental protocol and the recognition rates achieved in different evaluation scenarios.

III. Experimental Results and Analysis

As mentioned earlier, the system was trained to recognize 2,017 high frequency primary ligatures and 11 secondary components making a total of 2,028 components. For evaluation, 3 instances of each ligature were presented to the system constituting 6,084 ($3 \times 2,028$) query ligatures. The test set was generated by extracting ligatures from scanned Urdu books as well as from the CLE database itself (different images from those used in the training). The system reported an overall recognition rate of 97.93%. In addition to the overall recognition rate, we also separately computed the recognition rates as a function of the number of characters in the ligature. These results are summarized in Table I. It can be seen from these results that in general, the recognition rates are more or less consistent across ligatures with different number of characters. A relatively low recognition rate of 94.81% is realized on ligatures with only one character. This may be attributed to the fact that single character ligatures generally have a relatively smaller width resulting in fewer frames per ligature image leading to a reduced recognition rate. Overall, the high recognition rates achieved for different ligature classes validate the effectiveness of the proposed scheme.

TABLE I. Recognition rates as a function of number of characters per ligature

Characters/ Ligature	Query Ligatures	Recognized Ligatures	Recog. Rate
1	135	128	94.81%
2	408	394	96.57%
3	1,713	1,680	98.07%
4	2,280	2,230	97.80%
5	1,158	1,142	98.62%
6	333	327	98.20%
7	45	45	100%
8	12	12	100%
Total	**6,084**	**5,958**	**97.93%**

In addition to the results presented above, we also carried out a study on the sensitivity of the recognition rate to different parameters of the system. These experiments were conducted by using the first 500 high frequency ligatures in

the CLE database by varying the number of states and the window size used for framing. The corresponding recognition rates are illustrated in Figure 9. It can be seen from these results that the recognition rate increases with the increase in the number of states in the HMM and begins to stabilize from 13 states onwards. These rates are relatively more sensitive to the frame width where smaller frame sizes produce higher recognition rates. Smaller frame widths yield a larger number of windows per ligature and the extracted features better characterize the ligatures. In all evaluations, for a frame width of 'n' pixels there is an overlap of $(n+1)/2$ pixels.

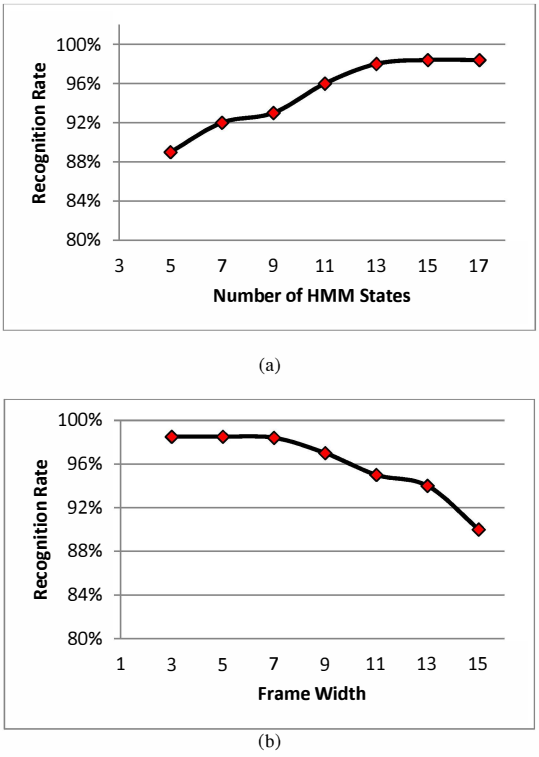

Fig. 9. Recognition rates on 500 ligatures as a function of (a) Number of states in the HMM (b) Frame width

We also carried out a comparison of the proposed recognition scheme with some notable contributions on Urdu ligature recognition. The recognition rates of these studies are summarized in Table II. A meaningful comparison can be made with the work of Javed et al. [12] and Akram et al. [16] as both the studies have reported their results on the same dataset as the one used in our evaluations. The recognition rate realized by the proposed method is comparable to the best recognition rate of 97.87% reported in [16]. However, this recognition rate is achieved on 1,475 ligatures and the technique works on fixed font sizes only. The proposed system reports a better recognition rate on 2,028 ligatures and is font size invariant. It should, however, be noted that we separately recognize the primary and secondary ligatures and the reported recognition rates are prior to association of secondary components with their respective primary components. The recognition rate of our system is likely to drop once the diacritics are associated with the primary ligatures. Nevertheless, a recognition rate of around 98% on more than 2000 classes is indeed promising.

IV. Conclusion

This study proposed a holistic approach for recognition of frequently occurring Urdu ligatures. The proposed methodology employs HMM based recognizers where a separate model is trained for each ligature considered in our study. The models are trained by extracting a set of features using overlapped sliding windows from multiple instances of each ligature. A total of 2,028 high frequency ligatures from the standard CLE database covering a major proportion of Urdu vocabulary are considered in our study. Experimental evaluations using 3 query instances of each ligature realized a high recognition rate of around 98%.

The presented approach does not cover the association of diacritics with the primary ligatures which is an important aspect from the view point of practical recognition engines and will form the subject of our subsequent communication. Presently, we have considered only the 11 most frequent secondary components. We intend to cater the entire set of diacritics in our further work on this subject. In addition, the system is presented with query ligatures for recognition. It would be a good idea to extend the system to be able to recognize complete words and subsequently complete pages of Urdu text. This in turn would require development of effective segmentation techniques and will be the focus of our future study.

References

[1] Aamir Wali and Sarmad Hussain, *Context sensitive shape-substitution in nastaliq writing system: Analysis and formulation*, Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, pages 53-58, 2007.

[2] Hamna Malik, Fahim and Muhammad Abuzar, *Segmentation of printed urdu scripts using structural features*, Second International Conference in Visualization, pages 191-195, 2009.

[3] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan, *Urdu nastaleeq optical character recognition*, In Proceedings of world academy of science, engineering and technology, pages 249-252, 2007.

[4] Zaheer Ahmed, Jehanzeb Khan Orakzai, and Inam Shamsher, *Urdu compound character recognition using feed forward neural networks*, In Proceedings of 2nd IEEE International Conference on Computer Science and Information Technology, pages 457-462, 2009.

[5] U. Pal and Anirban Sarkar, *Recognition of printed urdu script*, In Proceedings of 12th International Conference on Document Analysis and Recognition, volume 2, pages 1183-187, 2003.

[6] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, and Ching Y. Suen, *Holistic urdu handwritten word recognition using support vector machine*, In Proceedings of 20th International Conference on Pattern Recognition, pages 1900-1903, 2010.

[7] Sohail A. Sattar, Shamsul Haque, and Mahmood K. Pathan, *Nastaliq optical character recognition*,In Proceedings of the 46th Annual Southeast Regional Conference, pages 329-331, 2008.

[8] Shuwair Sardar and Abdul Wahab, *Optical character recognition system for urdu*,In Proceedings of International Conference on Information and Emerging Technologies, pages 1-5, 2010.

[9] Misbah Akram and Sarmad Hussain, *Word segmentation for urdu ocr system*,In Proceedings of the 8th Workshop on Asian Language Resouces, pages 88-94, 2010.

[10] N. Sabbour and F. Shafait, *A Segmentation Free Approach to Arabic and Urdu OCR*, In SPIE, Volume 8658, 2013.

TABLE II. NOTABLE STUDIES ON URDU OCR WITH RECOGNITION RATES

Study	Dataset	Recognition Rate	Remarks
Z. Ahmad et al. [3]	Synthetic and real-world images of Urdu	93.4%	Ignores diacritics
T. Nawaz et al. [11]	Isolated characters in different font sizes	89%	Tested only on isolated characters
N. Sabbour & F. Shafait [25]	UPTI Online Arabic e-book	Urdu:99% Arabic: 86%	Ignores diacritics
Javed et al. [12]	1,282 ligatures	92%	Fixed font size
Akram et al. [16]	1,475 ligatures	97.87%	Fixed font size
Proposed Method	2,028 ligatures	97.93%	Font size invariant

[11] Tabassam Nawaz, Syed Ammar Hassan Shah Naqvi, Habib ur Rehman, and Anoshia Faiz, *Optical character recognition system for urdu (naskh font) using pattern matching technique*, International Journal of Image Processing, pages 92-104, 2009.

[12] Sobia Javed, Sarmad Hussain, Ameera Maqbool, Samia Asloob, Sehrish Jamil, and Huma Moin, *Segmentation free nastalique urdu ocr*, In Proceedings of World Academy of Science, Engineering and Technology, volume 46, pages 456-461, 2010.

[13] Gurpreet Singh Lehal, *Choice of Recognizable Units for Urdu OCR*, In Proc. of the workshop on Document Analysis and Recognition Pages 79-85, 2012.

[14] Sobia Tariq Javed and Sarmad Hussain, *Improving Nastaliq Specific Pre-recognition Process for Urdu OCR*, In Proc. of the 13th Intl Multitopic Conference, 2009.

[15] Gurpreet Singh Lehal, *Ligature Segmentation for Urdu OCR*, In Proc. of the 12th International Conference on Document Analysis and Recognition (ICDAR), pages 1130-1134, 2013.

[16] Qurat-ul-Ain Akram, Sarmad Hussain, Aneeta Niazi, Umair Anjum and Faheem Irfan, *Adapting Tesseract for Complex Scripts- an Example for Nastalique*, In Proc. of the 11th IAPR International Workshop on Document Analysis Systems, 2014.

[17] Sherif Abdel Azeem and Hany Ahmed, *Effective technique for the recognition of offline Arabic handwritten words using hidden Markov models*, International Journal on Document Analysis and Recognition (IJDAR), Volume 16, pages 399-412, 2013.

[18] Khurram Khurshid, Claudie Faure, and Nicole Vincent, *Word spotting in historical printed documents using shape and sequence comparisons*, Pattern Recognition, 45(7), pages 2598-2609, 2012.

[19] Saeeda Naz, Khizar Hayat, Imran Razzak, Waqas Anwar, Sajjad Madani, and Samee Khan, *The optical character recognition of Urdu-like cursive scripts*, Pattern Recognition, 47(3), pages 1229-1248, 2014.

[20] Plotz, T. and Fink, G. A, *Markov models for offline handwriting recognition: a survey*, International Journal on Document Analysis and Recognition (IJDAR), 12(4), pages 269-298, 2009.

[21] Al-Hajj, R., Likforman-Sulem, L., Mokbel, C., *Combining slantedframe classifiers for improved HMM-based arabic handwriting recognition*, IEEE Trans. Pattern Anal. Mach. Intell. 31(7), pages 11651177, 2009.

[22] Al-Hajj, R., Likforman-Sulem, L., Mokbel, C., *Combination of HMM-based classifiers for the recognition of arabic handwritten words*, In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, 2007.

[23] Imran Siddiqi and Nicole Vincent, *Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features*, Pattern Recognition, 43(11), pages 3853-3865, 2010.

[24] Imran Siddiqi and Nicole Vincent, *A set of chain code based features for writer recognition*, In: Proc. of 10th Intl. Conference on Document Analysis and Recognition, pages 981-985, 2009.

[25] Nazly Sabbour and Faisal Shafait, *A segmentation-free approach to arabic and urdu ocr*, In IS&T/SPIE Electronic Imaging, 2013.