

Segmentation Based Urdu Nastalique OCR

Sobia Tariq Javed¹, Sarmad Hussain²

¹National University of Computer and Emerging Sciences, Lahore, Pakistan

sobia.tariq@nu.edu.pk

²Al-Khawarizmi Institute of Computer Science, University of Engineering and Technology, Lahore, Pakistan.

sarmad.hussain@kics.edu.pk

Abstract. Urdu Language is written in Nastalique writing style, which is highly cursive, context sensitive and is difficult to process as only the last character in its ligature resides on the baseline. This paper focuses on the development of OCR using Hidden Markov Model and rule based post-processor. The recognizer gets the main body (without diacritics) as input and recognizes the corresponding ligature. Accuracy of the system is 92.73% for printed and then scanned document images at 36 font size.

Keywords: Nastalique, Urdu OCR, Urdu Segmentation

1. Introduction

Urdu is written using Arabic script in Nastalique writing style. Urdu has an extended Arabic character set as given in the figure below [13, 14, and 15]. Urdu characters are constituted by a main body with zero or more diacritics for specifying the consonants and (optionally) vowels. Nastalique writing style is very cursive with context sensitive shaping [9, 10]. The characters join together to form a *Ligature*. One or more ligatures form a word. For example word Pakistan shown in the Figure 1 (b) has three ligatures. Moreover, Urdu is bidirectional [3].

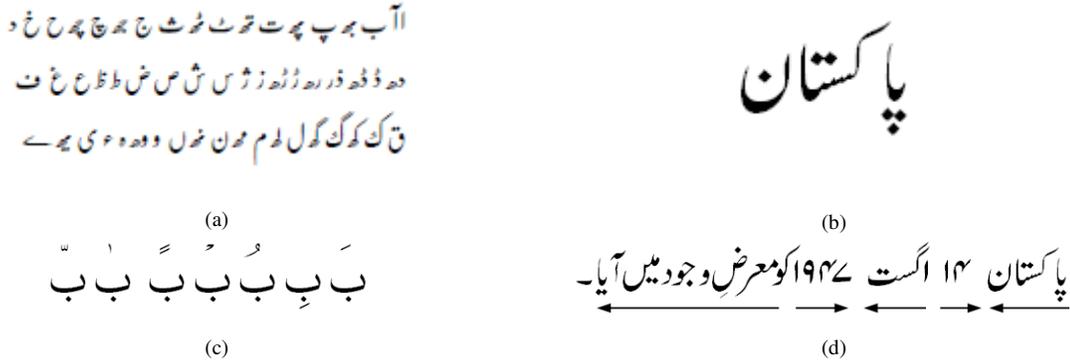


Fig. 1. Different characteristics of Urdu Writing System (a) Urdu character Set [3] (b) Word *Pakistan* written in Nastalique (c) Diacritical marks on letter bay (d) Bidirectional Urdu script

The character set of Urdu shown in Figure 1 (a) can be sub-categorized into classes which contain same base forms (if the dots and marks are ignored). This categorization is given in Table 1, and is the basis of segmentation and recognition. Some letters belong to multiple classes because they contain isolated and final forms different from initial and medial forms. For example, letter *Fay* has same initial and medial forms as *Qaf* (e.g. *فب* *بفب* compared with *قب* *بقب*) but different isolated and final forms (e.g. *ف* *بف* compared with *ق* *بق*). These letters are listed in the last row, in addition to being listed with other classes. In the rest of the paper we refer to classes and not the individual characters.

Table 1 : Classification of Urdu letters based on their shapes

Member Letter(s)	Class	Member Letter(s)	Class	Member Letter(s)	Class
م	م	ص	ص	ا آ	ا
و	و	ط ظ	ط	ب پ ت ٹ ث ن	ب
ہ	ہ	ع غ	ع	ج چ ح خ	ج
ھ	ھ	ف ق	ف	د ڈ ذ	د
ی	ی	ک گ	ک	ر ر ژ ز	ر
ے	ے	ل	ل	س ش	س
ق	ق	ف	ف	ن ن	ن

2. Methodology

The current system uses HMMs for pattern matching, as it can accurately handle large data sets and can be trained to handle noise and distortion to some extent [5, 6, 12, 16, 17, 18]. The recognition process is shown in the Figure 2.

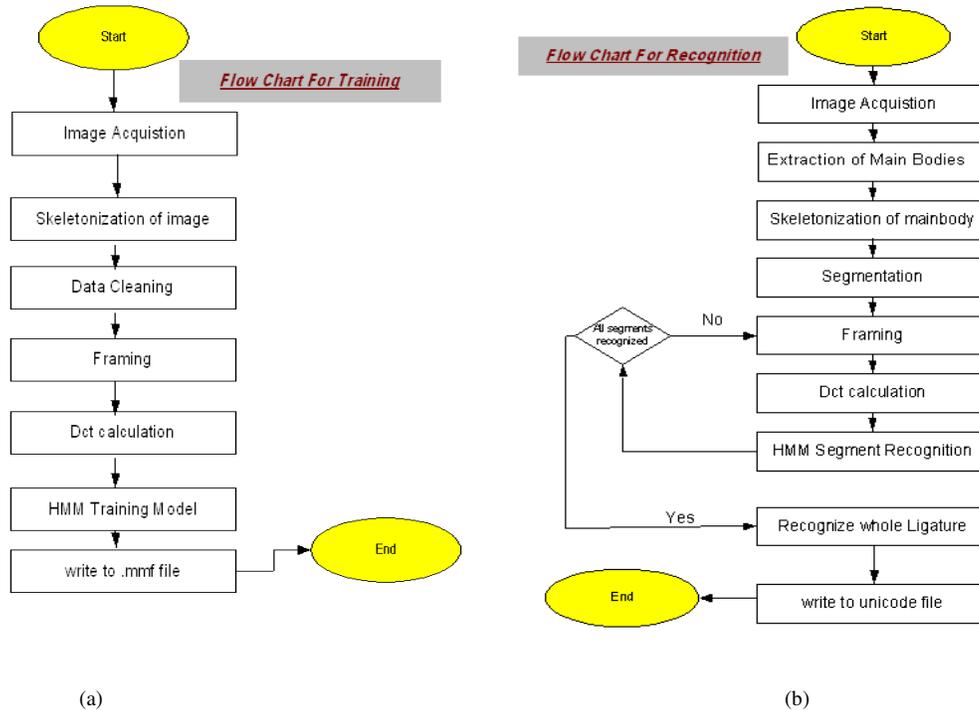


Fig. 2. Flow Charts for (a) Training and (b) Recognition

The system takes a monochrome scanned image with 150 dpi containing Urdu text as input. As the document images are generated by the authors in the lab to test the process, there is no pre-processing and it is assumed that the document images are not skewed and with minimum noise and distortion. Main bodies are extracted by first

separating lines of text within the page, then identifying the baseline and finally separating the main bodies from diacritics using the baseline [2, 3]. After extracting the main body, they are skeletonized using Jang Chin algorithm [4], as shown in Figure 3.

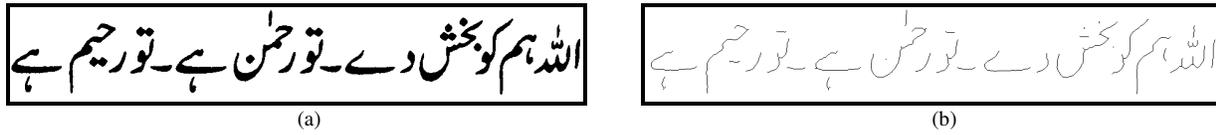


Fig. 3. (a) Original Image and (b) Sketetonized Image

The skeletonized image is then segmented after determining the ending point of the ligature. In Nastalique it is very difficult to determine the exact starting point of the ligature so instead of that we start with the ending point of the ligature [1] which is more deterministic, as shown in Figure 4.

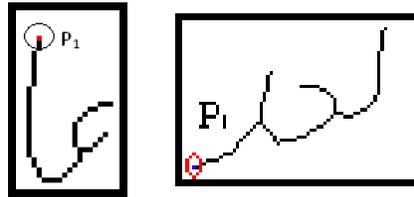


Fig. 4. The Ending Points P1 of the Ligatures are Circled

Therefore, as Urdu is written from right to left, the ligatures are traversed from left to right. During the traversal the ligatures are segmented at branching points as shown in Figure 5 below. This process results in multiple segments from a ligature, obtained in the reverse writing order as shown in Table 2.

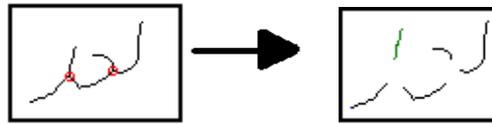
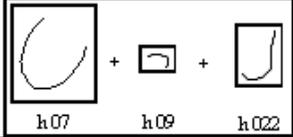
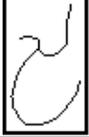
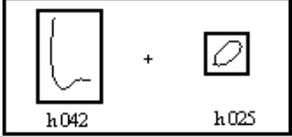
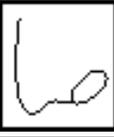
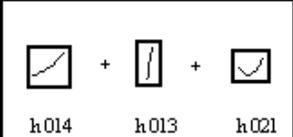
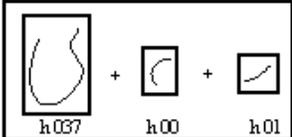
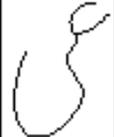


Fig. 5. Segmentation of a ligature using branching points [1]

Table 2 : Segmentation of the Ligatures

Sr.	Segments	Ligature	Sr.	Segments	Ligature
1	 h07 h09 h022		3	 h042 h025	
2	 h014 h013 h021		4	 h037 h00 h01	

These skeletonized ligature segments are framed and used to train the HMMs, as shown in Figure 6. The system is tested with non-overlapping frame sizes of 5x5, 8x8, 9x9, 12x12 and 16x16 pixels. 8x8 is found to give the best results for the 36 font size.

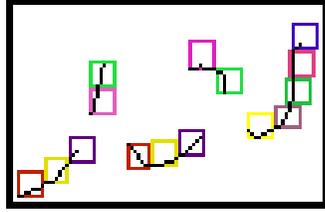


Fig. 6. The Framing of Segmented Word باء (baad)

When this pre-processing is complete, before starting training process, the HMM parameters are initialized with training data in order to allow convergence of the training algorithm. Each segment is considered as a separate HMM. Sixty HMMs were extracted from all shapes (isolated, initial, final and medial) of a sub-set of six classes of Urdu characters, including *Alif*, *Bay*, *Dal*, *Swad*, *Ain* and *Yeh* classes which are given in the Figure 7 (a). In order to cater variations in the image 100 samples of each shape are collected for training the HMMs. Samples of isolated *Bay* are given below in Figure 7 (b).

After training the model, the recognition process is performed. In recognition process the skeletonized ligature is first segmented and then each segment is fed to the HMM for recognition.

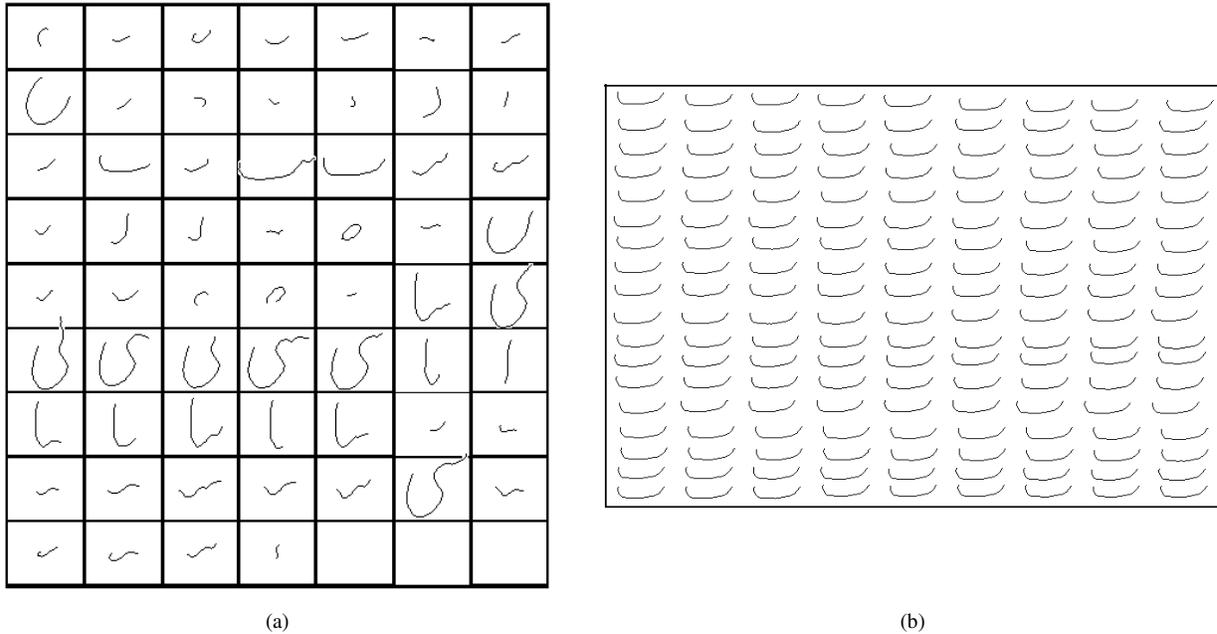


Fig. 7. (a) Segments modeled by HMMs and (b) Variation in training samples used for the HMMs

As the segments are of varied sizes, and the framing window size is fixed, for more precisely modeling each segment different number of states are defined for the different segments, with some examples illustrated in Table 3.

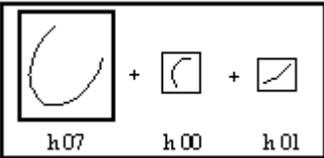
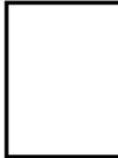
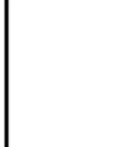
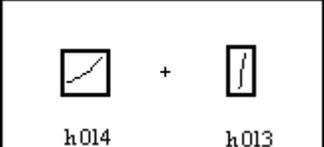
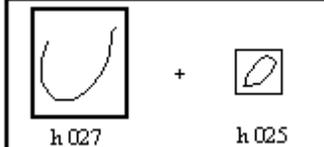
Table 3 : HMM State Analysis

Sr.	HMM Name	HMM	No. of Frames	No. of states	No. of samples
1	h00	ع	3	5	100
2	h01	ب	2	4	100
3	h02	و	3	5	109
4	h03	ا	3	5	136

5	h04		3	5	100
6	h05		3	4	110
7	h06		3	4	122
8	h07		10	11	108
9	h08		2	4	137
10	h09		2	4	114

After training, the model is developed, the recognition process is performed. In the recognition process the skeletonized ligature is first segmented and then each segment is sent to the HMM for recognition. Once the constituent segments are recognized, rules are applied to order them to form the corresponding ligature, as shown in Table 4.

Table 4 : Rules for Forming Ligatures from Constituent Segments

Sr.	Segments	Letter	Sr.	Segments	Letter
1	 h07 h00 h01	 Isolate d Ain class	4	 h09	 Medial Ain class
2	 h023	 Initial bay class	5	 h041	 Medial Alif class
3	 h014 h013	 Final Daal class	6	 h027 h025	 Isolate d Swad class

3. Results

A total of 1692 ligatures, which are formed from the six base forms mentioned above, are extracted from the 18600 high frequency words in a corpus-based dictionary [7]. These classes are used in these ligatures in a variety of contexts. The Urdu words were written in font Noori Nastalique and font size 36. The pages are printed and then scanned at 150 DPI. Out of these 1692 ligatures, 1569 ligatures were identified correctly giving an accuracy of 92.73%.

4. Discussion

The focus of the paper has been to explore segmentation based system capable of recognizing Urdu Nastalique font. The results are promising; however, some letters are not recognized correctly due to the following problems. The variation in the images affects the output of the recognizer. The variation may be introduced due to scanning, binarization and thinning processes, giving wrong recognition results, as shown in Figure 8.

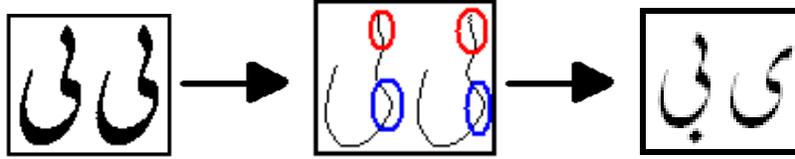


Fig. 8. Variation in the images causing ligature misrecognition

This problem may be resolved by increasing the number of training samples and by giving original segment as HMM input instead of giving skeletonized segment.

The similarity in the shapes of different characters can also lead to the recognizer confusion. For example, the shape of the letter *Bay* and last stroke in *Swad* (in Figure 9) are similar to each other when written in Noori Nastalique. Diacritics can disambiguate such cases.



Fig. 9. Similarity in Shapes *Swad* and *Bay* in Different Contexts

Inconsistency in font can also cause some variation causing recognition errors. The Noori Nastalique font used shows such behavior in some cases, e.g. main bodies of ligatures change with change in diacritics as shown in Figure 10. This variation is because the font uses hand written ligatures.



Fig. 10. Dissimilarity in Shapes of Same Ligature with Different Diacritic Placement

5. Conclusion

The Nastalique style used to write Urdu language is complex due to its diagonal, context sensitive and cursive nature. In this paper we have presented a technique to develop a segmentation based OCR for Nastalique. A ligature is first segmented and each segment is recognized using an HMM based recognizer. Then a set of rules are used to identify the ligature corresponding to the sequence of recognized segments. The accuracy of system is 92.73% for six base forms using fabricated documented images at 36 font size. The technique still needs to be tested on real data and extended to cover the entire set of Urdu letters at a variety of font sizes.

Reference

- [1]. Javed, S.T., Hussain, S.: Investigation into a Segmentation Based OCR for the Nastalique Writing System. Master's thesis report at National University of Computer and Emerging Sciences, Lahore (2007). Available at (<http://www.cle.org.pk/resources/theses.htm>)
- [2]. Javed, S.T., Hussain, S., Maqbool, A., Asloob, S., Jamil, S., Moin, H.: Segmentation Free Nastalique Urdu OCR. In: International Conference on Computer, Electrical, and Systems Science, and Engineering, Paris (2010).
- [3]. Javed, S.T., Hussain, S.: Improving Nastalique Specific Pre-Recognition Process for Urdu OCR. In: the Proceedings of 13th IEEE International Multitopic Conference 2009 (INMIC 2009), Islamabad, Pakistan (2009).
- [4]. Jang, B-K. and Chin, R.T.: Analysis of thinning algorithms using mathematical morphology. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (1990).
- [5]. Rabiner, L., Juang, B-H.: Theory and Implementation of Hidden Markov Models. In the book, "Fundamental of Speech Recognition", chapter 6 (1993).

- [6]. Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (1995).
- [7]. Ijaz, M., Hussain, S.: Corpus Based Urdu Lexicon Development. In the Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan (2007).
- [8]. Pal, U., Sarkar, A.: Recognition of Printed Urdu Text. In the Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR) (2003).
- [9]. Hussain, S.: www.LICT4D.asia/Fonts/Nafees_Nastalique. In the Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore (2003).
- [10]. Wali, A., Hussain, S.: Context Sensitive Shape-Substitution in Nastaliq Writing system: Analysis and Formulation. In the Proceedings of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE) (2006).
- [11]. Lu, Z., Bazzi, I., Kornai, A., Makhoul, J.: A Robust, Language-Independent OCR System. In the 27th AIPR Workshop: Advances in Computer Assisted Recognition, SPIE (1999).
- [12]. Bojovic, M., Savic, M. D.: Training of Hidden Markov Models for Cursive Handwritten Word Recognition. In the Proceedings of the 15th International Conference on Pattern Recognition (ICPR) vol.1, (2000).
- [13]. Hussain, S., Afzal, M.: Urdu Computing Standards: UZT 1.01. In the Proceedings of the IEEE International Multi-Topic Conference, Lahore, Pakistan (2001).
- [14]. Hussain, S.: Letter to Sound Rules for Urdu Text to Speech System. In the Proceedings of Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Switzerland (2004).
- [15]. Hussain, S., Durrani, N.: Urdu. In A Study on Collation of Languages from Developing Asia, Center for Research in Urdu Language Processing, NUCES, Pakistan (2007).
- [16]. El-Hajj, R., Likforman-Sulem, L., Mokbel, C.: Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling. In the 8th International Conference on Document Analysis and Recognition (ICDAR), South Korea (2005).
- [17]. Elms, A.J.: A Connected Character Recognizer Using Level Building of HMMs. In the Proceedings of 12th International Conference on Pattern Recognition (1994).
- [18]. Safabakhsh, R., Abidi, P.: Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM. In the Arabian Journal for Science and Engineering, (2005).
- [19]. Shah, Z., Saleem, F.: Ligature Based Optical Character Recognition of Urdu, Nastaliq Font. In the Proceedings of International Multi Topic Conference, Karachi, Pakistan (2002).
- [20]. Husain, S.A., Amin, S.H.: A Multi-tier Holistic approach for Urdu Nastaliq Recognition. In the Proceedings of International Multi Topic Conference, Karachi, Pakistan (2002).
- [21]. Ahmad, Z., Orakzai, J. K., Shamsheer, I., Adnan, A.: Urdu Nastalique Optical Character Recognition. In the Proceedings of World Academy of Science, Engineering and Technology (2007).
- [22]. Shamsheer, I., Ahmad, Z., Orakzai, J. K., Adnan, A.: OCR for Printed Urdu Script Using Feed Forward Neural Network. In the Proceedings of World Academy of Science, Engineering and Technology (2007).
- [23]. Malik, S., Khan, S.A.: Urdu online handwriting recognition. In Proceedings of the IEEE Symposium on Emerging Technologies (2005).