Choice of Recognizable Units for Urdu OCR

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University, Patiala
+91-9815473761
gslehal@gmail.com

ABSTRACT

There has been considerable work on Arabic OCR. However, all that work is based on Naskh style. Urdu script is based on Arabic alphabet, but uses Nastalique style. The Nastalique style makes OCR in general and character segmentation in particular, a highly challenging task, so most of the researchers avoid the character segmentation phase and go in for higher unit of recognition. For Urdu, the next higher recognition unit considered by researchers is ligature, which lies between character and word. A ligature is a connected component of one or more characters and usually an Urdu word is composed of 1 to 8 ligatures. A related issue is identification of all possible ligatures for recognition purpose. For this purpose, we have performed a statistical analysis of Urdu corpus to collect and organise the Urdu ligatures. The number of unique ligatures comes to be more than 26,000, and recognition of such a huge class is again a Herculean task. It becomes necessary to reduce the class count and look for alternative recognition unit. From OCR point of view, a ligature can further be segmented into one primary connected component and zero or more secondary connected components. The primary component represents the basic shape of the ligature, while the secondary connected component corresponds to the dots and diacritics marks and special symbols associated with the ligature. To reduce the class count, the ligatures with similar primary components are clubbed together. Further statistical analysis is performed to count and arrange in descending order the primary components and a manageable class of around 2300 recognition units has been generated, which covers 99% of Urdu corpus.

Keywords

OCR, Urdu script, segmentation, connected components.

1. INTRODUCTION

Character recognition is one of the most important fields of pattern recognition has been around since the development of first version of OCR in 1950's. Since then several character recognition systems have been proposed for English, Chinese, Japanese and other similar languages that use isolated characters. There also has been considerable work on Arabic OCR also but however very little work has been done for Urdu OCR. The Nastalique style makes OCR in general and character segmentation in particular, a highly challenging task, so most of the researchers avoid the character segmentation phase and go in for higher unit of recognition. For Urdu, the next higher recognition unit considered by researchers is ligature, which lies between character and word. A ligature is a connected component of one or more characters and usually an Urdu word is composed of 1 to 8 ligatures. Most of the papers in literature related to Urdu OCR, deal with a very limited number of Urdu ligatures [1,2] or recognize only isolated Urdu characters[3]. The only work reported which deals with a sizable of ligatures is by Javed et. al. [4]. The authors have trained their system for recognition of 1282 ligatures, but their system suffers from the drawback that it cannot differentiate between similar looking ligatures, which have same primary connected component.

As suggested by most of the researchers, the ligature is taken as recognition unit, though no frequency analysis has been provided by any researcher regarding the exact count of frequency of ligatures and which ligatures have to be trained. In this paper, we have looked at character segmentation issues in Urdu and suggested alternative recognition units. The number of these units is much bigger than Urdu characters but lesser than the ligatures, though still their count is high. To arrive at a manageable number of these units, we performed frequency analysis of an Urdu corpus and the top most frequently occurring units, which cover 99% of Urdu corpus have been suggested for OCR training and classification.

2. INTRODUCTION TO URDU

Urdu is an important South Asian language spoken by nearly 250 million people in India, Pakistan and other neighboring countries. Urdu was derived from the Farsi alphabet, which itself is derived from the Arabic alphabet. There also has been considerable work on Arabic OCR also but however very little work has been done for Urdu OCR. Urdu script uses superset of Arabic alphabet, but uses Nastaliq writing style, while Arabic uses Nashak style. Nastaliq script is highly cursive, context sensitive and is written diagonally from top right to bottom left with stacking of characters, which makes it very hard to process for OCR. Machine printed Arabic (Naskh style) text usually follows a horizontal base-line where most characters, irrespective of their shape, have a horizontal segment of constant width. If we can separate these horizontal constant-width segments from a ligature, the remaining components of the ligature could be recognized. However, this is not the case with Nastaliq which has multiple base-lines, horizontal as well as sloping, making Nastaliq a complex style for optical recognition. As in Persian and Arabic, short vowels are not written in Urdu except in specialized texts such as dictionaries and textbooks. Urdu is written from right to left just like Arabic and Persian. From OCR point of view, Urdu has 40 basic letters as shown in figure 1.

DAR '12, December 16 2012, Mumbai, IN, India

In addition Urdu characters may have diacritics which are written as strokes and change the pronunciation and the meaning of the word. Diacritics may appear as strokes above the character as *zabr, pesh, shadda, tanween, maddah* or below the characters as *zer* etc. There are 10 commonly used diacritical marks in Urdu. As can be seen in figure 1, the Urdu letters are composed of a basic stroke and optionally dots or special characters. The letters can be classified into 21 classes based on their basic strokes (1). For development of a complete Urdu OCR system, besides the above characters, the OCR should also recognize other commonly used characters/symbols in Urdu. These include:

1. Digits, both Roman and Urdu digits are used in Urdu text (20)
2. Punctuation Marks (7)
3. Honorific (5)
4. Poetic marks (2)

Like Arabic, in Urdu each letter can have upto four different shapes: isolated, beginning connection from the left, middle connection from the left and right, and end connection from the right. Most of the letters can be connected from both sides; the right and the left. However there are eleven letters, designated as non-joiners, that can be connected from one side only; the right as shown in figure 2. Thus these eleven letters can have two shapes while the 27 letters can have four shapes. The special character *hamza* does not join with any character and instead lies above or below a character and has single shape only.

1				آ	ا
2	ٹ	ٹ	ت	پ	ب
3		خ	ح	چ	ج
4			ذ	ڈ	د
5		ژ	ز	ڑ	ر
6				ش	س
7				ض	ص
8				ظ	ط
9				غ	ع
10					ف

11		ق
12	گ	ک
13		ل
14		م
15	ں	ن
16		و
17		ہ
18		ھ
19		ء
20		ی
21		ے

Figure 1. Urdu Character Set classified on basic strokes.

آ ا د ڈ ذ ر ڑ ز ژ و ے ں

Figure 2. Non-joiners in Urdu.

ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ

ف ق ک گ ل م ن ہ ی ھ

Figure 3. Joiners in Urdu.

From recognition point of view, we have 131 possible shapes for the basic Urdu letters, which combined with daicritic marks, punctuations marks, digits and other special characters total 175. Thus the minimum number of recognizable Urdu character/symbol shapes are 175, provided they can be segmented properly. But this as we shall see in later sections, is a really tough job.

A group of joiners and/or non-joiner joined together form a ligature. A ligature ends either with a space or with a non-joining character. We can easily identify a ligature as a connected component of characters. A word in Urdu is a collection of ligatures and isolated characters. As for example, the word (ہندستان) (*Hindustan*)is composed of two ligatures : ہند , ستا and single character ن

The two ligatures are further composed of three characters.

ہند = د + ن + ہ and ستا = ا + ت + س

As mentioned in literature, a ligature usually has upto 8 Urdu characters. But we observed that for words borrowed from foreign languages such as English, the number of characters per ligature were even more than 8.

From OCR point of view Urdu is a very tough script. The main challenges are:

- Urdu script is written diagonally from top right to bottom left with stacking of characters (figure 4). The ligatures are tilted at a certain angle towards the right side. Numerals add to the complexity as they are written from left-to-right Due to this diagonal nature the Nastaliq consumes less horizontal space as compared to Naskh, but from OCR point of view it creates problems in word and character segmentation.
- Context sensitive nature of Nastaliq compels the letters to adapt different shapes. The shape of a character not only depends on its position (at the start, in the middle or at the end) in the word but also depends on surrounding characters in the word. Usually, the shape of a character is mostly dependent on the shape of the character that follows it, while the shape of a final character in a ligature is dependent on the second to last character. In fact a character may have upto 45 different shapes. For example *bay*, which is the second letter in the Urdu alphabet has 23 different shapes for its initial form [5].

حسین فلسفہ کے حق (1797- 1869ء)

Figure 4. A Sample Urdu text.

- Line and word segmentation are non-trivial tasks. We have frequently merging lines and words in adjacent lines merged (figure 5). For word segmentation, there is not much difference between inter-word and intra-word vertical gap and it is very difficult to judge if the two adjacent ligatures belong to same or different words. As for example, the text in figure 6 is composed of six ligatures and three words, but it is very difficult to make out from vertical gap, which ligature belongs to which word. There is vertical overlapping both within ligatures and among ligatures, which further complicates word segmentation (figure 7).

Figure 5. Overlapping text lines and merged ligatures in adjacent lines.

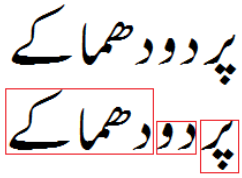


Figure 6. Confusing Inter-ligature and inter-word gap.

- Character segmentation poses another big challenge, as the characters are completely merged and it is very difficult to detect the character segmentation points.
- The letters are usually written in thick style and the diacritic marks and dots are usually written in small size, many times the holes of the diacritic marks or normal characters get filled or the parallel lines get merged, resulting in distorted shapes (figure 8). Besides frequent merging of dots and diacritic marks, the adjacent ligatures also get joined or even the ligatures in consecutive lines get joined (figure 5). This poses a big challenge in ligature and line segmentation.

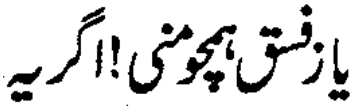


Figure 7. Vertically over lapping Urdu ligatures.

- Several Urdu characters (17 out of 40) are differentiated by the presence of dots placed over, below or within them. Three situations of ambiguity arise because of this. In the first instance, one character may have a dot while the other does not Fig 9(a). In the second case, two similar characters have different numbers of dots to distinguish their different sounds Fig 9(b). Lastly, two characters may be different only because of the difference in the position of dots Fig 9(c). Dots may appear as two separated dots, touched dots, hat or as a stroke.

Figure 8. Merged dots and special characters.

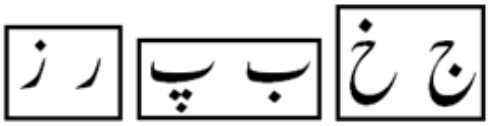


Figure 9. (a, b, c) Similar shaped ligatures distinguished by dots.

Even though the smallest recognizable units for Urdu are only 175, but practically it is almost impossible to break Urdu text into those 175 recognizable units, because of issues in character segmentation.

3. CHARACTER SEGMENTATION

For character segmentation first the Urdu word is broken into ligatures and isolated characters. The ligatures are then further segmented into characters. Segmentation of word into ligatures and isolated characters is easily achieved by connected component analysis. But segmentation of characters from ligatures is quite a challenging task. The main challenges are:

- It is very difficult to get the segmentation cue points in the ligature. In majority of the scripts such as Latin, Chinese etc, characters are separated by space and can easily be segmented by looking vertical gaps. For Indic script and Arabic script, the segmentation points are located on the headline and baseline respectively. But there are no similar cue points. As can be seen in figure 10, the ligatures have to be segmented into the individual character pieces as shown in adjacent table, but it is almost impossible to define the cutting points.
- The character segmentation has to be performed both in horizontal and vertical directions, as shown in figure 10(b).
- The size of the initial and middle forms of Urdu letters are much smaller as compared to the final form or isolated forms, which makes it difficult to identify and segment these forms.

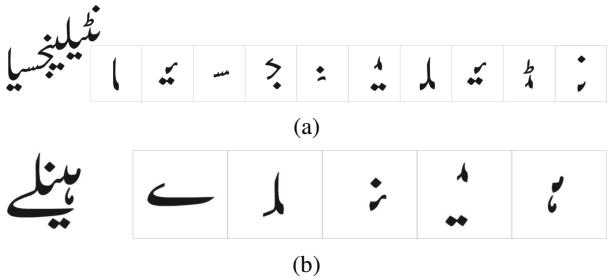


Figure 10. Segmentation of ligature into characters.

To avoid the difficulties of the character segmentation stage, researchers follow the holistic (or global) approach in which the recognition is globally performed on the whole representation of words and where there is no attempt to identify characters individually. In particular for Urdu, almost all the researchers have suggested going in for holistic recognition of Urdu ligatures, though no one has suggested how many and which ligatures have to be considered for training and recognition. We also propose taking the connected components of the ligature along with the diacritical marks and isolated forms of Urdu letters as the basic recognizable units.

The number of ligatures to be trained for recognition is not exactly known. As mentioned in literature, there are around 20,000 ligatures in Urdu. But we observed, from the analysis of Urdu corpus, that some of the valid ligatures had not been considered and the actual count turned out to be even higher. The results of the analysis are shown in Table 1 and 2.

Table 1. Some Statistics Generated from Urdu Corpus.

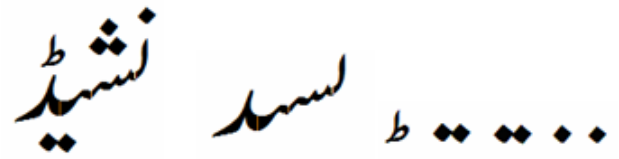
Total Urdu words	6,533,057
Total Ligatures	11,585,993
Unique Ligatures	25,957
Ligatures per word	1.77
Maximum Ligatures in a word	9 (انورادھاپورہ)
Maximum characters in a ligature	10 (سٹیبلشمینٹ)

Table 2. Characters count in ligatures.

Characters in Ligature	Unique Ligature Count	% Frequency Occurrence
2	1053	61.58
3	5842	27.45
4	9893	8.23
5	6168	2.36
6	2193	0.33
7	588	0.04
8	104	0.008154
9	32	0.001126
10	11	0.000299

4. PRIMARY AND SECONDARY CONNECTED COMPONENTS

The development of nearly 26,000 class character recognition system is a big challenge and some methods have to be devised to reduce the class count. From OCR point of view, a ligature consists of one primary connected component and zero or more secondary connected components. The primary component represents the basic shape of the ligature, while the secondary connected component corresponds to the dots and diacritics marks and special symbols associated with the ligature. As for example, the ligature in figure 11a, has one primary connected component (figure 11b) and five secondary connected components (figure 11c).



(a) (b) (c)

Figure 11. Urdu Ligature segmented into primary and secondary components.

Breaking ligature into primary and secondary components will reduce the recognizable units. There are multiple ligatures having same primary connected components and they are distinguished by secondary connected components. As an example, we have 17 ligatures (figure 12) sharing the same primary connected component. These ligatures can be identified by the secondary connected components. Thus instead of having 17 classes for these ligatures, we can have a single class for the primary component.

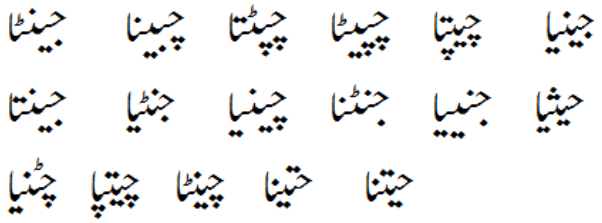


Figure 12. Some samples of ligatures containing same primary connected components.

For developing the complete ligature recognition system, we have to identify all the possible primary and secondary connected components. The number of secondary connected components is 22. A sample of secondary connected components is in figure 13.

Figure 13. Samples of secondary connected components.

But there is no available statistics regarding the count and frequency of primary connected components. The count and frequency of primary connected components is obtained by analysing all the 26,000 ligatures and separating out the ligatures having same primary connected components. After identifying the ligatures sharing same primary components, we found that 14815 primary components were needed to cover all the 26,000 ligatures. We also observed that 70.19% of primary components correspond to a single ligature only, while 15.32% of primary components are common for two ligatures and 1.48% of primary components are common for 10 or more ligatures. From coverage point of view, 11.05% of primary components are similar for 10 or more ligatures. The highest number of ligatures having same primary component is 71 (figure 14).

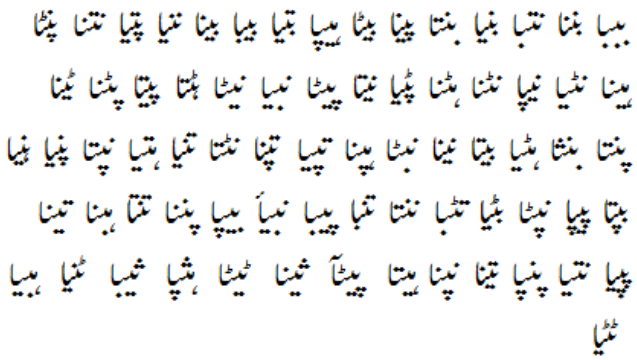


Figure 14. Ligatures with same primary connected components.

We also did coverage analysis to determine the least amount of ligatures and primary components for different corpus coverage. The results are tabulated in Table 3. From the table it is clear that 27 ligatures and 18 primary components account for 50% of ligature and primary components respectively. Similarly 138 ligatures and 73 primary components account for 75% of ligature and primary components respectively, while 632 ligatures and 297 primary components account for 90% corpus coverage. For 95% corpus coverage 1364 ligatures and 652 primary components are needed, while 2114 ligatures and 1026 primary components account for 97% of corpus. For 100% corpus coverage, 25957 unique ligatures and 14815 unique primary components are needed.

This is useful information for classifier design, as this gives idea about the total number of classes that need to be recognized. One can use this information to decide the classifier design. As for example, instead of developing a single classifier for all the 14815 primary components, one can design a two-tier classifier system, in which the first classifier can be trained to identify the first 652 primary component classes, which covers 95% of the corpus data, while the second classifier can be trained to the first classifier fails to recognize the image with high confidence. Alternatively, we may increase the number of primary components in first classifier from 652 to 2190, which will cover 99% of the primary components in the corpus.

We use this coverage analysis, to develop Urdu OCR system using 2212 classes, corresponding to 2190 primary connected components and 22 secondary components. Besides ligatures, the 40 isolated forms of Urdu characters have also to be considered for recognition. These 40 isolated Urdu letters can be further broken into 21 primary connected components and secondary components. Thus the total number of classes needed for recognition of Urdu text for 99% corpus coverage are 2233(2211 primary components and 22 secondary components).

The ten most frequently occurring primary components, along with their percentage frequency of occurrence, number of characters present in them and the number of ligatures sharing these primary components are displayed in Table 4. We can observe from the table that the most frequent primary component has 4.19% frequency of occurrence, contains two characters and 16 ligatures share this component.

Table 3. Ligature and Primary Connected Component Coverage.

Coverage	Ligature	Primary Component
50%	27	18
75%	138	73
80%	212	108
85%	350	170
90%	632	297
95%	1364	652
96%	1667	804
97%	2114	1026
98%	2883	1385
99%	4657	2190
100	25957	14815

Table 4. Ten most common primary connected components.

Primary Connected Component	Percentage Frequency	No. of characters	Ligatures sharing
	4.19	2	16
	3.33	2	1
	2.79	2	8
	2.76	2	23
	2.64	3	7
	2.48	2	1
	2.45	2	1
	2.16	2	6
	2.07	2	12
	2.04	2	2

5. TOUCHING PRIMARY AND SECONDARY COMPONENTS

The secondary components are placed very close to the primary components and this along with diagonal writing style of Nastaliq script frequently results in secondary components touching other components. After a detailed analysis on images taken from 40 documents, we found that 3.28% of secondary components were touching other primary or secondary components.

The touching secondary components can be categorized as:

a. **Secondary components touching primary components in same ligature:** This is the most common cause of touching components, as the secondary components are usually placed very close to their associated primary components. The secondary components frequently touch their associated primary components and in the process get merged with the primary component. It was found that the most

commonly touching components are double dots, parallel line and Urdu letter *hey* (figure 15a).

b. **Secondary components touching primary components in adjacent ligatures:** As the adjacent ligatures frequently vertically overlap, many times it happens that the secondary components of one ligature touch the primary component of adjacent ligature (figure 15b).

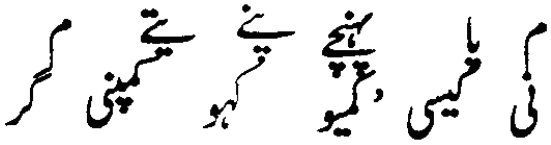
c. **Secondary components touching primary components in adjacent rows:** As Nastaliq is written diagonally, so it frequently happens that the components of ligatures lying in adjacent rows touch each other. As, we can see in figure 15c, some of the secondary components of ligatures get merged completely with primary components of ligatures lying in adjacent row.

d. **Secondary components touching secondary components in adjacent rows:** It was also observed that many times the secondary components ligatures lying in adjacent rows get merged and it is very difficult to identify and separate the individual components from the merged cluster (figure 15d).

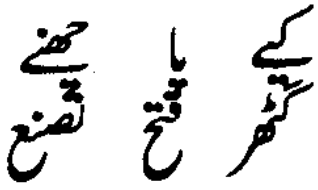
All these issues further complicate the recognition process, which makes recognition of Urdu script a real challenge. We have tried to alleviate the problem of touching secondary and associated primary components by creating a separate class of such frequently occurring components. We identified 95 such touching components.

(a) Touching primary and secondary components in same ligature

(b) Touching primary and secondary components in adjacent ligatures



(c) Touching primary and secondary components in adjacent rows



(d) Touching secondary components in adjacent rows

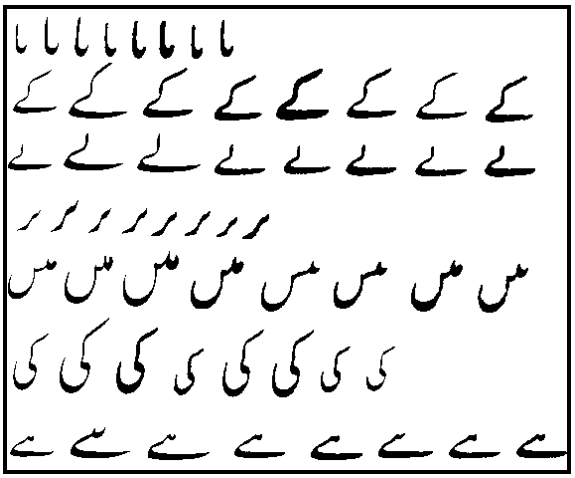
Figure 15. Touching primary and secondary components.

6. TRAINING DATA

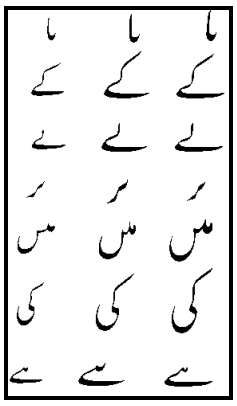
The next important task is creating the training data corresponding to 2233 connected components, for training and testing the OCR. But development of the training data for these components turned out to be quite a time consuming process. Urdu typography has been a great challenge for the printing and publishing industry. Because it is composed of complex and shifting letters, typesetting technology based on letter analysis could not render the richness of the script. So until 1995, all the Urdu books and newspapers were hand written by masters of calligraphy known as *katibs*. Thus the Urdu text printed in books and newspapers can be divided into two generations. The books printed before 1995 are all hand written while majority of the books published after 1995 use computer generated Nastaliq fonts such as Noori Nastaliq. Thus for our experimentation, we collected a corpora consisting of two sets of images (figure 16) to cover both the generations:

1. Real-world images consisting of scans of commonly available hardcopy documents published before 1995 (8 samples each). But collection of training data from hand written books is not an easy task, as one has to manually search and scan for the 2000 primary components, which has turned out to be quite a cumbersome task.
2. Computer generated, i.e. synthetic, images using the Nastaliq Urdu fonts (3 samples each)

Besides the samples corresponding to the 2233 connected components, we also created training set for the 95 touching components, as discussed in previous section (figure 16).



(a)



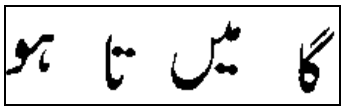
(b)

Figure 16. Samples of: (a) Handwritten and (b) Machine Printed Urdu primary components.

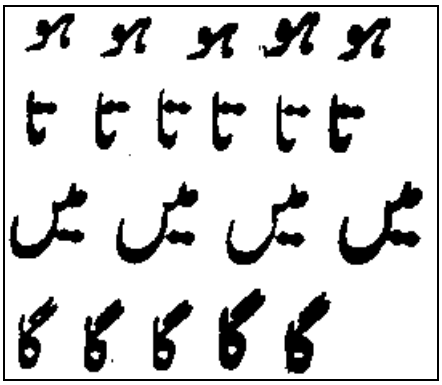
The inclusion of these 95 touching components raises the count of recognizable classes to 2328.

7. CONCLUSION

In this paper we have discussed, the derivation of recognizable units for Urdu OCR. An Urdu word is composed of ligatures and isolated characters. For character segmentation, the ligature has to be broken into individual characters. The number of recognizable units in Urdu OCR is 175, but as segmentation of ligatures into characters is a Herculean task, so most of the researchers avoid the segmentation stage and go in for holistic recognition of ligatures. But there is no data available on the exact number of ligatures to be trained for recognition and their coverage. From the statistical analysis of Urdu corpus, it was found that for developing an Urdu OCR system, at least 26000 classes had to be considered if we take ligatures and isolated letters as recognition unite. From ligature coverage statistics, It was found that for 99% coverage, 4657 ligatures had to be trained. To reduce the recognition classes, we have segmented the ligatures into primary and secondary connected components by connected component analysis.



(a) Some of the common ligatures with touching components



(b) Training files for touching components of ligatures in figure 17a

Figure 17. A sample training data for touching secondary components.

The segmentation of ligatures into primary and secondary components reduces the number of classes from 4657 to 2212 (2190 primary connected components and 22 secondary components) for 99% corpus coverage. Besides ligatures, the 41 isolated forms of Urdu characters have also to be considered for recognition. These 41 isolated Urdu letters can be further broken into 21 primary connected components and secondary components. Thus the total number of classes needed for recognition of Urdu text are 2233(2211 primary components and 22 secondary components). It is also found that, as the secondary components are placed very close to the primary components and this along with diagonal writing style of Nastaliq script frequently results in secondary components touching other components and 3.28% of secondary components were touching other primary or secondary components. To take care of these touching components, we also created training set for the 95 touching components. The inclusion of these 95 touching components raises the count of recognizable classes to 2328, which will be covering 99% of Urdu corpus data.

8. ACKNOWLEDGEMENT

This research work is sponsored by Ministry of Communications and Information Technology under the project : Development of Robust Document Analysis and Recognition System for Printed Indian Scripts**.**

9. REFERENCES

[1] S. A. Husain and S. H. Amin. A Multi-tier Holistic Approach for Urdu Nastaliq Recognition. *IEEE INMIC,* Dec. 2002, Karachi.

[2] S. A. Sattar, S. Haque, Shamsul Haque and M. K. Pathan. Nastaliq Optical Character Recognition. *Proceedings of the 46th Annual Southeast Regional Conference on XX,* 2008, pp 329-331.

[3] Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan. Urdu Nastaleeq Optical Character Recognition. *Proceedings of World Academy of Science, Engineering and Technology*, Volume 26, 2007, pp. 249-252.

[4] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Moin. Segmentation Free Nastalique Urdu OCR. *Proceedings of World Academy of Science, Engineering and Technology*, 46, 2010, pp. 456-461.

[5] S. T. Javed, S. Hussain. Improving Nastaliq Specific Pre-Recognition Process for Urdu OCR. *In the Proceedings of 13th IEEE International Multitopic Conference 2009 (INMIC 2009)*, Islamabad, Pakistan, 2009, pp. 1-6.