Urdu Writing Rules for Online Input in PDA's

Fareeha Anwar

Lecturer, Department of Computer Science International Islamic University Islamabad fareehaanwar@iiu.edu.pk

Abstract

For online input, stroke squence based recognition is generally employed. For this method, the stroke sequence must be uniquely defined for every character/Ligature. Normally, every language has unique writing rules, which are followed by experienced users and recognition engines. However, there are variations in writing style from person to person and place to place. Langauges which are written from right to left e.g. Urdu, Arabic, Persian etc. are complex and have a lot of variations due to fonts and writing style. If rules are not followed properly, the recognition engine is bound to fail. Therefore, proper writing rules are necessary for online recogniton of any language based on stroke sequence. There are no published and acknowledged rules available for online input in PDA's so far for Urdu language.

This paper is an effort in accumulating writing rules for 'Nastalique' font for online urdu recogniton engine.

1. Introduction

Writing on tablet, PDA or any online input device, generates a sequence of strokes. These stokes are sent to recognition engine which then accepts or rejects the strokes based on predefined rules. Different fonts are available for single language just like roman bases languages. Each font has its own writing rules which vary from each other and also has influence on each other which causes confusion in reception. People writing Nastalique can use Nasakh rules and vice versa is also possible. Arabic S. Afaq Husain

Professor, Department of Electrical Eng. Swedish College of Engg & Tech drafaqh@gmail.au.edu.pk

has great influence on Urdu. There are different Arabic calligraphic styles having different level of difficulties [9,10,11]. The most common styles in Urdu language are Nasakh [8] and Nastalique [9]. Each has its own rule of writing. If stroke sequence is not followed according to predefined writing rules, the rejection rate increases. If user changes the direction of writing stroke; e.g if a diagnol line starting from top to bottom is writtin from bottom to top, its shape seems accurate but stroke sequence totally changed so online recogniton engine will not be able to recognize it. Therefore to improve the performance of recogniton engine, user should follow writing rules. We faced similar problems when developing online recognition engine for Urdu handwritten characters and ligatures[4]. Most common mistakes were pen up in middle of ligature/ character, writing ligature statring from opposite sequence *f*, when pen up is required continue writing generating dublicate sequence etc. Therefore need for devising predefined writing rules was felt. Unless these rules are known and the user are trained according to these rules, the recognition engine is bound to fail or give very high faliure rate. While searching for writing rules for Urdu language we came across a lack of well published or publicized writing rules. We have devised rules for Urdu language, which will increase recogniton rate. These rules will also be guide for new users to learn writing urdu language or to use computing device for automatic learning just like writing tutorial. Urdu is a cursive languages and very difficult to recognize as discussed in [6]. There are 38 characters in the Basic Urdu Character set given in figure 1.

Figure 1: Basic Urdu alphabets [3]

According to the revised extended character set in Urdu, there are a total 58 Urdu alphabets [3]. The new alphabet set of Urdu is shown in figure 2.

1.	ъ	ٹ	ٹھ	ٹ	à	ت	á,	Ų	ø,	Ļ	1	I	
ڙھ	ţ	æ	,	j	ڑھ	ţ	زھ	,	i	ι	ą	Ъ	
ک	J	ن	t	t	ظ	٩	ض	ص	Ĵ	U	¢	;	
وھ	,	ž	υ	à	ن	s	٢	لھ	J	\$	ſ	à	
								ے	ى	,	õ	,	
Figure	2:	Ch	ara	icte	r S	et (58	alp	hał	oet	s) of	f Ur	du

Script.[3]

Most of the Urdu characters have same shape in ligature provided the same context, so for easy and efficient recognition 38 characters are divided into 18 classes, shown in Table 1. In given table there is a class known as "Kashti" i.e. ship. We have given this name because basic shape of characters present in this class is very similar to ship — i.e. "Kashti" in urdu. In the Nastalique way of writing script, Urdu assumes four different locations of a character i.e. whether the character is isolated, in the beginning, at the end or connected from both the sides in a word [2]. Therefore, each character has different shape according to position in a given ligature

Each character has multiple shapes, varying from 4 to 20+ depending upon location of a character with in a ligature and also on previous and next character of the ligature. Therefore the combination of each of the alphabet with the defined 18 classes will have its own contextual shape. For example if "Kashti" is at start position then it has two different shapes when

followed by "Jeem" Eand "Aien" Eclass.

Diacritics: Diacritics are very important in Urdu language. These include diacritics such as *Dots*, *Ttaay*, *Hamzaa*, *Diagonal and Madaa*, etc. In this paper diacritics are not discussed

Figure 3: Diacritics/Aerab of Urdu [5]

Table 1: Classification of Urdu characters

Table 1: Classification of Ordu characters									
ف ق	10	Alif J	1						
Kaaf J.	11	Kashti ب پ ت ٹ ث	2						
Laam しじ	12	Jeem えてきひ	3						
Meem	13	و ژ ژ Daal	4						
Wao 9.	14	Ray プリプノ	5						
Gool hay	15	یں ش Seen	6						
Do chashmi hay 20	16	ص ض Swad	7						
Choti Yaa G	17	طظ Tua	8						
Bari Yaa 🖌	18	Aien É C	9						

2. BASIC RULES:

Nastalique is actually written from top right to bottom left.

(1)
$$\frac{1}{K_a}$$
 $\frac{1}{K_a}$ Ba (Basic phonems)

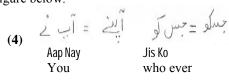
Each ligature that starts with or or and ends with is tilted at approx. 45 degree. This is of particular significance as there is no fixed level or height for any character with respect to base line.



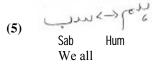
Angle of each part of stroke can be measured from starting points and horizontal baseline as shown in figure below. 180°

Angles: Downward 210+20 and upward 30+20

Write words separately if it is possible as shown in figure below.



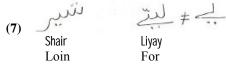
Normally writers give distance between two words.



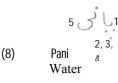
Long spaces should not be given between ligatures within a single word



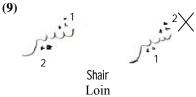
All cusps/"shosha" should be drawn properly. Length of cusp should be proper so that it differentiates the alphabets with in ligature.



Secondary strokes of first ligature are drawn before second one if word is composed of more than one ligature.



Secondary strokes should follow proper sequence.



Stroke written using full qat (length of nib" never joins another full qat, rather it always joins with a half *qat* glyph

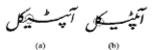


Figure 3: (a) Full qat kashish joined with half qat connector, (b) Full qat kashish joined with full qat circle

3. Shapes of Classes

The Urdu calligraphy is measured with respect to pen width, i.e., whether it is written with the full width of the pen or half width i.e., vertical nib. In online writing we use skeleton or thinned version. It is just the stroke order that is important and not the thickness. In online writing we use tablet PC, in which we consider resolution. The measurement of resolution is dpi (dots per inch) i.e. pixels per inch. So we use pixel as unit of measurement in online writing. Each tablet has different resolution varying from 600 dpi to 1200 dpi depending upon hardware. In tablet or PDA writing is done in a box or window. Each window has particular size. If a stroke length is greater than 3/4 of maximum window size then we call it long line and if length is less than 1/4 of window size than it is considered as short line. If stroke length is between 1/4 and 3/4 of window size then it is considered as medium line. In preprocessing step, writing is rescaled to standard size which makes it scale invariant.

In each class, there is one or more than one character. Each differs in number and shape of secondary strokes.

3.1. Alif class:

Shape of Alif is long vertical line; direction is

from top to bottom.

- Shape Alif remains same at the isolated, start, middle, or end position of any ligature.
- It remains isolated when comes at start of any ligature.
- Stroke direction changes when it comes after other character i.e., bottom to top instead from top to bottom.
- ➤ When Laam is followed by Alif, IJ it will be slanting line from top to bottom joining the base of Laam rather than the usual vertical line.

3.2. Kashti class:

Five characters are present in this class named "Bay", "Pay", "Taay", "Taay", "Ttay", and "Saay".

<u>Basic</u> short vertical line (top to bottom) followed by a long horizontal line (right to left), then

short vertical line (bottom to top).

If "jeem class" follows "kashti" ^L then its shape is simple middle diagonal line at 210±20 degrees (top right to bottom left).



If "kashti" is followed by characters having loop in them e.g. "fay", "Aien", "qaaf", "wao" etc., ray" and "yay" class, then its shape is simple "ray" or half "kashti" i.e., short vertical line (top to bottom, 90±20) followed by short horizontal line (right to left, 180±20). There is no cusp after it.

If "meem" follows Kashti then its shape will be short curve or diagonal line (top to bottom) at

230<u>+</u>20 degrees.

If "Alif", "Kaaf" or "Laam" follows Kashti then its shape is semi circle (right + downward + left

> For rest it is semi circle (from right + downward

+ left + upward). There is cusp at the end. Us Shape at middle position

- If Kashti comes in middle, then it is semi circle (right + downward + left + upward). There is cusp at start as well at end.
- All rules for the shape at stating position are same for shape at middle position, including cusp at start.

Shape at end position

At the end position, shape is same as isolated but cusp at start.

3.3. Jeem class:

Four characters are present in this class named "Jeem" $\mathbf{\tilde{C}}$, "Chay" $\mathbf{\tilde{C}}$, "Hay" $\mathbf{\tilde{C}}$ and "Khay" $\mathbf{\tilde{C}}$.

<u>Basic Shape</u>: medium horizontal line from left to right, followed by downward curve from left to right (angle: downward 210 ± 20 and upward 30 ± 20).

• Alternatively, short vertical line from bottom to

top at start and rest shape is same as above.

Half Jeem (before curve) either of two given ways followed by other character.

Shape at middle position

- Middle diagonal line of 225±20 degrees, followed by sharp edge and diagonal line at 330±20 degrees followed by sharp edge and diagonal line/curve at 210±20 degrees
- If "Jeem" class follows "Kashti", then its shape is half "Jeem" starts from horizontal line, not from vertical line. *

Shape at end position

When it comes at end, its shape is same as isolated, starts from horizontal line, not from vertical line.

3.4. Daal class:

Three characters are present in this class named

"Dal", "Ddal" and "Zal.

Basic Shape: Medium diagonal at 315 ± 20 degrees from left to right, followed by either downward curve or medium diagonal line at 210 ± 20 degrees from

right to left. 🤰

Shape at start position

Remain same as of isolated.

Shape at middle position

Never comes in middle position of ligature

Shape at end position

At the end of any ligature it starts with cusp followed by medium vertical line (top to bottom) and then medium horizontal line

3.5. Ray class:

Four characters are present in this class named,

"Ray", "Aray", "Zay" and "ZYaay"

Basic Shape: Starting from right to left, its shape is medium vertical line of 250+20 degrees, followed by

horizontal line of 180<u>+</u>20 degrees

Shape at start position

Remain same as of isolated.

Shape at middle position

> Never comes in middle position of ligature

- Shape at end position
- At the end of any ligature, its shape is medium vertical line at 225+20 (top to bottom) and then

medium horizontal line 🚣 🦯

3.6. Seen class:

Two characters are present in this class named

ش "Seen" C and "Sheen"

Basic Shape: Two small semi circles (right + down+ left + up) having short diameter, with two cusps followed by a big semi circle having long diameter.

<u>Shape at start position</u>

- First part of shape i.e. two semi circles remains same with two cusps There is no last big semi circle.
- If "Jeem" follows "Seen", "Meem" or "yaa" class then after second semi circle there is no cusp but start the shape of next character.
 If "E

Shape at middle position

- Start shape remain same but instead of two, three cusps are present including starting cusp as well.
- All conditions of start position remain same. \triangleright

Shape at end position

Remain same as of isolated. Instead of two, it has three cusps (start cusp as well).

3.7. Swad class:

Two characters are present in this class named

"Swad" U and "Zwad"

Basic Shape: Start from left its shape is short diagonal line making an angle of 30 ± 10 followed by small downward curve and move back towards starting location making a loop. After starting loop there is small vertical line followed by a cusp and a semi circle (down+ left + up) having long diameter same as last part of "Seen"

Shape at start position

First part of shape i.e. loop with cusp, remains same. Instead of last semi circle, the shape is

small "Ray" with cusp.

Shape at middle position

- Whenever it comes in the middle of ligature, there are two ways of writing it
- 1. There is pen up and shape remains same as of start shape.

(Pen up)

- There is continuous stroke connected with previous character. Shape remains same as of start shape with no pen up.
- > There is cusp at the end. All conditions of start position remain same.

Shape at end position

- Whenever it comes in the end of ligature, again there are two ways of writing it as discussed in "shape at middle position", i.e. with pen up and without pen up.
- \triangleright Shape remains same as of isolated.

3.8. Tua class:

Two characters are present in this class named "Tua" and "Zua"

Basic Shape: Straight vertical line from top to bottom making an angle of 270+20; same as "Alif"; followed by a cusp and stroke same as "Dal" with curve (not with diagonal line) then move back towards starting location making a loop. After loop,

there is small horizontal line d

Shape at start position

- Shape remains same as of isolated. \geq Shape at middle position
- Whenever it comes in the middle of ligature,
- there is pen up and shape remains same as of وط

- (Pen up)
- \geq All conditions of isolated position remain same.

Shape at end position

Whenever it comes in the end of ligature, there is pen up and shape remains same as of

3.9. Aien class:

Two characters are present in this class named "Aien" \mathcal{E}_{and} "Ghaien" $\dot{\mathcal{E}}$

Basic Shape: Starting from top right making a diagonal line/curve in downward left direction making angle of 225+20 followed by small curve then move back towards right (shape same as "Dal" but in opposite direction) and same as small semi circle having medium diameter. Then there is a cusp and big semi circle (down-left + right + up) having long diameter.

Shape at start position

- First part of shape (small semi circle with cusp) remains same. Instead of second curve, there is small horizontal line/curve toward left.
- \triangleright If "Meem" or "yaa" class follows "Aien" then after first part of shape next character starts at once. 2, 3, - 3
- If "Aien" is followed by "Jeem" class then after \triangleright first part there is small diagonal line at angle 225+20 2

Shape at middle position

Whenever it comes in the middle of ligature, there is a small upward diagonal line/curve making an angle 135+20 followed by small horizontal line from left to right and finally small downward diagonal line/curve making an angle 335+20. This will make a loop. There is no cusp after it.

Shape at end position

At end of ligature, first part remains same as "shape at middle" and after loop there is a semi circle(down left + right + up) having long diameter (same as second part of Aien shape without cusp)

3.10. **Fay class:**

Two characters are present in this class named "Faay" نفand "Qaaf" تن

Basic Shape:

- "Faay" Start from down and move towards left, up, right then down make a circle of short radius (loop). After loop, the shape is same as "Kashti".
- "Qaaf": loop remains same and after loop there is a big semi circle having long diameter (right +down +left+ up). Ü

Shape at start position

- If both characters occur at start then the shape is only first part i.e. loop.
- ► If "Fay" is followed by "Meem" or "yaa" class

لى في في في المعاد then after loop next character starts at once.

> If "Fay" is followed by" U then after loop there

is small diagonal line at angle 225+20

Shape at middle position

- Shape at middle is same as shape at start position only connected with previous character. There is no cusp after it.
- > All conditions remain same as of start position.

Shape at end position

 At the end of ligature the shape is same as of basic shape بقب طبق

3.11. Kaaf class:

Two characters are present in this class named "Kaf" \checkmark and "Gaaf" \checkmark

Basic Shape: At start, there is medium vertical line; instead of short vertical line; rest part is same as basic shape of "Kashti" class. After this "Kashti" like shape there is a pen up and we draw a secondary

stroke i.e. "Kash" ***** (long diagonal line). There is single "Kash" in case of "Kaf"(single pen up) and two "Kash"s in case of "Gaaf" (two pen ups).

"Kash" is a diagonal line of angle 220+20 starting from top right to bottom left.

We will write "Kash" after writing whole ligature in all positions.

Shape at start position

If "Jeem", "Ray", "yay" class and all the classes having loop follows "Kaaf" then its shape is long vertical line (top to bottom, 270+20) followed by short horizontal line (right to left, 180+20).

There is no cusp after it $2 \sqrt{2} \sqrt{2}$. If "Alif" or "Laam" follows "Kaf" class then its

If "Alif" or "Laam" follows "Kaf" class then its shape is same as start position of "Fay" class, i.e. loop. 6

- All conditions of "Fay" at start position remain same.
- For rest it is semi circle (from right + downward + left + upward) with straight vertical line. There

Shape at middle position

- If Kaaf comes in middle, then there is a vertical line from top to bottom making an angle of 90±20 then downward stroke making an angle of 270±20 then small semi circle (right + downward + left + upward). There is cusp at start as well at end. جكب
- If "Alif" or "Laam" follows "Kaf" class then its shape is same as middle position of "Fay" class, i.e. loop
- All rules for the shape of Fay at middle position دیکل. لگا

Shape at end position

At the end position, shape is same as isolated but cusp at start.

3.12. Laam class:

Two characters are present in this class named "Lam" Jand "Nun" U

Basic Shape:

"Nun": it is same as last part of "Seen" i.e. semi circles (right + down+ left + up) having long diameter \mathbf{U}

"Lam": At start, there is vertical line from top to bottom making an angle of 270 ± 20 , remaining part is

same as "Nun". U

<u>Shape at start position</u> Nun": At start of any ligature, its shape is same as

shape of "Kashti" at start position.

All conditions associated with shape of "Kashti" at start position remain same.

"Lam": At start of any ligature, its shape is same as

shape of "Kaaf" at start position. لع , لا

- All conditions associated with shape of "Kaaf" at start position remain same. Except
- When "Alif" follows "Lam, its shape is vertical line followed by short horizontal line (right to left).

Shape at middle position

Nun": At middle of any ligature, its shape is same as

shape of "Kashti" at middle position. بيند

All conditions associated with shape of "Kashti" at middle position remain same.

"Lam": At middle of any ligature, its shape is same

as shape of "Kaaf" at middle position, بلكل

- ➢ All conditions associated with shape of "Kaaf" at middle position remain same.
- > Except When "Alif" follows "Lam, its shape is vertical line followed by short horizontal line

(right to left).

Shape at end position

At the end position, shape is same as isolated but سن فل cusp at start.

3.13. Meem class:

One character is present named "Meem"

Basic Shape Start from left its shape is short diagonal curve making an angle of 30+10 followed by small downward curve and move back towards starting location making a loop. After starting loop there is small horizontal line followed by a large vertical line from top to bottom same as "Alif".

Shape at start position

Start from up and move towards left, down, right then up make a circle of short radius (loop).

If "Alif", "Jeem", "Mem" or "Yaa" class is follows "Mem" class then the shape of next \geq

ی بی کا دharacter start at once.

For rest after loop, there is diagonal line making \triangleright an angle of 225+20

Shape at middle position

- Shape same as starting position joined with previous character
- All conditions of shape at start location remain \triangleright same.

Shape at end position

Loop remains same as of middle position. After

loop, shape is same as basic shape after loop.

3.14. Wao class:

One character is present named "Wao" Basic Shape: First, there is loop same as "Fay" and after loop there is a curve same as "Daal" facing towards left.

Shape at start position

Remain same as of isolated.

Shape at middle position

Never comes in middle position of ligature Shape at end position

Remain same as of isolated

3.15. Gol Hav class:

One character is present in this class i.e Gol Hay **Basic Shape:**

Its shape is same as circle, starting from right moving left and downward then right and

upward back to starting position making a loop of medium radius.

Alternatively, a curve of angle 235+20 then \geq moves right and at the end curve of angle 60 ± 20 . There is an intersection point making a loop. 🎽

Shape at start position

At start, its shape is same as shape of "Kashti" at start position U

Shape at middle position

A line of angle 235 ± 20 followed by a curve of angle 60+20. There is cusp point between these two curves.

Shape at end position

At the end of ligature, its shape is a small line of angle 235+20 followed by a small curve and a medium line of angle 150+20. When height of this line approaches starting point then again a curve followed by small line of angle 235+20 *****

3.16. **Do Chashmi Hay class:**

One character is present in this class named "Do chashmi Hay" 🌌

Basic Shape: There is a small line of angle 235+20 followed by a loop same as loop of "Swad". After first loop, there is another loop same as "Swad" intersecting previous one and moving back to starting position. At the end, there is small horizontal line.

Shape at start position

Same as isolated shape Shape at middle position

Same as isolated shape

Shape at end position

Same as isolated shape 4

3.17. Choti Yaa class:

One character is present in this class named ی "Choti Yaa"

Basic Shape: A small line/curve of angle 245+20 followed by a curve of angle 315+20. Then there is semi circle same as "Fay" class,

Shape at start position

At starting position its shape remain same as

shape of "Kashti" at starting position.

 \triangleright All conditions of shape of "Kashti" at starting position remain same here

Shape at middle position

At starting position its shape remain same as

shape of "Kashti" at middle position.

- All conditions of shape of "Kashti" at middle position remain same here
- Shape at end position
- If "ChotiYaa" follows "Kashti", "Jeem", "Fay", "Kaaf" or "Laam" class then its shape is only last part i.e. semi circle
- For rest its shape remains same as of isolated.

3.18. Bari Yaa class:

One character is present in this class named "Bari Yaa" -

<u>Basic</u> Shape: A small line of angle 245 ± 20 followed by big horizontal line from left to right making an angle of zero.

 \succ Or at start there is small vertical line from top to

bottom and rest part is same as previous 4

Shape at start position

Same as shape of "Choti Yaa" at starting position. *Shape at middle position*

Same as shape of "Choti Yaa" at middle_position. *Shape at end position*

Its shape remains same as isolated \leq

4. Results and Conclusion

We have tested the online Urdu OCR developed earlier [4] by using 10 native Urdu writers and found out that there were a number of ambiguities in variations which result in failure in recognition by the engine. Moreover, the standard writing rules described above have not been devised keeping in view the online input and as such are not efficient in writing using a stylus or digital pen. It is therefore recommended that writing rules be modified keeping in view stroke based convenience and efficiency for online input devices.

Urdu language is difficult to learn and write. If proper writing rules are not followed in online input then recognition rate decreases as shape seems same but stroke sequence is completely changed.

We also analyzed that there are many pen ups required in writing different ligatures which make recognition process slow, e.g. if we want to write \checkmark , first we write \checkmark then a pen up and write the stroke. \checkmark However, if we write this stroke with out pen up then this will be more convenient and efficient

5. Future Direction of research

Future work includes devising efficient and convenient rules for online input. This will make

recognition engine more efficient. Also, secondary strokes i.e., diacritics are not discussed in this paper, so a future direction of research would be to devise rules including diacritical marks.

6. Acknowledgements

We acknowledge the contribution and useful input of Dr. Sarmad Hussain, Kiran Khurshid and Mr. Jamil of NUCES for their invaluable assistance in conducting this research.

7. References

[1] Mukhtar Omer, Setlur, Srirangaraj "Experiments on Urdu Text Recognition" Guide to OCR for Indic Scripts, Advances in Pattern Recognition, Volume . ISBN 978-1-84800-329-3. Springer-Verlag London, 2009, p. 163

[2] Reza Safa bakhsh and Peyman Adibi, "Nastaaliq handwritten word recognition using a continuous-density variable-duration Hmm". The Arabian Journal for Science and Engineering, Volume 30, Number 1B. April 2005. ttp://ajse.kfupm.edu.sa/articles/301B_07P.pdf

[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 volume 26 December 2007.

http://www.waset.org/journals/waset/v32/v32-49.pdf

[4] S.A Husain, A Sajjad, F Anwar, "Online Urdu Character Recognition Engine" - MVA2007 IAPR Conference on Machine 2007 http://init.org.pk/papersandpublications/Paper5.pdf

[5] Aamir, Atif, Ayesha, Ahmad, Irfan, Saqib, Sara and Sheraz. "Contextual Shape Analysis of Nastalique", CRULP Annual Report 2001. http://crulp.org/Publication/Crulp report/CR04 01E.pdf

[6] http://eprints.ecs.soton.ac.uk/16510/1/ASattar_85. doc

[7] http://en.wikipedia.org/wiki/Nastalīq_script

[8] http://islamicart.com/main/calligraphy/styles/naskh.html

[9] http://www.al-bab.com/arab/visual/calligraphy.htm

[10] http://islamicart.com/main/calligraphy/styles/index.html

[11] http://www.sakkal.com/ArtArabicCalligraphy.html