

FONT SIZE COMPUTATION

Urdu Nastalique

Optical Character Recognition System

Qasim Ali

qasim.ali@kics.edu.pk

11-December, 2012

مرکز تحقیقات لسانیات



Center for Language Engineering
Al- Khwarizmi Institute of Computer Sciences
University of Engineering and Technology, Lahore, Pakistan

PEN WIDTH OR QAT

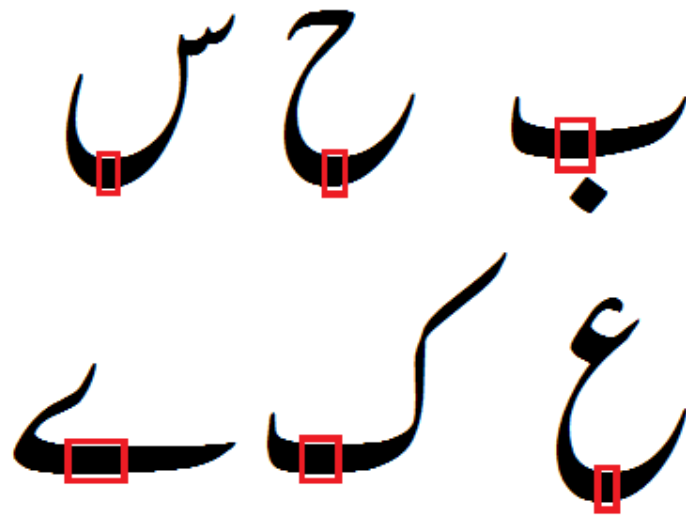


Figure 1 Qat indication in Main bodies

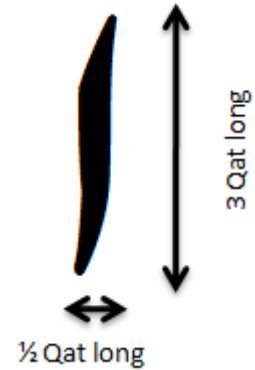


Figure 2 Qat information in character Alif

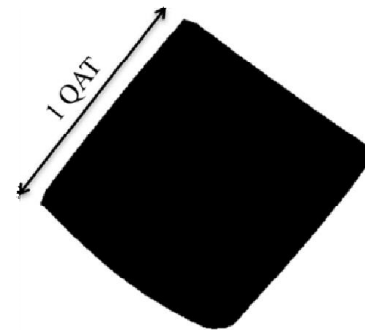


Figure 3 Qat information in 1 dot diacritic

DOTS IN URDU NASTALIQU

ب پ ت ث ج چ خ
ذ ز ژ ش ض ظ غ ف ق ن

Figure 4 Urdu characters with dot diacritics attached

Contd...

- Ratio of lines with dots to lines without dots is around 97.24%

(Tested over 587 pages)

- Ratio of dots to other diacritics in a line is around 95.71%

(Tested over 10,510 lines)

METHODOLOGY

(9) دائیں کریں کیس پر لگا ہوا آئل پمپ، کلچ اسمبلی اور دوسرے تمام پرزے علیحدہ کر لیں۔

Figure 5 A sample from image corpus of font size 16

(9) دائیں کریں کیس پر لگا ہوا آئل پمپ، کلچ اسمبلی اور دوسرے تمام پرزے علیحدہ کر لیں۔

Figure 6 Extracted dots

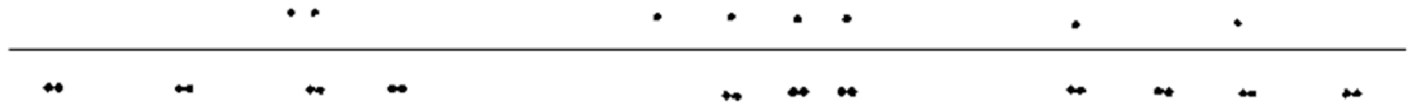


Figure 7 Separation of one dots and two dots

Contd...

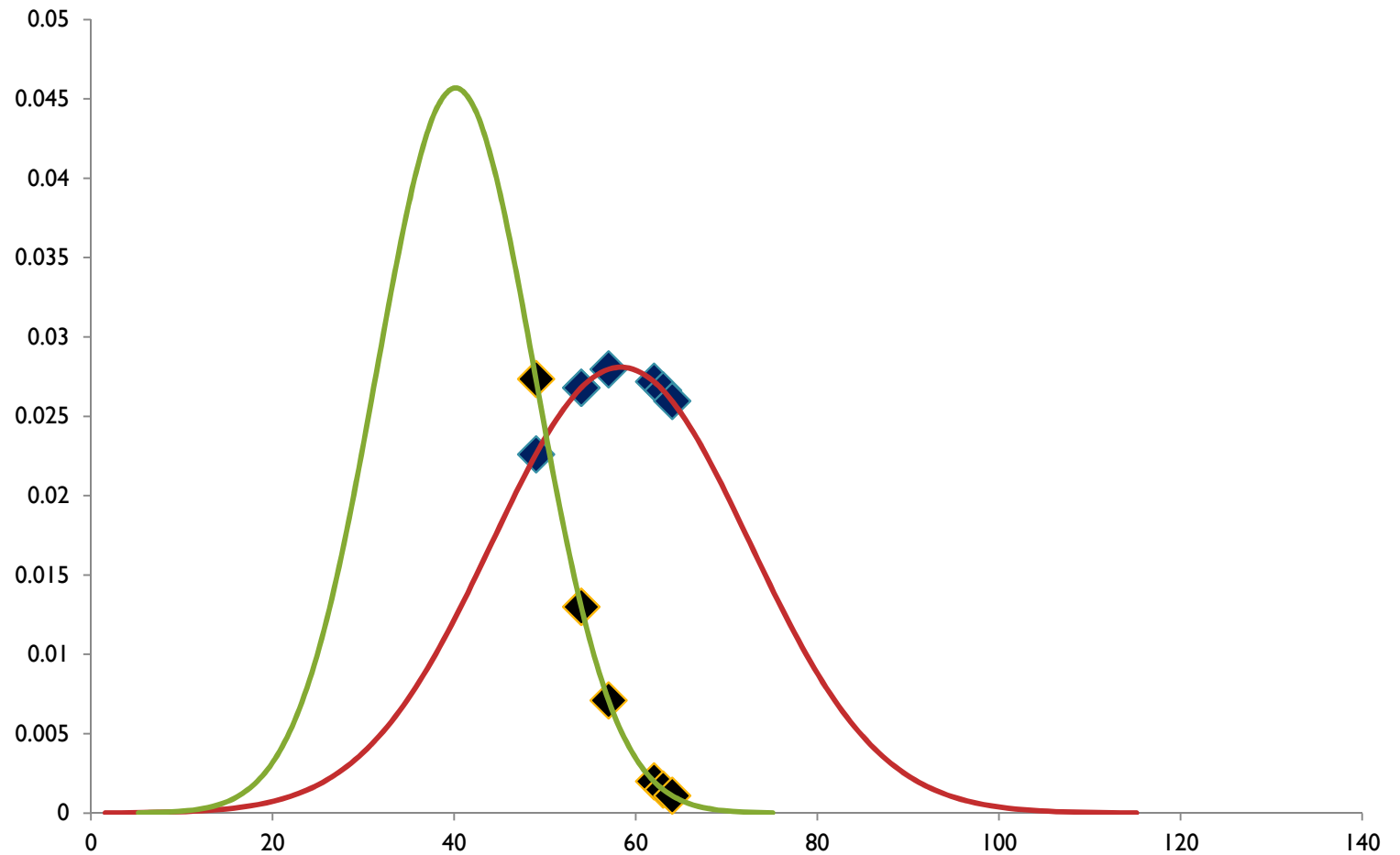


Figure 8 Extracted diacritics of 1 dots mapped on curves of font sizes 14 and 16

Contd...

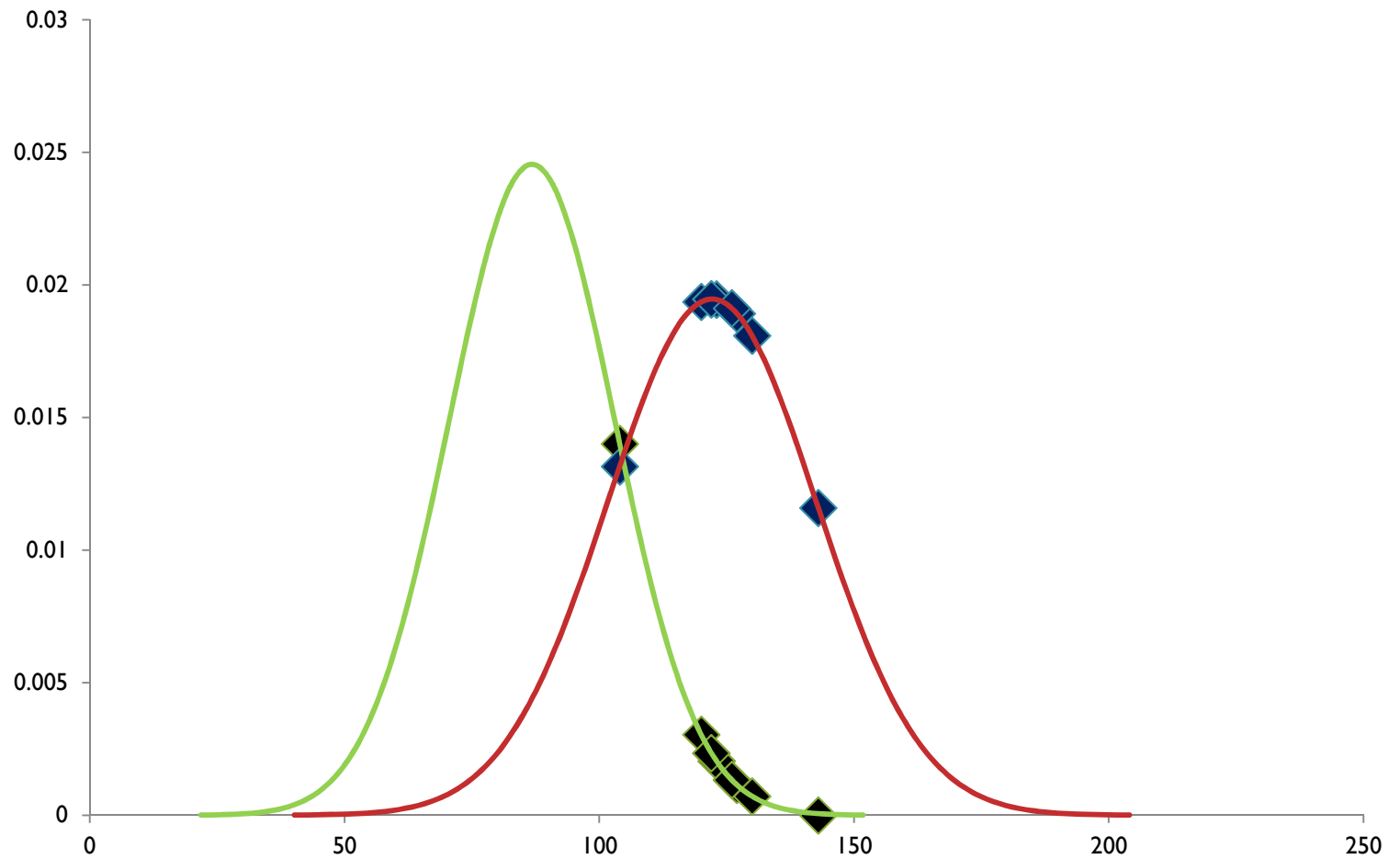


Figure 9 Extracted diacritics of 2 dots mapped on curves of font sizes 14 and 16

PROBABILITY GRIDS

Black Pixel Count	Font Size 14	Font Size 16
57	0.0071065	0.0279673
62	0.0019979	0.0271898
49	0.0273583	0.0226061
63	0.0014903	0.0266376
54	0.0129996	0.0268007
54	0.0129996	0.0268007
64	0.0010971	0.0259675

1 Dot Grid

Black Pixel Count	Font Size 14	Font Size 16
120	0.0030426	0.0193630
122	0.0023482	0.0194682
123	0.0020512	0.0194514
123	0.0020512	0.0194514
126	0.0013365	0.0191259
104	0.0140140	0.0131586
122	0.0023482	0.0194682
127	0.0011499	0.0189282
126	0.0013365	0.0191259
143	0.0000001	0.0115947
130	0.0007160	0.0180871

2 Dots Grid

RESULTS

(9) دائیں کریں کیس پر لگا ہوا آئل پمپ، کلچ اسمبلی اور دوسرے تمام پرزے علیحدہ کر لیں۔

Figure 10 The sample from image corpus of font size 16

Method	Detected Font Size	Actual Font Size
Normalization	16	16
Median	16	16
Mode	16	16

TESTING

Font Size	Total Lines	1 Dot Images	2 Dots Images
14	222	1,849	1,079
16	202	1,337	1,383

TESTING

	Font Size	Precision	Recall	Accuracy	f-measure
Normalization Method	14	92.99%	89.64%	91.04%	91.28%
	16	89.05%	92.57%	91.04%	90.78%
Median Method	14	91.28%	89.64%	90.09%	90.45%
	16	88.83%	90.59%	90.09%	89.71%
Mode Method	14	92.92%	88.74%	90.57%	90.78%
	16	88.21%	92.57%	90.57%	90.34%