

Ground Truth Data for Performance Evaluation of Urdu Nastalique OCR

Aneeta Niazi
Research Officer

Ground Truth Data

- **Definition:**

The term "ground truthing" refers to the process of gathering the proper objective data to prove or disprove research hypotheses.[1]

It serves as the highly representative reference data for continued research.[2]

For Optical Character Recognition, the characters of an image along with their aligned text constitute the ground truth data.

Applications

- Detailed performance evaluation of an OCR System.
- Accuracy comparison of different OCR techniques.
- Text to image mapping.
- Connected Component image extraction.
- Extraction of erroneous subsets of data for system analysis and improvement.

Properties of Ground Truth Data:

- The ground truth data must be at least one order of magnitude more accurate than the expected output of the system [3].
- A large amount of ground truth data has more significant impact on the overall success of an optical character recognizer [4].
- The ground truth data must be realistic and comprehensive [5].
- The ground truth data must be able to support an in-depth evaluation methodology for an OCR [5].
- The ground truth data set should also be flexibly structured, so that it can be easily searched for selecting subsets with different layout conditions, for more focused evaluation [5].

Existing Ground Truth Datasets

- A fast recursive text alignment scheme (RETAS) [6] has been used to align the ground truth e-texts, obtained from Project Gutenberg website with their corresponding OCR output. The OCR accuracy of real scanned 100 books in English and 20 books in French, German and Spanish respectively has been evaluated by using this approach.
- Sofia-Munich Corpus [7] has been reported for Eastern European languages.(text along with metadata)
- An automatic layout generation system for newspapers [8] has been used to generate synthetic ground truthed images.
- A recognition based ground truthing approach has been used for annotating Chinese handwritten document images, for text line segmentation, character segmentation and labeling [9].
- A database for handwritten Arabic script [10] has been presented, which contains ground truth information for 26459 Tunisian town/village names, written by 411 writers.(metadata and text)

Existing Ground Truth Datasets for Urdu

- The development of ground truth data has been carried out for a handwritten Urdu database [11] containing isolated digits, numerical strings with/without decimal points, 5 special symbols, 44 isolated characters, 57 Urdu words and Urdu dates in different patterns.(includes metadata information only).
- An Urdu handwritten sentence database[12] has been developed, with line level ground truth data for 400 handwritten forms, written by 200 different writers and contains 23833 printed Urdu words in 2051 lines of text.(line level coordinates information only).

Complexities of Nastalique Writing Style

ہر شخص پر اس کی استعداد کے مطابق جدا گانہ تجلیات ہوتی ہیں

Vertical Overlapping between ligatures

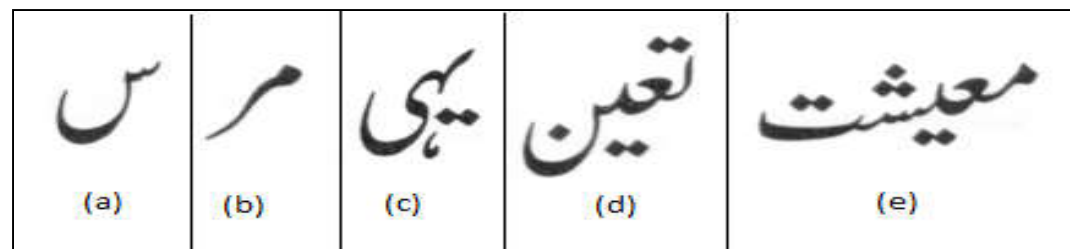


(a) Character shaping of ب class in Naskh writing style (b) Contextual character shaping of ب class in Nastalique writing style.

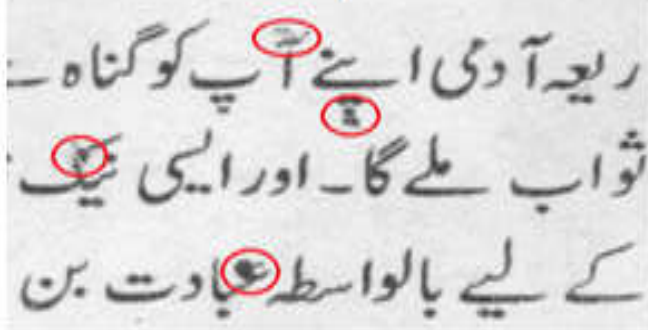
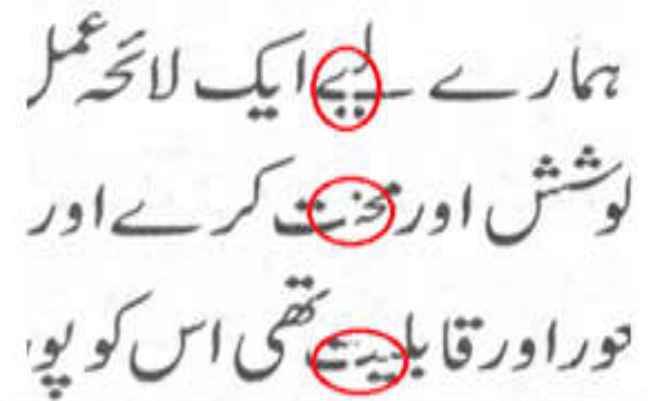
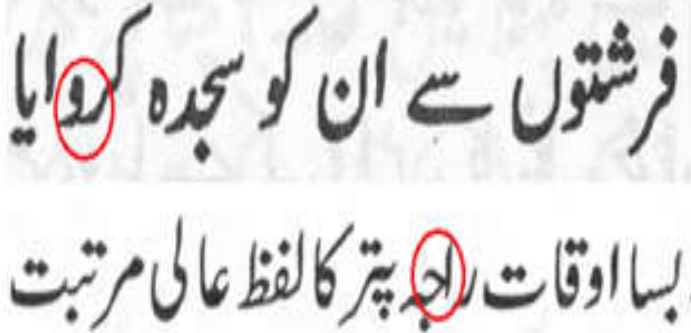
A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

Ligature	Main body Shape	Confusing Diacritic	Diacritic Shape (enlarged)
ر	ر	Diacritic of نی	ر
با	با	Diacritic of ہم	با
ط	ط	Diacritic of ٹ	ط

Diacritics and main bodies confusion

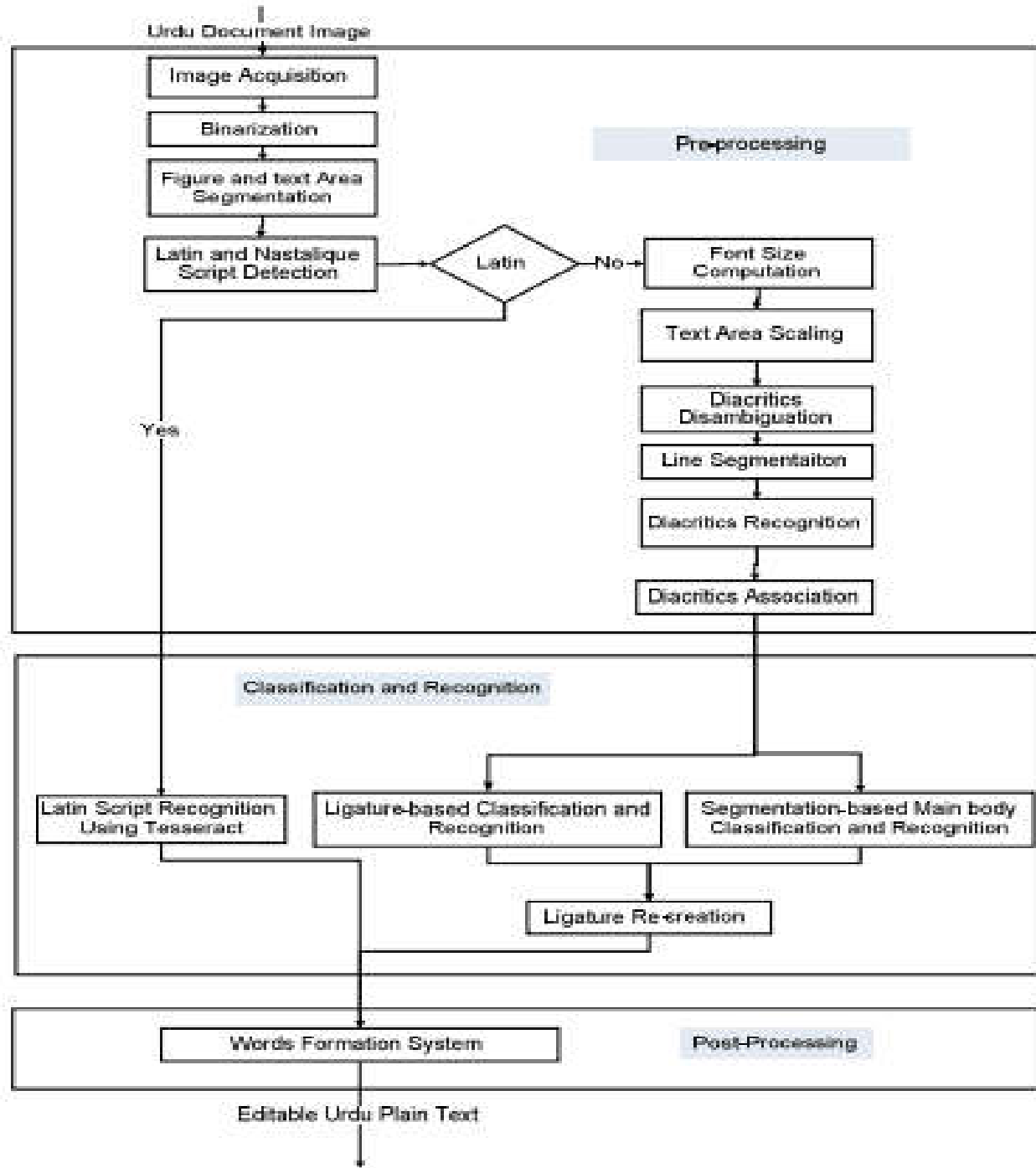


Thick-thin stroke variation across characters in ligatures having (a) one character (b) two characters (c) three characters (d) four characters (e) five characters.

Special Cases	Examples
Connected Components with Attached Noise	 <p>ربیعہ آدمی اپنے آپ کو گناہ۔ ثواب ملے گا۔ اور ایسی بیک کے لیے بالواسطہ عبادت بن</p>
Broken Ligatures	 <p>ہمارے لیے ایک لائحہ عمل لوشش اور محنت کرے اور نور اور قابلیت تھی اس کو پورا</p>
Joined Ligatures	 <p>فرشتوں سے ان کو سجدہ کروایا بسا اوقات راجہ پتر کا لفظ عالی مرتبت</p>

Portions of text encircled with red color indicating special cases found in real Urdu Nastalīq. The document images are for research purposes only. The quality of the document images is due to poor printing.

Nastalique OCR



آئے ہوئے تمام لوگوں کا شکریہ ادا کرتا ہے

OCR

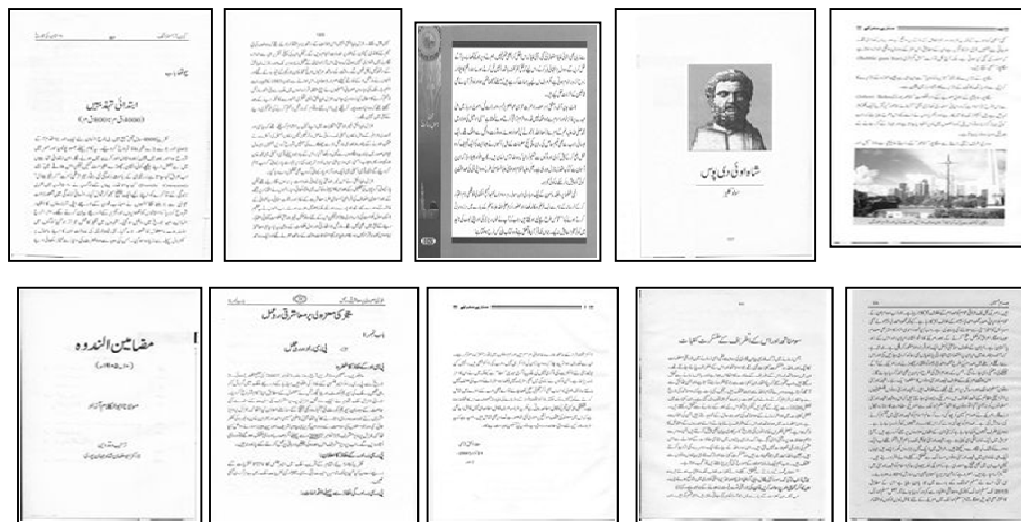
آئے ہوئے تمام لوگوں کا شکریہ ادا کرتا ہے



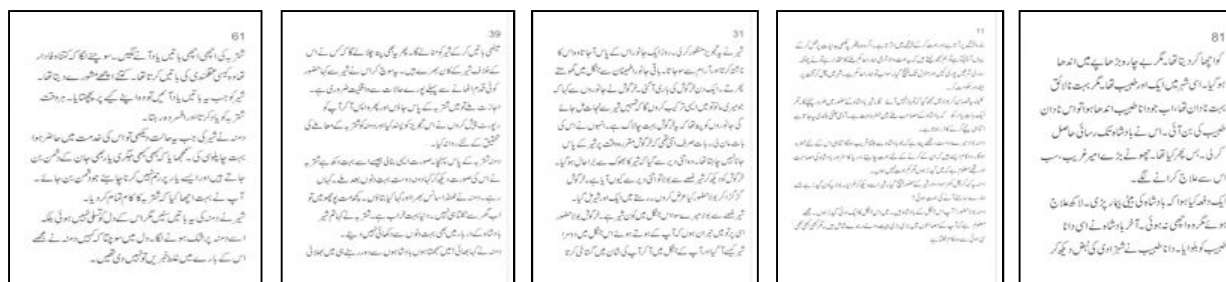
A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

Methodology

- Data Collection



Scanned document image collected from books



Synthesized Document Images (for 26 30 34 38 42 and 44 font sizes)

Methodology

- **Naming Convention:**

The naming of scanned images has been done in such a way that their meta data information i.e. book identifier, page number and font size of the printed text, can be obtained from the image name.

G(Grayscale)_E(Edited)_C(Cropped)_B<Book ID>_P<Page Number>_F.jpg

Typed Text Files:

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

- For each scanned image, a typed text file has been prepared, which contains typed text of the corresponding scanned image.
- The typed text file is in UTF-8 .txt format, which is an open format and can be easily accessed on different platforms.
- Each typed file has been assigned the same name as that of its corresponding scanned image.



141 لنکا ڈھانے

یتیم خوش تھے

بڑا آدمی یتیم ہو جائے تو یتیم نہیں لگتا۔ یہ اصطلاح غریب کے لئے مخصوص ہے۔

اس کا باپ زندہ ہو تب بھی یتیم کہلاتا ہے۔

تیسری دنیا کے عوام کو یتیم کہا جاتا ہے۔ اس لیے امریکہ ان کے سر پر ہاتھ رکھتا ہے۔ لاکھ شور مچایا جائے، کوئی یتیم ہوگا تو کسی نہ کسی کو سر پر ہاتھ رکھنا ہوگا۔

ترقی پذیر ملکوں کے سربراہ اصرار کر کے امریکہ کا ہاتھ اپنے سر پر رکھواتے ہیں۔

اتنے عادی ہو جاتے ہیں سال میں ایک بار امریکہ جا کر اپنا سوائٹ ہاؤس کی چوکھٹ پر

ضرور لگاتے ہیں۔ انسان اور جانور کے درمیان روزِ اوّل سے تعلق ہوتی ہے۔ کون کسے

یتیم بنائے۔ انسان جانور کے بچوں کو یتیم کرتا ہے۔ تو اس پر اپنا پاؤں رکھ کر تصویر کھنچواتا

ہے۔ جانور ایسے موقعوں پر تصویر کھنچوانے کے تکلف کے بجائے پیٹ میں ڈال لیتا ہے۔

جانوروں میں یتیم ہونے کا رواج نہیں۔ بعض جانور ایسے ہوتے ہیں جن کے لئے



یتیم کا لفظ استعمال کیا ہی نہیں جاسکتا۔ کبھی نہیں سنا کوئی شیر یتیم ہو گیا۔ گینڈے کے سر

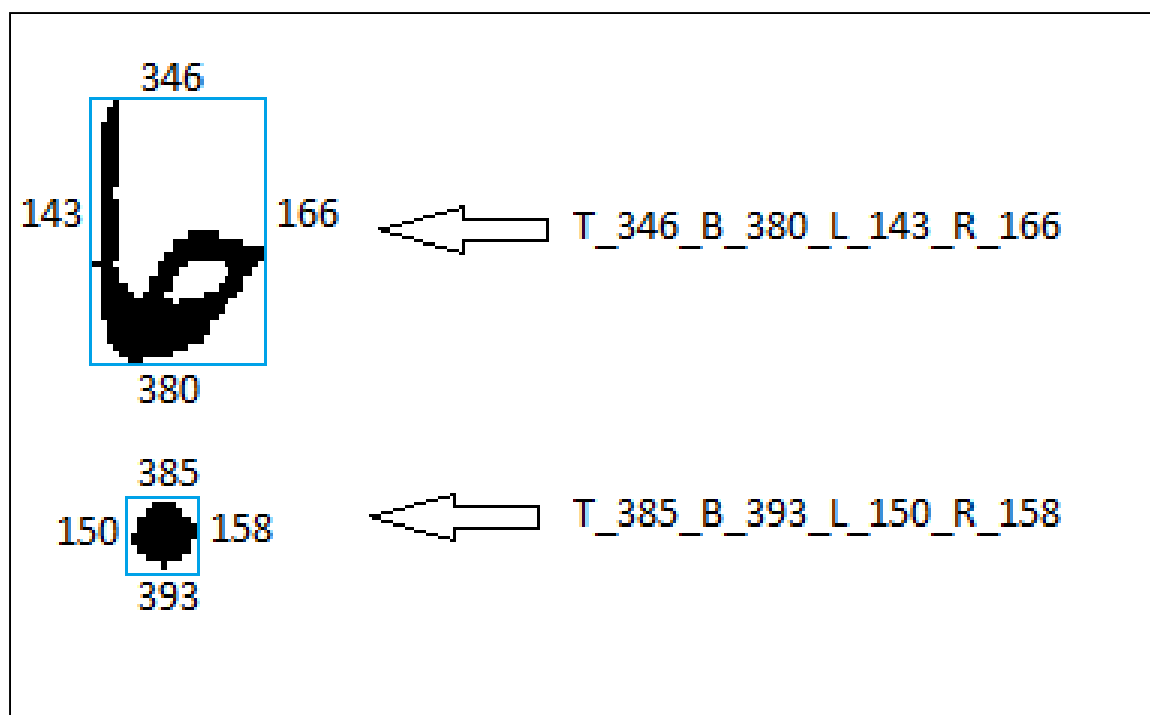
سے ماں باپ کا سایہ اٹھ گیا۔ چیتا لاوارث ہو گیا۔ ان جانوروں کے سامنے بہادر سورما

for I

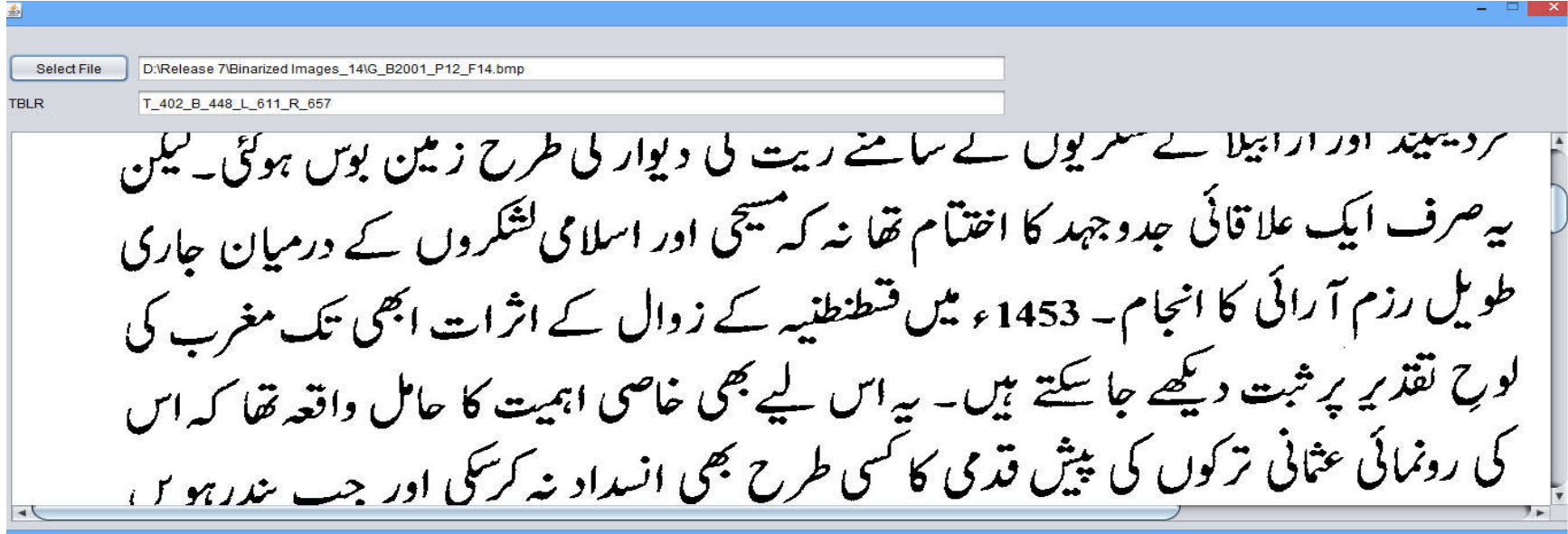
A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

Ground Truth File Format:

Line Number	Ligature Number	Font Size	Ligature TBLR	Base Ligature	MBID	Recognizer ID	Diacritics TBLR	Diacritics Sequence	Ligature ID	Ligature	Error Code
1	31	F14	T_1366_B_1 415_L_1283 _R_1345		4775	1	T_1359_B_1 366_L_1319 _R_1326	1001	643		11



Verification



Utility for Automatic TBLR Extraction

”یہ زندگی کا اصول بنا لیں کہ آپ کے ہاتھوں سے جو کام بھی ہوا چھا
ہو۔ کام پر آپ کی چھاپ ہونی چاہیے، یہی کام آپ کی عظمت کا نشان
ہوگا۔“

Special Cases

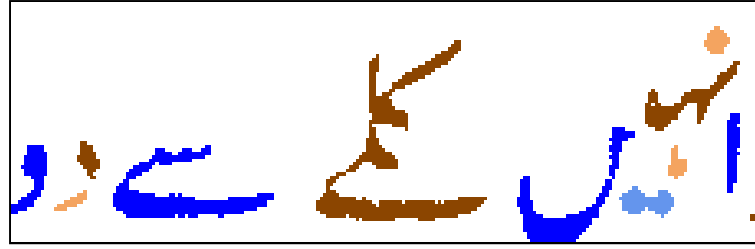
- **Broken Connected Components:**
 - i. Broken Main Body
 - ii. Broken Diacritics
- **Joined Connected Components:**
 - i. Joined Main Bodies
 - ii. Main Bodies Joined with Diacritics
 - iii. Main Bodies Joined with Incorrect Diacritics
 - iv. Joined Diacritics
- **Special Symbols**
- **Noise Attached with Connected Components**

Broken Connected Components

Broken Main Body:

1. Get TBLR of the bounding box containing all pieces of complete main body stroke from TBLR Extractor utility.
2. Write the desired ligature string in the respective column.
3. Enter the tag, "**Broken_MB**" in the respective column.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



Distorted shape of گلے due to broken main body.



The main body of ل has two colors instead of one color in color coded image, indicating that it has a broken main body.



The broken piece of سکھا is associated with its main body

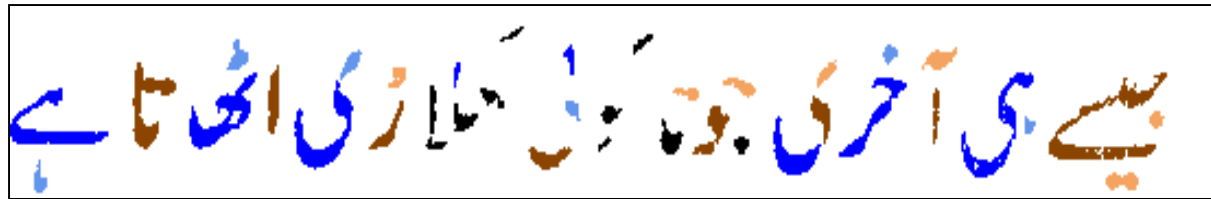
A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



The broken piece of گو is associated with the main body of لو as a diacritic.



The pieces of the broken main body of تا are marked as noise (in black color).



The shape information is almost lost due to poor printing quality for the main bodies of ٹھا, کھلا, کو, ئی, تا and جو.

Broken Connected Components

- **Broken Diacritics:**

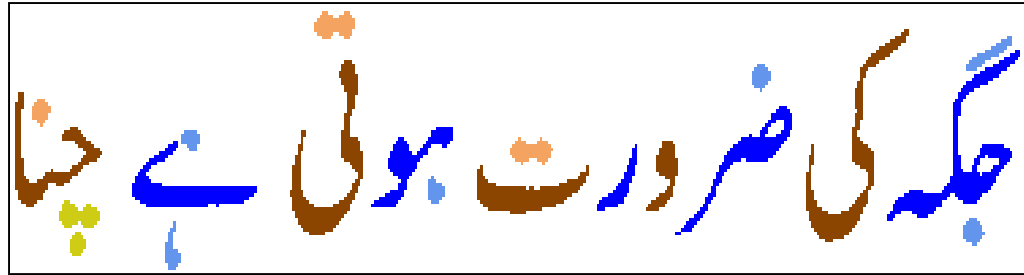
1. Get TBLR of the bounding box containing all pieces of complete diacritic stroke from the TBLR Extractor Utility.
2. Write the desired diacritic identifier in the respective column.
3. Enter the tag, "**Broken_Dia**" in the respective column.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



اپنے اپنے دائیں ہاتھ کی

The broken diacritic piece of ئیں is marked as noise due to small size (in black color).



جگہ کی ضرورت ہوتی ہے چنا

The broken diacritic of ہو gets incorrectly recognized as one dot due to shape similarity

Joined Connected Components

- **Joined Main Bodies:**

1. Get TBLR of the bounding box containing joined main bodies from the TBLR Extractor Utility.
2. Write the ligature strings of all joined main bodies in the respective column.
3. Enter the tag, "**Joined_MB_MB**" in the respective column.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



عام انسانوں کی سطح پر لے آتا ہے۔ اسی وجہ سے

Joined main bodies of و and جہ are incorrectly marked as a single main body (brown color instead of blue and brown).



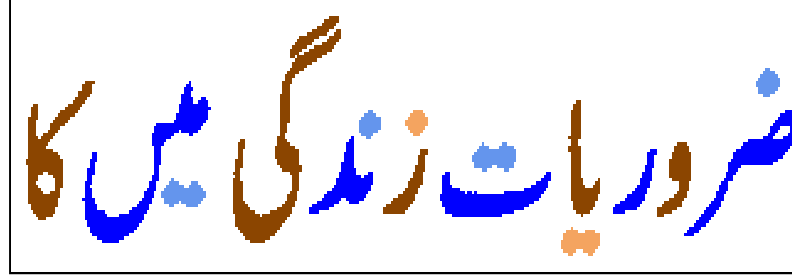
جب دوسری جنگ عظیم شروع ہوئی تو وہ
بعد انگریز افسران کی کارکردگی سے اتے

Joined main bodies of شر and گی in different lines of a document image, incorrectly marked as noise (black in color).

Joined Connected Components

- **Main Body with Joined Diacritics:**
 1. Get TBLR of the bounding box containing the complete stroke of the main body with joined diacritics from the TBLR Extractor Utility.
 2. Write the ligature string of the ligature having joined diacritics in the respective column.
 3. Enter the tag, "**Joined_MB_Dia**" in the respective column.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



The main body of گے is joined with its diacritic (14 font size).



The main body of ہا has a joined diacritic in the synthesized image of a larger font size (30 font size), indicating the property of Nastalique

Joined Connected Components

- **Main Body Joined with Incorrect Diacritics:**
 1. Get TBLR of the bounding box containing the complete joined stroke of the main body with incorrect diacritics from the TBLR Extractor Utility.
 2. Write the ligature string of the ligature having incorrect joined diacritics in the respective column.
 3. Enter the tag, "**Joined_MB_IncorrectDia**" in the respective column of the ligature entry having incorrect joined diacritics.
 4. Write the ligature string of the ligature having incomplete number of diacritics in the respective column.
 5. Enter the tag, "**Joined_MB_IncorrectDia**" in the respective column of the ligature entry having incomplete number of diacritics.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



The diacritic of بی is joined with the main body of مغر, making بی an invalid ligature, and distorting the main body shape of مغر.

Joined Connected Components

- **Joined Diacritics:**

1. Get TBLR of the bounding box containing the complete stroke of the joined diacritics from the TBLR Extractor Utility.
2. Write diacritic identifiers of all diacritics, separated by "_" (e.g. One Dot_Two Dots), in the respective column.
3. Enter the tag, "**Joined_Dia_Dia**" in the respective column.



The joined diacritics of منظم are incorrectly marked as noise.

Special Symbols

- Latin Script Main Bodies.

رہنے لگے مگر بعد کو لوٹ ڈرائل (Lavendar Hill)

- Connected Components of other writing styles of Urdu.

حیات: وقت کی طاقت

- Arabic Connected Components.

ادْعُونِي أَسْتَجِبْ لَكُمْ)۔ دوسری جگہ فرمایا

- Bullets and numbering etc.

☆ ایک مشرقی۔
خطرہ دور ہو چکا تھا۔ (LE)
5th May, 2

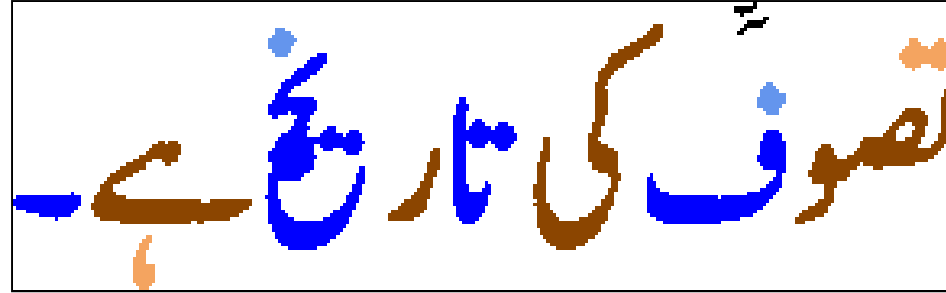
Special Symbols

1. Get TBLR of the bounding box containing the complete stroke of the special symbol from the TBLR Extractor Utility.
2. Write the ligature string of the special symbol in the respective column. If the ligature string of the symbol cannot be typed from key board, write "Symbol" in the respective ligature string column.
3. Enter the tag, "**Special_Symbol**" in the respective column.

Noise attached with Connected Components

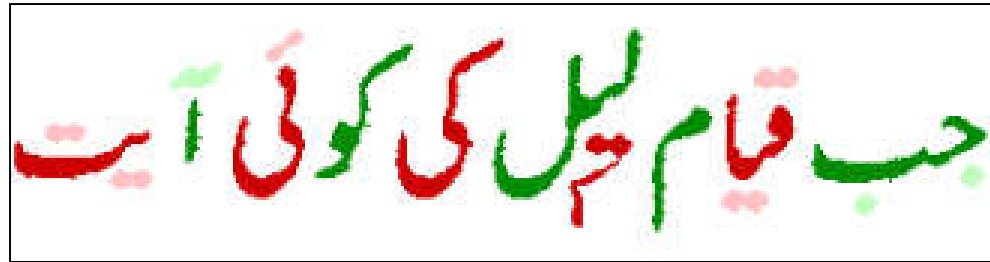
1. Get TBLR of the bounding box containing the main body/diacritic with attached noise from the TBLR Extractor Utility.
2. Write the ligature string of the main body/diacritic identifier in the respective column.
3. Enter the tag, "**Noise_Attached**" in the respective column.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR



تصوف کی تاریخ ہے

Noise attached with the main body of یخ.



















































جب قیام لیل کی کوئی آیت

Noise attached with the diacritic of لیل.

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

2nd Verification Pass

 G_E_C_B2010_P34 5_F14_T_586_B_61 9_L_525_R_536	 G_E_C_B2010_P34 5_F14_T_862_B_89 6_L_925_R_935	 G_E_C_B2010_P34 5_F14_T_1047_B_1 080_L_1139_R_115	 G_E_C_B2010_P34 5_F14_T_1140_B_1 173_L_825_R_836	 G_E_C_B2010_P34 5_F14_T_1507_B_1 540_L_1536_R_154	 G_E_C_B2012_P9 F14_F24_T_939_B_1 975_L_672_R_684	 G_E_C_B2012_P9 F14_F24_T_1041_B_1 1077_L_360_R_3	 G_E_C_B2012_P9 F14_F24_T_1527_B_1 1563_L_1321_R_1	 G_E_C_B2012_P9 F14_F24_T_1628_B_1 1665_L_1246_R_1	 G_E_C_B2012_P9 F14_F24_T_1930_B_1 1967_L_1174_R_1	 G_E_C_B2012_P9 F14_F24_T_1950_B_1 1972_L_328_R_3	 G_E_C_B2012_P9 F14_F24_T_2028_B_1 2064_L_1404_R_1
		0		6		73	1333	1259	1187	41	1416
 G_E_C_B2012_P9 F14_F24_T_2036_B_1 2071_L_414_R_4	 G_E_C_B2012_P10 F14_F18_T_262_B_1 298_L_366_R_377	 G_E_C_B2012_P10 F14_F18_T_267_B_1 302_L_888_R_900	 G_E_C_B2012_P10 F14_F18_T_562_B_1 597_L_758_R_769	 G_E_C_B2012_P10 F14_F18_T_566_B_1 602_L_1217_R_12	 G_E_C_B2012_P10 F14_F18_T_664_B_1 699_L_1006_R_10	 G_E_C_B2012_P10 F14_F18_T_763_B_1 798_L_1112_R_11	 G_E_C_B2012_P10 F14_F18_T_859_B_1 894_L_802_R_814	 G_E_C_B2012_P10 F14_F18_T_957_B_1 993_L_657_R_669	 G_E_C_B2012_P13 9_F14_T_297_B_33 3_L_242_R_254	 G_E_C_B2012_P13 9_F14_T_1009_B_1 045_L_443_R_456	 G_E_C_B2012_P13 9_F14_T_1459_B_1 495_L_954_R_966
27				29	18	24					
 G_E_C_B2012_P13 9_F14_T_1547_B_1 584_L_541_R_552	 G_E_C_B2014_P22 F14_F16_T_306_B_1 338_L_266_R_277	 G_E_C_B2014_P22 F14_F16_T_307_B_1 340_L_389_R_399	 G_E_C_B2014_P22 F14_F16_T_396_B_1 429_L_163_R_174	 G_E_C_B2014_P22 F14_F16_T_400_B_1 433_L_620_R_630	 G_E_C_B2014_P22 F14_F16_T_402_B_1 434_L_715_R_725	 G_E_C_B2014_P22 F14_F16_T_490_B_1 523_L_560_R_571	 G_E_C_B2014_P22 F14_F16_T_763_B_1 794_L_413_R_424	 G_E_C_B2014_P22 F14_F16_T_948_B_1 981_L_681_R_691	 G_E_C_B2014_P22 F14_F16_T_953_B_1 986_L_1093_R_11	 G_E_C_B2014_P22 F14_F16_T_954_B_1 988_L_1222_R_12	 G_E_C_B2014_P22 F14_F16_T_1127_B_1 1159_L_240_R_2
									03	33	51
 G_E_C_B2014_P22 F14_F16_T_1130_B_1 1161_L_507_R_5	 G_E_C_B2014_P22 F14_F16_T_1137_B_1 1170_L_1212_R_1	 G_E_C_B2014_P22 F14_F16_T_1312_B_1 1344_L_349_R_3	 G_E_C_B2014_P22 F14_F16_T_1315_B_1 1346_L_495_R_5	 G_E_C_B2014_P22 F14_F16_T_1407_B_1 1440_L_792_R_8	 G_E_C_B2014_P22 F14_F16_T_1687_B_1 1719_L_1212_R_1	 G_E_C_B2014_P21 8_F14_F16_T_413_B_1 444_L_1350_R_1	 G_E_C_B2014_P21 8_F14_F16_T_626_B_1 658_L_1105_R_1	 G_E_C_B2014_P21 8_F14_F16_T_628_B_1 660_L_759_R_77	 G_E_C_B2014_P21 8_F14_F16_T_629_B_1 661_L_473_R_48	 G_E_C_B2014_P21 8_F14_F16_T_1084_B_1 1115_L_430_R_1	 G_E_C_B2014_P21 8_F14_F16_T_1174_B_1 1205_L_587_R_1
18	1223	57	06	03	1223	360	115	0	4	440	598

A folder for ۛ class, containing an instance image of ۛ, indicating a tagging error.

Example:

A Ground Truth Dataset for Performance Evaluation of Urdu Nastalique OCR

وطنیت کے مغربی تصور

Line Number	Ligature Number	Font Size	Ligature TBLR	Base Ligature	MBID	Recognizer ID	Diacritics TBLR	Diacritics Sequence	Ligature ID	Ligature	Error Code
1	34	F14	T_1378_B_1 398_L_1481 _R_1492	و	5189	42			3	و	
1	33	F14	T_1355_B_1 399_L_1411 _R_1482	طیب	2911	1	T_1369_B_137 4_L_1459_R_1 465 T_1398_B_140 4_L_1444_R_1 457 T_1382_B_138 8_L_1422_R_1 436	1001 2002 1002	4093	طنیت	
1	32	F14	T_1353_B_1 399_L_1348 _R_1393	کے	3868	1			7	کے	
1	31	F14	T_1366_B_1 415_L_1283 _R_1345	معر	4775	1	T_1359_B_136 6_L_1319_R_1 326	1001	643	مغر	11
1	30	F14	T_1356_B_1 406_L_1269 _R_1293	لی	4306	1			113	بی	11
1	29	F14	T_1370_B_1 399_L_1217 _R_1257	بصو	1241	1	T_1359_B_136 5_L_1252_R_1 267 T_1368_B_137 6_L_1218_R_1 227	1002 1005	486	تصو	
1	28	F14	T_1380_B_1 399_L_1192 _R_1206	ر	2317	4			2	ر	

Data Counts

Number of Pages	490
Number of Books	176
Authors	151
Domains	19
Publishers	80

References

1. http://en.wikipedia.org/wiki/Ground_truth
2. **Muhlberger, Gunter.** *TranScriptorium D2.1: Data Collection and Ground Truth Annotation.* s.l. : ICT Project 600707, funded by European Community, 2013.
3. *Ground Truth Design Principles.* **Kondermann, Daniel.** Petersburg, Russia : s.n., 2013. International Workshop on Video and Image Ground Truth in Computer Vision Applications.
4. **Muhlberger, Gunter.** *TranScriptorium D2.1: Data Collection and Ground Truth Annotation.* s.l. : ICT Project 600707, funded by European Community, 2013.
5. *A Realistic Dataset for Performance Evaluation of Document Layout Analysis.* **A. Antonacopoulos, D. Bridson, C. Papadopoulos, S. Pletschacher.** Barcelona, Spain : s.n., 2009. 10th International Conference on Document Analysis Recognition. pp. 296-300.
6. *A Fast Alignment Scheme for Automatic OCR Evaluation of Books.* **Ismet Zeki Yalniz, R. Manmatha.** Beijing, China : s.n., 2011. 11th International Conference on Document Analysis and Recognition. pp. 754-758. *A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques.* **Stoyan Mihov, Klaus U. Schulz, Christoph Ringlstetter, Veselka Dojchinova, Vanja Nakova, Kristina Kalpakchieva, Ognjan Gerasimov, Annette Gotscharek, Claudia Gercke.** Seoul, Korea : s.n., 2005. 8th International Conference on Document Analysis and Recognition. pp. 162-166.
7. *A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques.* **Stoyan Mihov, Klaus U. Schulz, Christoph Ringlstetter, Veselka Dojchinova, Vanja Nakova, Kristina Kalpakchieva, Ognjan Gerasimov, Annette Gotscharek, Claudia Gercke.** Seoul, Korea : s.n., 2005. 8th International Conference on Document Analysis and Recognition. pp. 162-166.
8. *Automated Ground Truth Data Generation for Newspaper Document Images.* **Thomas Strecker, Joost van Beusekom, Sahin Albayrak, Thomas M. Breuel.** Barcelona, Spain : s.n., 2009. 10th International Conference on Document Analysis and Recognition. pp. 1275-1279.

References

9. *A Tool for Ground Truthing Text Lines and Characters in Offline Handwritten Chinese Documents.* **Fei Yim, Qiu-Feng Wang, Cheng-Lin Liu.** Barcelona, Spain : s.n., 2009. 10th International Conference on Document Analysis and Recognition. pp. 951-955.
10. *IFN/ENIT-Database of Handwritten Arabic Words.* **Pechwitz, Samia Snoussi Maddouri, Volker Margner, Nouredine Ellouze, Hamid Amiri.** Hammamet, Tunis : s.n., 2002. 7th Colloque International Francophone sur l'Ecrit et le Document. pp. 127-136.
11. *A New Large Urdu Database for Off-Line Handwriting Recognition.* **Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, Ching Y. Suen.** Vietri sul Mare, Italy : s.n., 2009. 15th International Conference on Image Analysis and Processing. pp. 538-546.
12. *. An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation.* **Ahsen Raza, Imran Siddiqi, Ali Abidi, Fahim Arif.** Bari, Italy : s.n., 2012. 13th International Conference on Frontiers in Handwriting Recognition. pp. 491-496.

Thank You