

Maturation Process of the Ligature Based Urdu Noori Nastalique Optical Character Recognizer

Presenter :Aneeta Niazi

What is Optical Character Recognition?

آئے ہوئے تمام لوگوں کا شکریہ ادا کرتا ہے



OCR

آئے ہوئے تمام لوگوں کا شکریہ ادا کرتا ہے

Ligature Based Recognizer

کھر ساہ کھانے طو کبو کم فاصد ح سلا حلا مد لو

Ligature Strings

کھر ساہ کھانے طو کو کم فاصد ح سلا حلا مد لو

Main bodies of ligatures

Training and Testing Data Division

Training and Testing data have been prepared for **5586** High Frequency Main body Classes.

- **For Training:** 35 tokens for each MB Class.
- **For Testing:** 15 tokens for each MB Class.

Training of MB Classes is done by using **Tesseract**, an open source multilingual OCR System.

Tesseract returns a list of best choices for each Main Body after recognition. If a Main Body exist in this ranked list of choices, it is considered correctly recognized.

Maturation Process of the Ligature Based Urdu Noori Nastalique Optical Character Recognizer

Superset of 5586 Main Bodies

Main Body Tokens
of a class

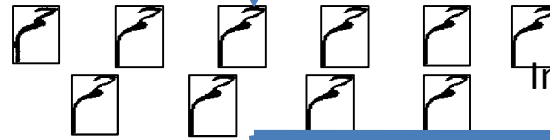
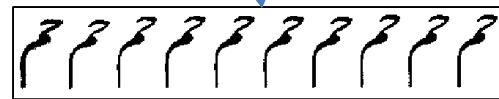


Image Creation Utility

.tiff Image



from command prompt

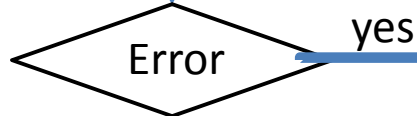
.box file (contains coordinates
of each Main Body's bounding
box)

```
A02155 3 10 45 82 0  
A02155 46 10 86 81 0  
A02155 87 10 128 85 0  
A02155 129 10 171 86 0  
A02155 172 10 214 86 0
```

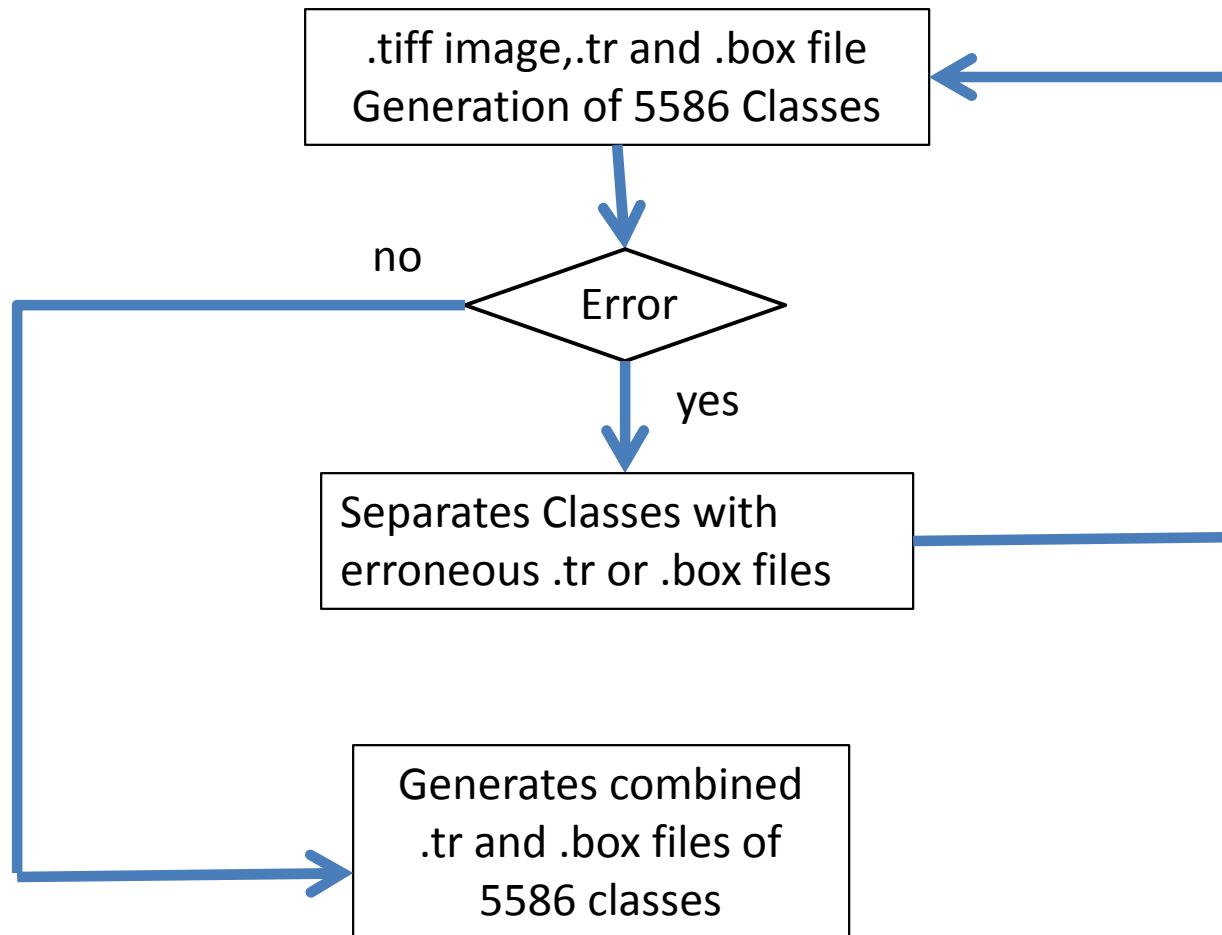
from command prompt

tr file (contains outline
features, size and
position information)

```
nas A02155  
2mf 20  
0.1923 0.0688866 0.0642724 0.609651 0 0  
0.183922 0.115051 0.0839869 0.895019 0 0  
cn 1  
0.429688 0.205078 0.207031 0.117188
```



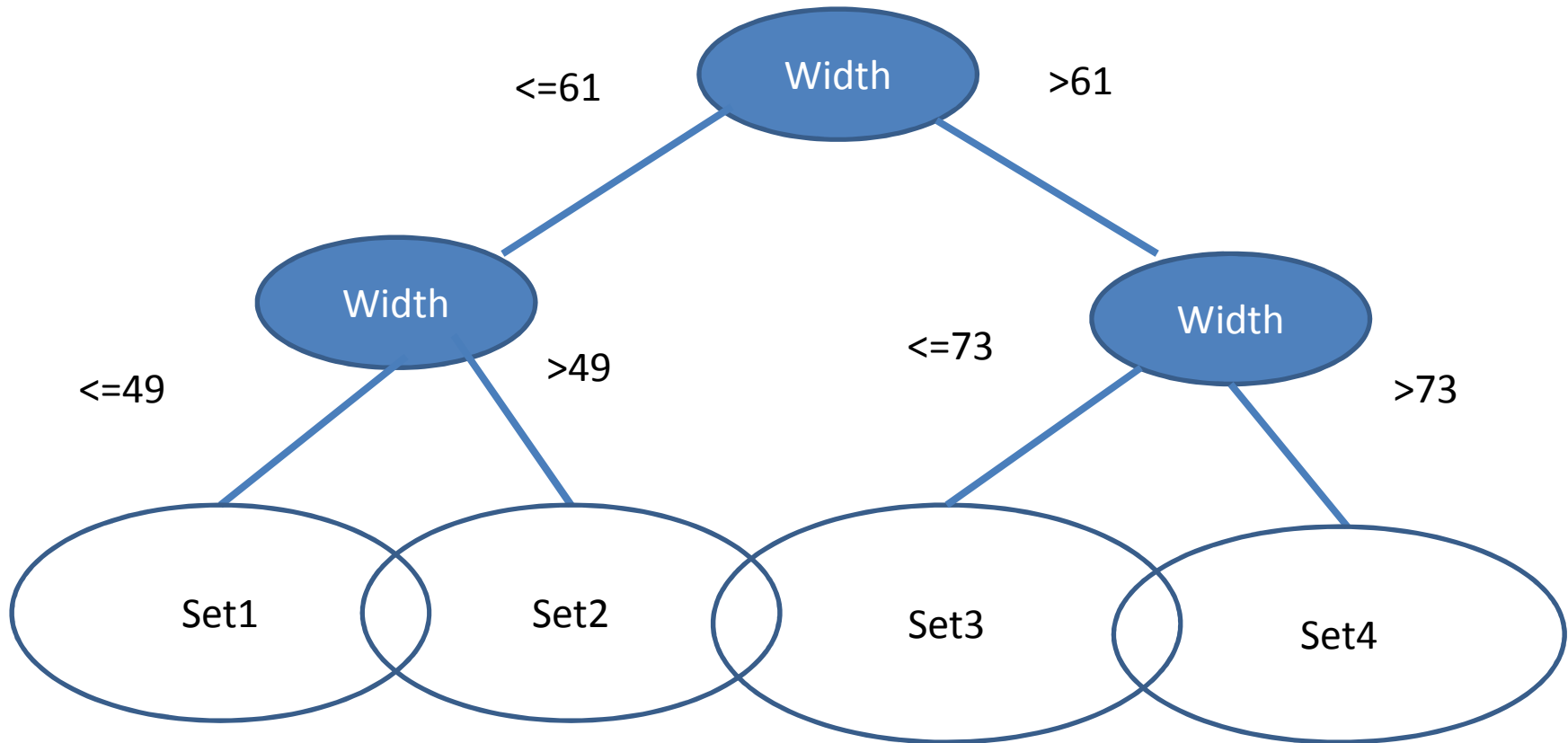
Automatic Generation of Training Files



Previously Used Testing Images:

سعر معلوم ۸ جا مہما فر طا طلا لک لکھا سک کھلے سکر حلو مسعن حیا صوو طمنا جمع لب مطر سکر مسا سکل عس سک مطر
 مٹلا فطر مسعل ۴ سا ۶ سو سک عس نع صظلا معا لہ عمد نکہ نہ محض حسب حر محر لک حسب ہر ہا سظا لک مسلم ملہ
 طمنا حیا عظم لعا سمب مظا سا معلو لکل عمد مل ۵ عس لس لکھر سد مسعل کس مد عمد سکل لہا سو محض ملا مطو
 ۷ حسو ہلا لعا لکھو جا ۶ فر لہر لہا لعا حے حل ۶ مسلم لہا مجموعہ ہلا ملانا طہر ود لہہ کسا لہ سظا کے لک محر مکا لعا
 حصف عسا ظلا مسد ح ۷ کھا صح صوو ک کھا لہہ لعا لفظ مطر ۸ عظم سحر ۴ معصد ملت رکا لہر مسکر مسد صلے
 مل سجد لب لعا لکھا حیا لہا لہ سلط مغل مسد رکا مہ رکا لک فر سک ہلو صوو کے محلے لکھ سحر مطر ۹ عد مطلب
 لعا مسعن فلسفہ سک ملکو مو سو کہا مطر ملک مسکر معلو کس حے سن لعا صلے سکے صل محصر محض حلا لکا حملہ سعا صوو
 رکا حلو محو محمد ۱۰ سہا سو لعا حیا طہ ۵ لعا کلا لصل ۱۱ سعال لب لجا مہ عمر لعا مسعن مجموعہ سو کھ لعا ح مطو حسب
 سک جمع نو صوو کھک لک صظلا لکل سظا حسو سکر کد سو کھلا کھلا عمو لکا سا لعا لک فکر لعا لہ کھلا لعا سنع عا سجد
 عس کد لعا ک لب معمو کھلا سحر سمب ملت سک لکا لہو عمو حصف لظم ک ہر سحر حلو معلو معا سس لعا عظم

Sets Division



Sigma Computation for Overlapping Sets

- The value of Sigma is computed by taking the Standard Deviation of the real data of each MB Class, and then taking the average of all Standard Deviations.
- For overlapping sets, the value of 2* sigma is used.

Font	Sigma	2*Sigma
F14	1.820	3.640
F16	1.656	3.313
F22	2.440	4.881
F36	1.109	2.218

Set Division Thresholds

	F14	F16	F22	F36
Threshold between Set1 and Set2	49	59	82	127
Threshold between Set2 and Set3	61	73	99	156
Threshold between Set3 and Set4	73	88	120	190

Testing Images after Sets Division

سماے مہہ سیا حائے سے سماے سر ر سے سیا سے سیا سے سیا سے سیا سے سیا سے سیا سے
حوتیے ماسرے موے سو سیا سے کہ سیاے ر مہہ ب ماب سے
ب ماس سماے سے لولا ماسے ب ماس سے سیا سے سیا

Set1

کے کے کہ لکاح مس کے سیا حمر مس مس لے م مک

Set2

کوئی با ہم کر کہ
ہم کو بھی کر کی کر
ہم کو کو کہ نظر

Set3

سی سی سی سی سی سی سی سی سی

Set4

F14 Overall Accuracy with a Single Trained data File	93.69323
F14 Overall Accuracy with 4 Trained data Files	94.65523

Addition of Scaled Data to the recognizers of 22 and 36 font sizes



Font	Sigma	2*Sigma
F22-Pivot (F18-F28)	5.429	10.858
F36-Pivot (F30-F44)	6.747	13.493

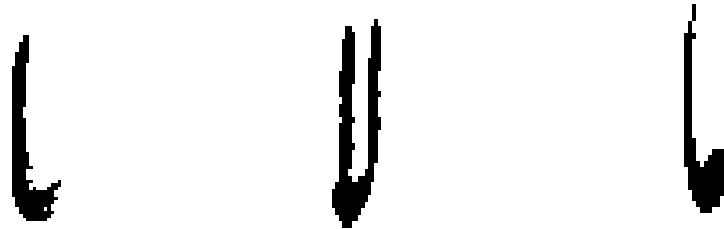
	F22-Pivot	F36-Pivot
Threshold between Set1 and Set2	76	122
Threshold between Set2 and Set3	94	150
Threshold between Set3 and Set4	117	186

Alif Recognition

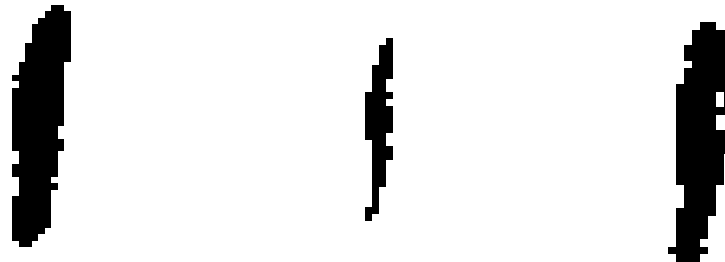
- Alif was not being trained by Tesseract.
- Alif has been recognized on the basis of height and width thresholds, as it has a unique shape.

	F14	F16	F22	F36
Alif's Mean Height	29	32	47	44
Alif's Mean Width	6	6	9	8
Alif's Height S.D.	5	7	6	4
Alif's Width S.D	3	2	2	2

- Testing on document pages from Urdu books showed that some Main Bodies were being misrecognized as Alif.



- Some Alifs were also being misrecognized



- Alif Thresholds have been updated

	F14	F16	F22	F36
Alif's Mean Height	29	32	47	72
Alif's Mean Width	6	6	9	12
Alif's Height S.D.	7	7	15	13
Alif's Width S.D	4	4	5	6
Alif's minimum Width	2	3	3	-

- Decision trees have been implemented for the disambiguation of Main Bodies that were being misrecognized as Alifs.

Addition of Main Bodies with attached Diacritics

ہی ہا میں گر بنا ٹا گے
گی گس گو کی میں تے پ

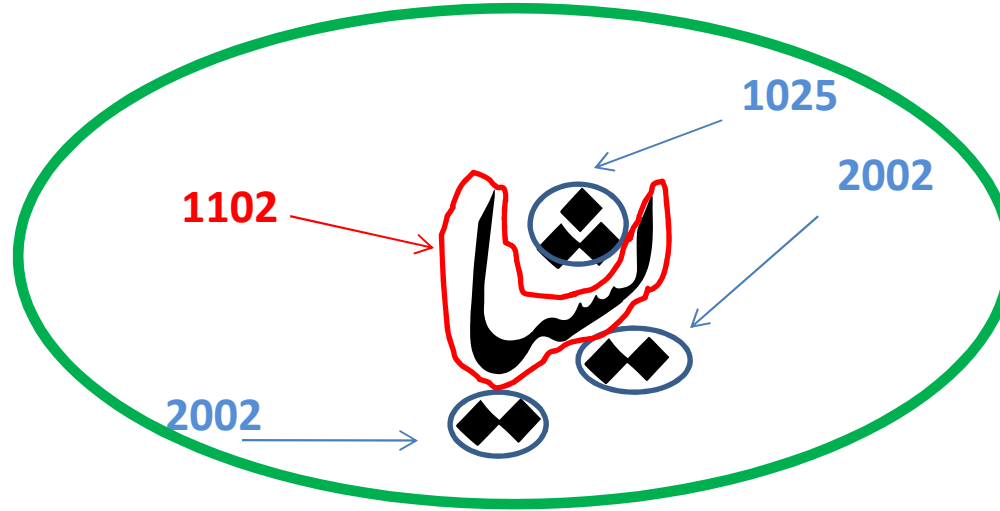
Addition of Latin Digits and Symbols

0 1 2 3 4 5 6 7 8 9) (؟
س

Final MB Testing Results

	F14 Previous Accuracies	F14 Final Accuracies	F16 Previous Accuracies	F16 Final Accuracies	F22 Previous Accuracies	F22 Final Accuracies	F36 Previous Accuracies	F36 Final Accuracies
Set1	99.22	99.27	99.19	99.80	97.96	99.66	99.35	99.59
Set2	99.06	99.34	98.36	99.09	98.76	98.67	98.62	98.74
Set3	98.02	98.56	98.86	98.88	96.52	97.42	97.54	97.55
Set4	96.92	97.36	96.10	97.23	95.77	97.15	94	96.47
Overall	98.30	98.63	98.13	98.75	97.25	98.22	97.38	98.09

Lookup Table



1119

Ligature ID	Ligature String	MBID	Diacritic Sequence
1	ا	623	
9	میں	4495	2002
10	ت	704	1002
1119	پیشا	1102	2002 1025 2002

Automatic Lookup Table Generation

Ligature Indexed List



Ligatures reduced
to MB Classes



Generation of
Ligature Diacritic
Sequences



Merging of Confused
MB Classes



Addition of Dia
Attached MB Classes



Lookup Table

Character	Position (initial, medial, final and isolated)	Mapping Character Class
ب پ ت ٹ ٹھ	All Positions	ب
ج چ ح خ	All Positions	ج
د ڈ ذ	All Positions	د
ر ژ ز ژر	All Positions	ر
س ش	All Positions	س
ص ض	All Positions	ص
ط ظ	All Positions	ط
ع غ	All Positions	ع
ف	All Positions	ف
ق	Final and Isolated	ق
ق	Initial and Medial	ف
ک گ	All Positions	ک
ل	All Positions	ل
م	All Positions	م
ن	Final and Isolated	ن
ن	Initial and Medial	ب
و	All Positions	و
ہ ء	All Positions	ہ
ھ	All Positions	ھ
ء	All Positions	ء
ی	All Positions	ی
ے	All Positions	ے
ئ	Initial and Medial	ب
ی	Final and Isolated	ی

22nd April, 2014

Center for Language Engineering (CLE)

صوبہ ہے ﴿ غلط نہیں۔ چمن میں واقعی زندگی کافی ﴿ ﴿ ﴿ اور یہاں کے لوگ ﴿ مخلصی اور ایماندار ہیں۔

پشین ضلع کے مقامی باشندے ﴿ مختلف قبیلوں میں ﴿ ہوتے ہیں۔ ان میں بڑے بڑے ﴿ ﴿ ﴿ اچکنئی عمیدزئی اور عشی زئی ہیں۔ ان کے علاوہ اور بھی ﴿ ﴿ ﴿ چھوٹے چھوٹے ﴿ آباد ہیں۔

عمیدزئی ﴿ کی ایک لڑکی جس کا نام مریم تھا پشین کے نزدیک ﴿ چھوٹے بھائی فرزا اور ﴿ بوڑھے وادا کے ساتھ رہتی تھی۔ اس کی عمر تقریباً تیرہ سال کی تھی۔ مریم کے ماں با ﴿ کچھ سال پہلے ﴿ کے ایک حادثے میں ہلاک ہو گئے تھے۔ ننھی مریم بھی حادثے کے وقت ان کے ساتھ تھی۔ مگر ﴿ اللہ رکھے اے کون ﴿ ﴿ گھر ﴿ کھڑ میں گر گئی۔ ماں ﴿ اسی وقت اللہ کو پیاری ہو گئی۔ مریم کو خراش بھی نہیں آئی۔ با ﴿ سرکاری اسپتال میں دو تین دن زندہ رہ کر چل بسا۔ ﴿ ﴿ کے

صوبہ ہے تو غلط نہیں۔ چمن میں واقعی زندگی کافی مشکل ہے۔ اور یہاں کے لوگ بہت مخلصی اور ایماندار ہیں۔

پشین ضلع کے مقامی باشندے مختلف قبیلوں میں بٹے ہوئے ہیں۔ ان میں بڑے بڑے قبیلے اچکنئی، حمیدزئی اور عشی زئی ہیں۔ ان کے علاوہ اور بھی بہت سے چھوٹے چھوٹے قبیلے آباد ہیں۔

حمیدزئی قبیلے کی ایک لڑکی جس کا نام مریم تھا، پشین کے نزدیک اپنے چھوٹے بھائی فرزا اور اپنے بوڑھے دادا کے ساتھ رہتی تھی۔ اس کی عمر تقریباً تیرہ سال کی تھی۔ مریم کے ماں باپ کچھ سال پہلے بس کے ایک حادثے میں ہلاک ہو گئے تھے۔ ننھی مریم بھی حادثے کے وقت ان کے ساتھ تھی۔ مگر جسے اللہ رکھے اسے کون چسکھے۔ بس گہرے کھڑ میں گر گئی۔ ماں تو اسی وقت اللہ کو پیاری ہو گئی۔ مریم کو خراش بھی نہیں آئی۔ باپ سرکاری اسپتال میں دو تین دن زندہ رہ کر چل بسا۔ باپ کے

Error Analysis

- The diacritic IDs for the middle position, starting with 3 were not included in the lookup table.
- The ranked list of misrecognized ligature contained Main Bodies that could be disambiguated with diacritics.

Ligature ID of لُق	Ligature String of لُق	MBID of لُق	Diacritic Sequence of لُق
2476	لُق	3931	1002

Maturation Process of the Ligature Based Urdu Noori Nastalique Optical Character Recognizer

Desired Ligature	Ranked List	Recognized MBID	Recognized Dia Sequence	Ligature Returned
مشکل	مسکل	4687	3025	null
	بیسکل	815	3025	null
	مبسکل	4393	3025	null
	جبکل	1921	3025	null
	سبکل	2450	3025	null
	بییکل	4350	3025	null
	مبیکل	4461	3025	null
	بھکل	1807	3025	null
	سمکل	2753	3025	null
	سرکل	2779	3025	null

22nd April, 2014

Center for Language Engineering (CLE)

Maturation Process of the Ligature Based Urdu Noori Nastaliq Optical Character Recognizer				
Desired Ligature	Ranked List	Recognized MBID	Sequence	Ligature Returned
ہے	ہے	1839	2302	null
	ہے	775	2302	null
بہت	بہت	1101	2001 2302 1002	null
	بہت	938	2001 2302 1002	null
	بہت	3325	2001 2302 1002	null
	بہت	1814	2001 2302 1002	null
	بہت	1698	2001 2302 1002	null
سے	سے	4953		سے
	سے	5025		null
	سے	775		null

صوبہ بے تو غلط نہیں۔ چمن میں واقعی زندگی کافی مشکل ہے۔ اور یہاں کے لوگ بہت محنتی اور ایماندار ہیں۔

پشین ضلع کے مقامی باشندے مختلف قبیلوں میں بٹے ہوئے ہیں۔ ان میں بڑے بڑے قبیلے اچکزئی، حمیدزئی اور عشی زئی ہیں۔ ان کے علاوہ اور بھی بہت سے چھوٹے چھوٹے قبیلے آباد ہیں۔

حمیدزئی قبیلے کی ایک لڑکی جس کا نام مریم تھا پشین کے نزدیک اپنے چھوٹے بھائی فراز اور اپنے بڑے دادا کے ساتھ رہتی تھی۔ اس کی عمر تقریباً تیرہ سال کی تھی۔ مریم کے ماں باپ کچھ سال پہلے بس کے ایک حادثے میں ہلاک ہو گئے تھے۔ ننھی مریم بھی حادثے کے وقت ان کے ساتھ تھی۔ مگر جسے اللہ رکھے اسے کون چکھے۔ بس گہرے کھڈ میں گونگتی۔ ماں تو اسی وقت اللہ کو پیاری ہو گئی۔ مریم کو خراش بھی نہیں آئی۔ باپ سرکاری اسپتال میں دو تین دن زندہ رہ کر چل بسا۔ باپ کے

صوبہ ہے تو غلط نہیں۔ چمن میں واقعی زندگی کافی مشکل ہے۔ اور یہاں کے لوگ بہت محنتی اور ایماندار ہیں۔

پشین ضلع کے مقامی باشندے مختلف قبیلوں میں بٹے ہوئے ہیں۔ ان میں بڑے بڑے قبیلے اچکزئی، حمیدزئی اور عشی زئی ہیں۔ ان کے علاوہ اور بھی بہت سے چھوٹے چھوٹے قبیلے آباد ہیں۔

حمیدزئی قبیلے کی ایک لڑکی جس کا نام مریم تھا، پشین کے نزدیک اپنے چھوٹے بھائی فراز اور اپنے بوڑھے دادا کے ساتھ رہتی تھی۔ اس کی عمر تقریباً تیرہ سال کی تھی۔ مریم کے ماں باپ کچھ سال پہلے بس کے ایک حادثے میں ہلاک ہو گئے تھے۔ ننھی مریم بھی حادثے کے وقت ان کے ساتھ تھی۔ مگر جسے اللہ رکھے اسے کون چکھے۔ بس گہرے کھڈ میں گر گئی۔ ماں تو اسی وقت اللہ کو پیاری ہو گئی۔ مریم کو خراش بھی نہیں آئی۔ باپ سرکاری اسپتال میں دو تین دن زندہ رہ کر چل بسا۔ باپ کے

Maturation Process of the Ligature Based Urdu Noori Nastalique Optical Character Recognizer

Testing Results with Initial Versions of Trained data and Lookup Table

Testing Results with Final Versions of Trained data and Lookup Table

Font	Total in Gold	Correct	%Accuracy CR
14	31458	24363	0.774
16	15366	12348	0.804
18	12392	10129	0.817
20	9299	7024	0.755
22	7105	6104	0.859
24	758	527	0.695
26	27	24	0.889
28	113	92	0.814
32	232	154	0.664
36	419	197	0.470
38	13	8	0.615
40	158	61	0.386
42	13	12	0.923
Average:			0.728

Font	Total in Gold	Correct	%Accuracy CR
14	31483	28017	0.890
16	15366	14107	0.918
18	12392	11294	0.911
20	9897	8337	0.842
22	7105	6799	0.957
24	758	568	0.749
26	27	26	0.963
28	113	100	0.885
32	232	183	0.789
36	419	221	0.527
38	13	13	1.000
40	158	64	0.405
42	13	12	0.923
Average:			0.828

Testing Results

CR Accuracy of 199 Document Pages (Initial Version)	CR Accuracy of 199 Document Pages (Final Version)
77%	87%

Challenges



Joined
MBs



Noise attached
with MB



Broken MBs




Different Font



Untrained Symbols

Thank you

Details of Tesseract Training Files

- .tiff Image: 
- .box File: lists the characters in the training image, with the coordinates of the bounding box around each character.
- .tr File: contains information about features that are polygon segments of the outline normalized to the 1st and 2nd moments, and features to correct for the moment normalization to distinguish position and size (eg c vs C and , vs ')

Details of Tesseract Training Files

- Unicharset File: lists the set of possible characters it can output, and character properties.
- Mftraining Files: contain information about shape prototypes, number of expected features for each character.
- Cntraining Files: contain information about the character normalization sensitivity prototypes

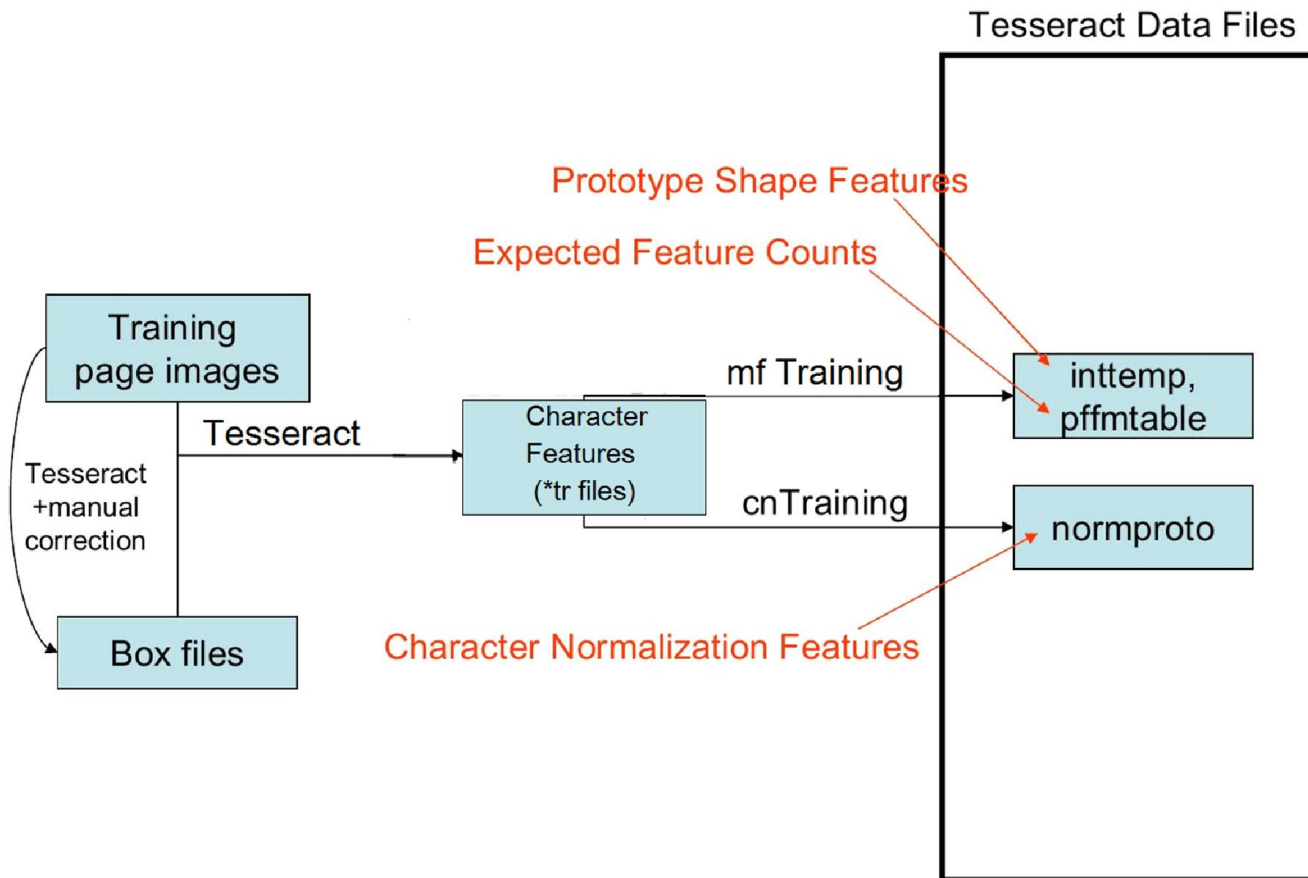
Manual Generation of Training Files

- .tiff image is to be created from the Main Body tokens of class.
- .box file is generated through command prompt.
- .tr file is generated through command prompt.
- Incase of .box or .tr file generation failure, .tiff image has to be edited, or regenerated.
- The above process has to be repeated for each MB class i.e. 5589 times.

Automatic Generation of Training Files

- Generates .tiff images, .tr and .box files of all 5589 classes in a single step.
- Creates a log file showing the success and failure of .tr and .box file generation for each class.
- Separates the classes with failed .tr and .box file generation, and the 1st step is repeated for them.
- Combines .tr and .box files of all classes into single .tr and .box files.
- Unicharset extraction, Mftraining and Cntraining are carried out on these combined .tr and .box files, and trained data is generated.

Training Tesseract OCR



Module Ligature Based Urdu Noori Nastalique Optical Character

