

Urdu OCR

Ahmed Muaz

Associate Development Engineer

Center for Research in Urdu Language Processing (CRULP)

PAN Localization Project

Urdu

- Written in Arabic script from right to left
- Bidirectional nature
 - Numbers are written from left to right
- 37 characters in Urdu Alphabet
 - Diacritical marks
- Most of the printed material is in “Nastaleeq” writing style/script

Character set and Classification

- 21 classes

ق	11.	آ	1.
ک گ	12.	ب پ ت ٹ ث	2.
ل	13.	ج چ ح خ	3.
م	14.	د ڈ ذ	4.
ن	15.	ر ژ ز	5.
و	16.	س ش	6.
ہ	17.	ص ض	7.
ھ	18.	ط ظ	8.
ء	19.	ع غ	9.
ی	20.	ف	10.
اے	21.		

Properties of Nastaleeq Script

- Diagonally written with stacking of Characters

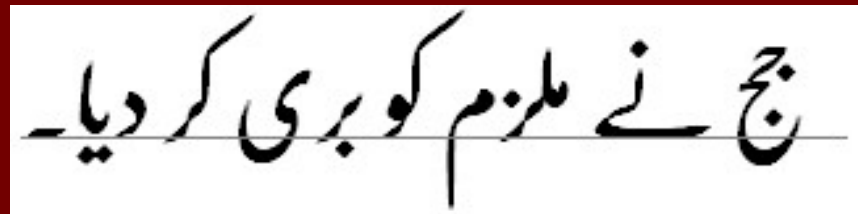


ن نصی نصیحت نصیحتیں

- Hidden baseline



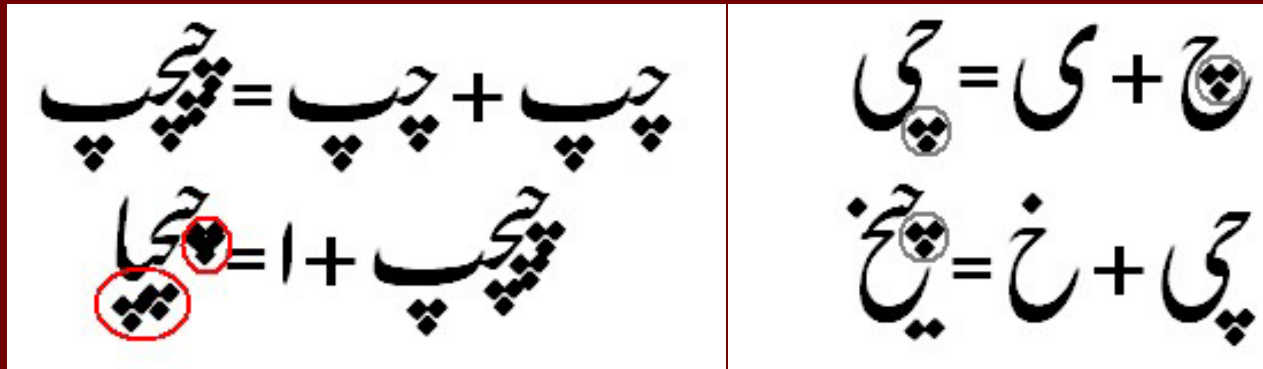
ابجدضطنی



حج نے ملزم کو بری کر دیا۔

Properties of Nastaleeq Script

- Diacritic re-placement during joining process

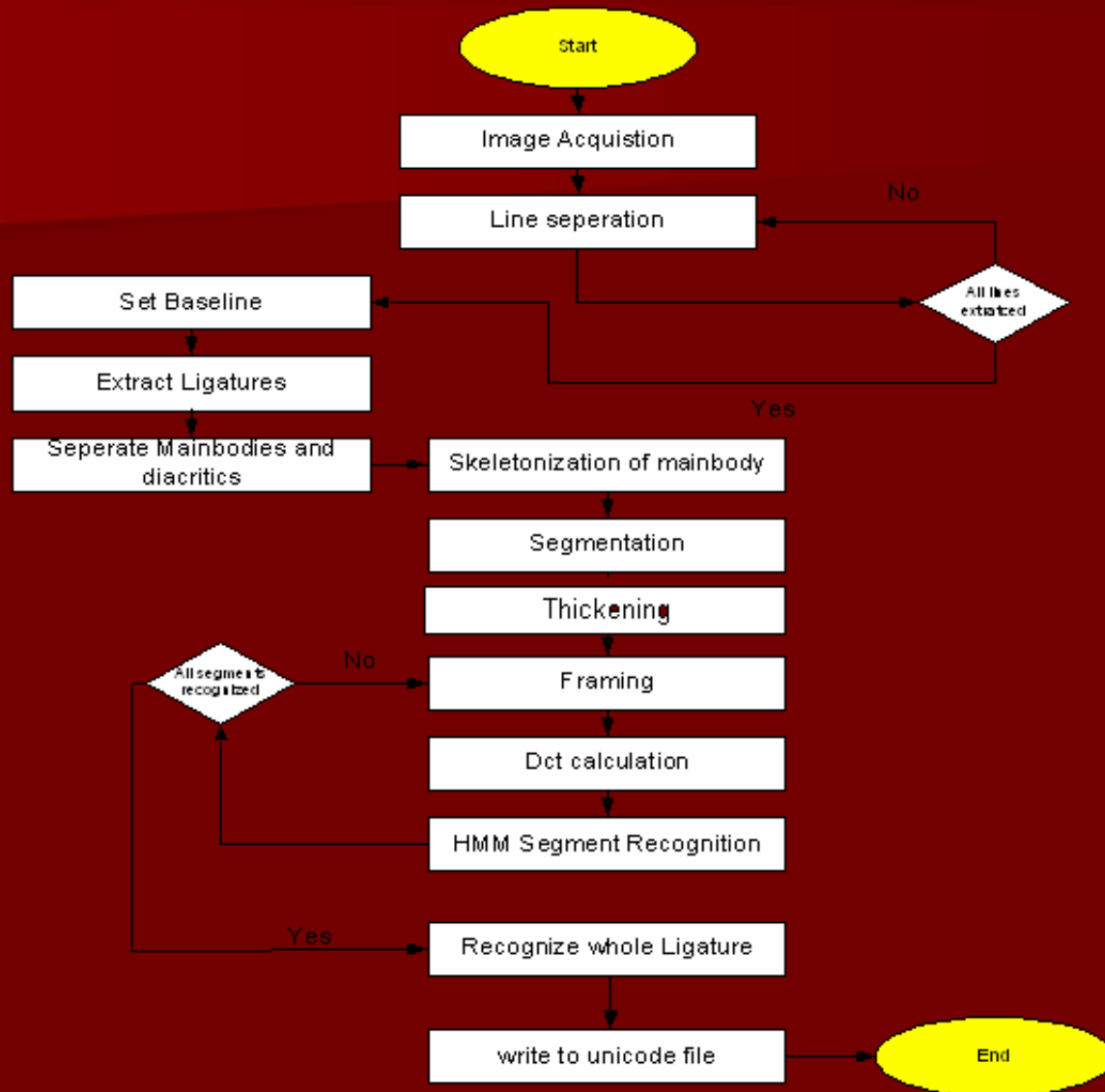


- Shifting with addition of every character

Urdu OCR development

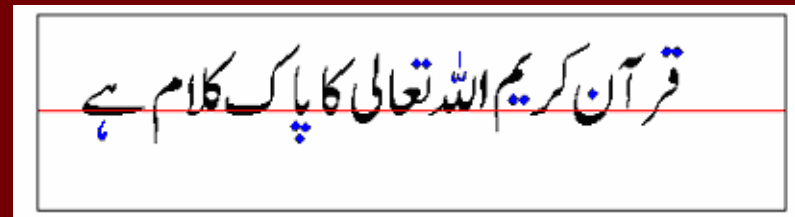
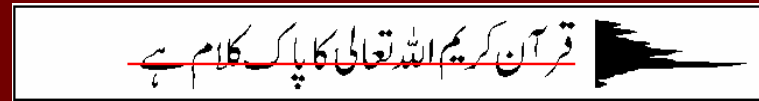
- Scope of current research
 - Noori Nastaleeq
 - Font size 36 Point
 - Scanned at 300 DPI

Urdu OCR Design



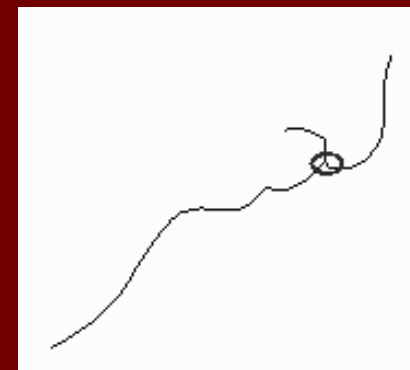
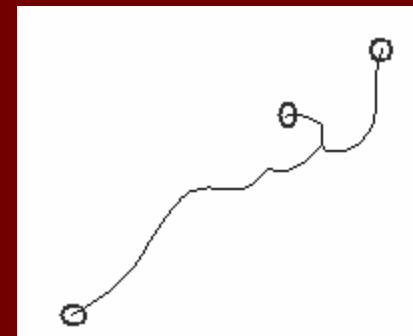
Segmentation Preprocessing

- Separate lines
 - Vertical Histogram
- Set Base Line
 - At maximum Intensity
- Main body Isolation
 - Attached to baseline



Segmentation I

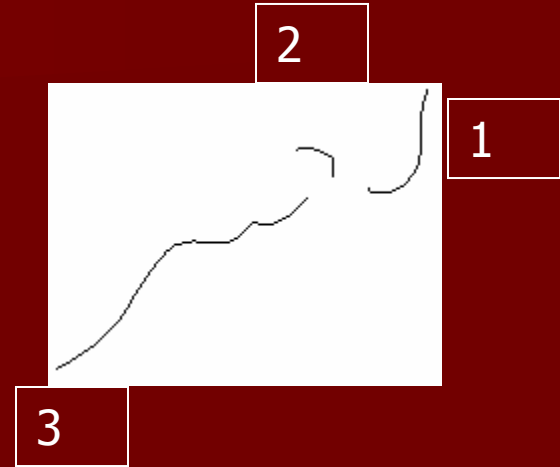
- Skeletanization (Thinning)
 - Jang Chin Masking Algorithm
- Starting point
 - Find all free points
 - Leftmost from bottom to top
- Chain code traversal
- Segmentation point
 - Stroke Junction



Segmentation II

■ Cutting

- At segmentation point
- Ordered from right to left, top to bottom



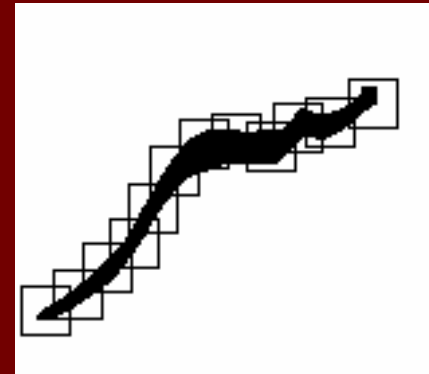
■ Thickening

- Restoration of original shapes
- Cutting original ligature along with segmented stroke



Framing

- Divide segment in equal sized windows
- Sliding window framing
 - 15*15 pixel frame
 - Overlapped
 - Sliding along thin stroke
- Frames are numbered according to their position
- Discrete Cosine Transformation



Recognition and Post processing

■ Recognition System

- Hidden Markov Model
- Used HTK for training HMM
- Frames are used as states of HMM
- The highest probable HMM ID is returned

■ Post Processing

- Deterministic Automata to get character ID
- Rule file for construction of Ligatures

Training Data Population

- Class based incremental training
 - Generating combinations of characters
 - Adding more classes to cover more ligatures
- Ligatures for training were taken out of 46K words clean corpus
- 13,628 unique ligatures
- Only 3,184 ligatures are frequent
- Used 30 samples per ligature

Current Status

- We have trained 13 classes (29 Characters)
- That covers 1,367 frequent Ligatures
- There are 376 segments trained
- Accuracy for different classes

Accuracy	Recognized	Tested Tokens	Class
98 %	1,087	1,100	ا
93%	714	765	ب
81%	664	814	ج
98.3	307	312	ر
71%	593	825	س
90%	588	656	ی

Challenges

- Various shapes of same character
- Humongous training data
 - Addition of on new class adds 150-200 Ligatures
 - Total ligatures 5,000-6,000
 - That yields segments 20-30K
- Clustering of segments for training
 - New clusters
 - Matching with already trained clusters
- Over and under segmentation
 - Variation in the shape of ligature can cause change in thinning and hence segmentation is affected

Future Work

- Diacritic association
 - Improvement in recognition
 - Character identification from class ID
- Development of (Semi) automatic training system
- Improvement in Segmentation Algorithm
- Preprocessing engine

Thank you

Suggestions/Comments
Questions