Recognition of Printed Devnagari Characters With Regular Expression in Finite State Models

Latesh Malik
Asst. Prof.
G.H.Raisoni College of Engineering, Nagpur
Dr. P.S. Deshpande
Asst. Prof.
V.N.I.T., Nagpur
Previous Work

Objective:

1. Devnagari is used by number of Indian languages including Sanskrit, Hindi and Marathi. Hindi is the world third most commonly used language after Chinese and English.

2. An OCR has a variety of commercial and practical applications in reading forms, manuscripts and their archival etc. Such a system facilitates a keyboard less user computer interaction. Also the text which is printed can be directly transferred to the machine.
Properties of Devnagari Script

- It is written and read from left to right
- Characters are distinguished by presence of matras
- Matras are dependent vowels used for representing a vowel sound that is not inherent to the consonants.
Properties Of Devnagari Script

- Header lines for words
- Upper modifiers
- Lower Modifiers
Work done for printed Devnagari script

- Chaudhari and Pal (ISI, Kolkata) have developed a devnagari OCR system which is being marketed as custom solution is not yet available as an off the shelf product.
- A feature based tree classifier is used to recognize the basic characters. Error detection and correction for the OCR based on dictionary search has led to the recognition accuracy of 91.25% at word level and 97.18% at character level.
Work done for printed Devnagari script

- Bansal et al proposed Hindi text recognition system by integrating knowledge sources. They achieved accuracy of 87% at character level.
Proposed Approach

- Image is converted into binary form.
- Apply operators to convert into string.
- Shape, joints are extracted in the form of string and design regular expressions.
- Matching of regular expression will recognize a character.
Feature Extraction

- **Joints**
  - Joint with vertical line (loop, curve, line joint)
  - Joint with header line (loop, curve, line joint)
- **Shapes**
  - Ascending curve
  - Descending curve
  - Straight line
  - Division of one curve into two
  - Merging of two curves into one
  - Starting of one curve from vertical (line, curve)
  - Starting of two curves from vertical (line, curve)
String Operators
Encoding in String

Shape of
seq. no. = 1

Segments

Encoded String = UUUQQQQDD

column no = 09
starting row no. = 2
length of segment = 5
connectivity code = D
seq no. = 1
Regular Expressions

- Sequence of geometrical properties of the character, like strokes and their directions, end points, or intersection of segments, and loops can be denoted with regular expressions. If a character have sequence like loop, intersection, loop and then end point can be represented as

\[
[^{\text{loop}}]*\text{loop}\{1\}[^{\text{intersection}}]*(\text{intersection}) + [^{\text{loop}}]*\text{loop}\{1\}[^{\text{end point}}]*(\text{end point}) +
\]
Designing of Regular Expression

\[ \text{[ } ^{\text{SSS}}*\text{(SSS)} + [ ^{\text{K}}]*\text{K} + \text{Q} + \text{(SS)} + \text{Q} + \text{LQ} + \text{]} \]

\[ \text{QQQQQSQQSQSSSQLKQQQQQQQQQQQQQQQQQQQQSSQQQLQQQQQQ+ (SQ) +} \]

\[ \text{QQQQQQSQSQSQYQUDQQDQLDQQLQQQQQLQQQQQQQQQQQQQQQQ} \]

\[ \text{QQQQQQSQSQSLUQLQQLQQLQI Q} \]

\[ \text{Q+(SQ)+Y+[ } ^{\text{QUD}}*\text{(QUD)}+\text{ [ } ^{\text{QQ}(D|L)}*\text{(QQ}(D|L)\text{)}]+[ ^{\text{SSS}}]*\text{SSS([ } ^{\text{QL}}*\text{(QL)})+(Q|I)+} \]
Size Invariance

Regular expression created for one size of a character matches with the string generated from any size of the same character even if generated string is not same. Normalization (thinning etc..) of character is not required.

क क क क क
Conclusion

- Achieved accuracy of 100% for printed characters
- Use of regular expression in the field of character recognition is found to be fruitful and new concept. In this paper modifiers are not taken into account, and they need to be processed differently.
- It can be applied to other languages as well as handwritten characters
Classifier 2 (Edit distance method)

Binary character image is scanned from left to right
Structural pattern is found.
Structural pattern is compared against all patterns stored in training file
Distance is calculated using edit distance method
**Classifier 2** (L to R scan)

<table>
<thead>
<tr>
<th>Classifier 2</th>
<th>Classifier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Classifier 2" /></td>
<td><img src="image" alt="Classifier 3" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier 4</th>
<th>Classifier 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Classifier 4" /></td>
<td><img src="image" alt="Classifier 5" /></td>
</tr>
</tbody>
</table>
Classifier Combination

- Classifier 1: regular expression matching
  - Trained patterns created from L to R scan
- Classifier 2: left to right scan and minimum edit distance method
  - Top 5 characters
- Classifier 3: right to left scan and minimum edit distance method
  - Top 5 characters
- Classifier 4: top to bottom scan and minimum edit distance method
  - Top 5 characters
- Classifier 5: bottom to top scan and minimum edit distance method
  - Top 5 characters

N matching characters

Count occurrence of each character and sort in descending order of occurrences

Fine classification
- Top most element in the list is the identified character

Coarse classification
- If character is found in list having occurrence >=2
Majority voting scheme

No of matches from regular expressions 1

*Character is K.jpg count 9*

Character is anta-jya.jpg count 2

Character is pha.jpg count 6

Character is danta-s.jpg count 2
Results of two different classifiers if applied in isolation

<table>
<thead>
<tr>
<th>Method</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum edit distance method (L to R scan)</td>
<td>70%</td>
</tr>
<tr>
<td>Regular expression matching</td>
<td>67 %</td>
</tr>
</tbody>
</table>
Top choices results of combined classifier

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Proposed method result (Combined 5 classifiers)</th>
<th>Accuracy obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Top 1 choice</td>
<td>85%</td>
</tr>
<tr>
<td>2</td>
<td>Top 2 choices</td>
<td>88%</td>
</tr>
<tr>
<td>3</td>
<td>Top 3 choices</td>
<td>91%</td>
</tr>
<tr>
<td>4</td>
<td>Top 4 choices</td>
<td>92%</td>
</tr>
<tr>
<td>5</td>
<td>Top 5 choices</td>
<td>95%</td>
</tr>
</tbody>
</table>
# Recognition at various level

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Level</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Horizontal line separation</td>
<td>98%</td>
</tr>
<tr>
<td>2</td>
<td>Word isolation</td>
<td>94%</td>
</tr>
<tr>
<td>3</td>
<td>Character isolation</td>
<td>85%</td>
</tr>
<tr>
<td>4</td>
<td>Character Recognition (Coarse classification)</td>
<td>95%</td>
</tr>
<tr>
<td>5</td>
<td>Character Recognition (Fine classification)</td>
<td>85%</td>
</tr>
</tbody>
</table>
References

Paper Presentation On Work

3. Dr. P.S. Deshpande, L. Malik, “Recognition of Hand written Devnagari Characters using Percentage Component Regular Expression and Tree Classifier” at International Conference on Image Processing organized by University Visvesvaraya College of Engineering Bangalore from 10-13 August 2007
5. “Fine Classification & Recognition of Hand Written Devanagari Characters with Regular Expressions & Minimum Edit Distance Method” accepted in Journal of Computers, ISSN : 1796-203X Volume : 3 Issue : 5 Date : May 2008 Academy publisher (JOURNAL).
6. “Recognition of Handwritten Devnagari Script” submitted to journal
Other related paper presentations

1. S.Arora, L.Malik,” Classification Of Gradient Change Features Using MLP For Handwritten Character Recognition “, International conference in emerging applications on IT , 10-11 FEB 2006, Organised by CSI Kolkata.


3. L.Malik, R. Welekar ,” Cursive character recognition using four connected segments and minimum edit distance “, First International conference on Information Technology, 19-21 March 2007, Haldia Institute of Technology, WB


RECEIVED GRANT

- Received grant from AICTE under RPS of Rs. 7.5 Lakhs for project “Recognition of handwritten manuscript in Devnagari Script”
- Duration of project is 2 years.
Thanking You