Hindi (Devanagari) OCR Notes

* Background on Devanagari
  + Devanagari script used for many languages
    - Used for Hindi, but also for Marathi, Sanskrit, and Nepali
  + 11 vowels and 33 consonants
  + Variety of diacritics called modifiers which can combine characters to make new sounds called conjuncts
  + New shape clusters called compound characters
  + Words joined with headline called Shirorekha
* Preprocessing methods
  + **Thresholding/Binarization** increases contrast by changing gray-scale (analog) images to discrete black and white (digital)
  + **Noise reduction** clears gaps in lines, segments, bumps, loops etc.
  + **Skew detection and correction** transforms the image based on the angle at which the picture was taken so that the processed image is viewed head on
  + **Size normalization** to make each character uniform
  + **Thinning** allows easier boundary detection
* Segmentationseparates sequences of characters into sub images, such as breaking apart characters in a word or isolating diacritics
* Feature Extraction
  + Continuous signal broken down into linear combination of series of well-defined functions
    - Fourier Transform
    - Gabor Transform
    - Wavelets
    - Moments
    - Karhunen-Loeve Expansion
  + Statistical Features
    - Zoning
    - Crossings and distances
    - Projections
  + Measuring geometrical and topological features of character set
    - Freeman’s chain coding
    - Graphs and trees
* Character Classification with variety of methods
  + Template matching
  + Statistical techniques
  + ANNs, SVMs
  + Combination Classifier
* Postprocessing
  + Dictionary for correcting minor mistakes
    - “Spell check”
  + Natural Language Processing to provide context

**References**

“A Overview of Devnagari Character Recognition” by V. J. Dongre and V. H. Mankar. <https://arxiv.org/pdf/1101.2491.pdf>

“Optical Character Recognition for Hindi Language Using a Neural-network Approach” by D. Yadav, S. Sánchez-Cuadrado and J. Morato. <https://pdfs.semanticscholar.org/42ab/98e6dddea52f5a84a4b8d631f8186b7e5169.pdf>

“Optical Character Recognition (OCR) for Printed Devnagari Script using Artificial Neural Network” by R. Singh, C. S. Yadav, P. Verma, V. Yadav. <https://pdfs.semanticscholar.org/1869/91bba497220adec7472ae9f31eec79fb86aa.pdf>