

FEATURE BASED OPTICAL BENGALI CHARACTER RECOGNITION USING MULTIPLE BACK PROPAGATION NEURAL NETWORKS

Fariha Nazmul

Roll - 1436

Kazi Wali Ullah

Roll - 1404

ABSTRACT

- Optical Character Recognition (OCR), an Important Area in Pattern Recognition and Image Processing
- Optical Bangla Character Recognition from Documents Written in Standard Bangla Font (25 characters)
- Use of Multiple Neural Networks for Character Recognition
- Implementation of the System Using MATLAB
- Performance Analysis of the OCR System

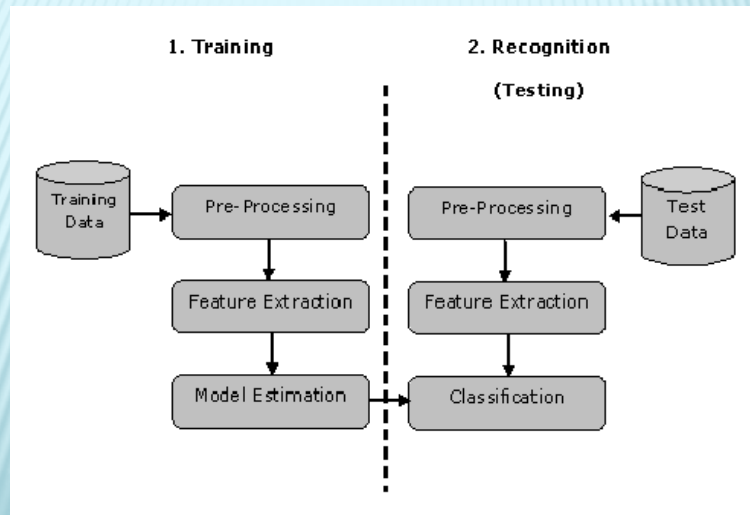
OUTLINE

- Introduction to OCR
- Design of the OCR
- Steps of the OCR
 - Preprocessing
 - Segmentation
 - Feature Extraction
 - Classification
- Performance Analysis
- Concluding Remarks

OCR

- Optical character recognition deals with recognition of characters acquired by optical means, typically a scanner or a camera.
- They can be printed or handwritten, of any size, shape, or orientation
- The printed or handwritten documents are transformed into ASCII files for the purpose of compact storage, editing, fast retrieval, and other file manipulations through the use of a computer.

STEPS IN A TYPICAL OCR SYSTEM



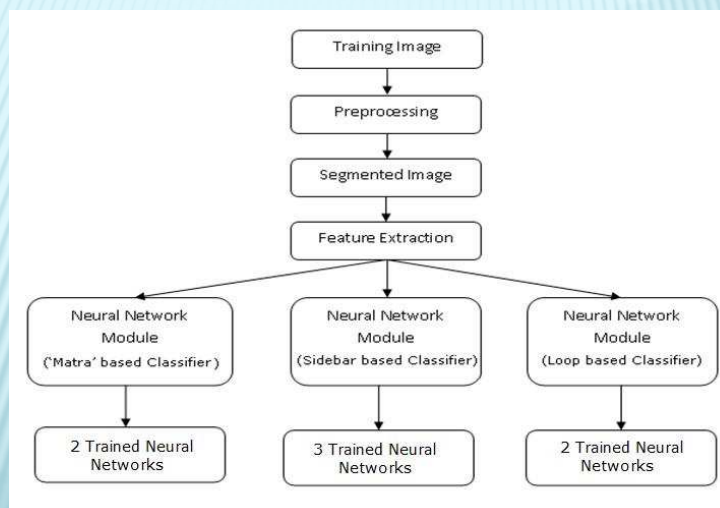
DESIGN OF THE OCR

MAJOR COMPONENTS

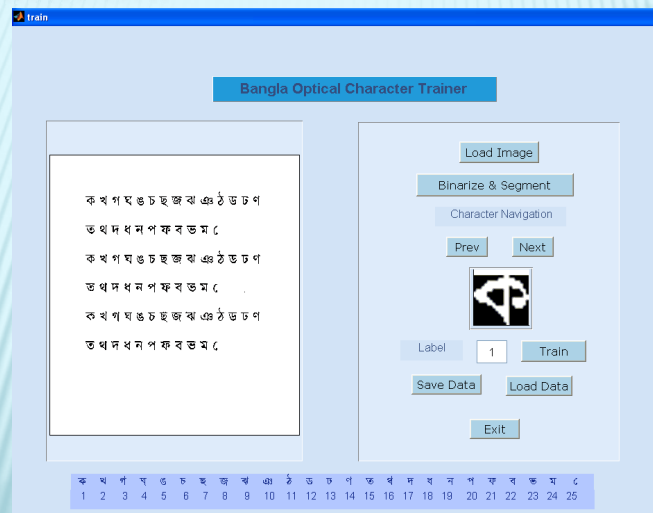
The OCR system consists of two major components:

- Training Component
- Recognizer Component

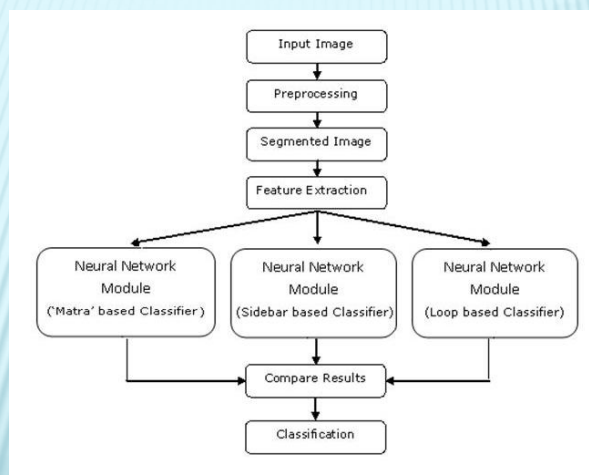
TRAINING COMPONENT



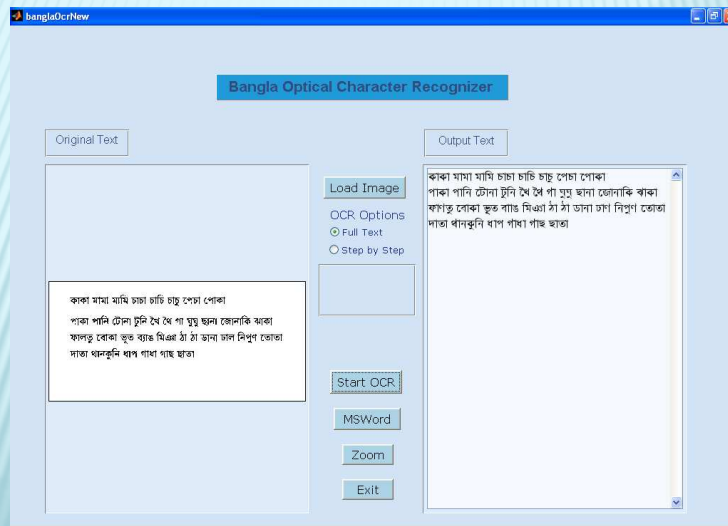
TRAINING GUI



RECOGNIZER COMPONENT



RECOGNIZER GUI



PREPROCESSING

PREPROCESSING STEPS

Preprocessing step deals with:

- **Translation:** To avoid the translation problem, the enclosing rectangles of patterns are detected. For example, *regionprops()* function of the MATLAB Image Processing Toolbox gives the smallest bounding box of each image segment.
- **Scaling:** The size of the patterns or characters is not always the same due to the use of different fonts sizes. Character images of different widths and heights are resized to the same dimension (20x20).
- **Rotation:** The orientation of the patterns is not always the same. If the neural network is trained with slightly rotated images, it should be able to classify such images correctly. But, image rotation may require a lot of CPU resources.
- **Noise:** Morphological operations, such as opening, closing, dilation, erosion are used to remove noises.

MORPHOLOGICAL OPERATIONS

- Technique of image processing based on shapes.
- **Dilation and Erosion**
 - Dilation adds pixels to the boundary
 - Erosion removes pixels from the boundary of objects in the image
- These operations are used in splitting characters that are joined by thin 'Matra' after 'Matra' removal operation.

MORPHOLOGICAL OPERATIONS...

- Effect of dilation and erosion in splitting characters that are joined by thin 'Matra' after 'Matra' removal operation:



SEGMENTATION

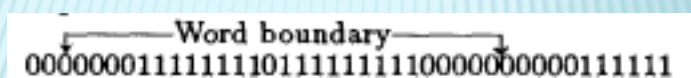
LINE SEGMENTATION

- The text lines are segmented by finding the valleys of the histogram computed by row wise sum of pixel values. A threshold t_i has been chosen at the training phase by observing a large number of images.

কখন কচু কাছাকাছি কাবা খাজনা খাটো
গজ গাছ গণিত গোমতী ঘনীভূত ঘোমটা
চিবুক চৌকাঠ মিশ্র ছাতা জাপান টোটা
জীবিকা জেদী ঘুঘু ঝামা ঢেকি জোনাকি
ডিম তোতা থানকুনি দানব দৈনিক ধান

WORD SEGMENTATION

- A segmented text line is scanned vertically. If in one vertical scan two or less black pixels are encountered then the scan is denoted by 0, else the scan is denoted by 1.

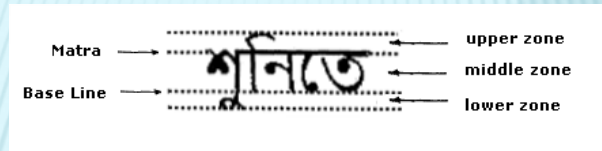

 Word boundary
 0000000111111101111111000000000000111111

- Now if a run of 0's is longer than a pre-specified integer threshold then the middle point of that run of 0's is considered as the boundary of a word.

কখন	কচু	কাছাকাছি	কাবা	খাজনা	খাটো
-----	-----	----------	------	-------	------

ZONE SEGMENTATION

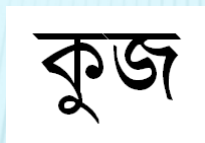
A Bangla word script can be partitioned into three horizontal zones.



- **The Matra/ Headline:** The row with the highest frequency value indicates the head line which has been denoted as 'Matra'.
- **The Base Line:** The average word height of the document is calculated when the text image is segmented into words. This denotes the location of the Base Line for each word.

CHARACTER SEGMENTATION

- Characters in Bengali words are written both horizontally and vertically.



- Almost all characters are connected to some other character.



CHARACTER SEGMENTATION (Cont.)

- When the 'Matra' region is deleted from middle zone, the characters in a word get topologically disconnected.

কথোপকথন

- But 'Matra' deletion fails to disconnect vertically connected characters. For this, the region below the Base Line is also removed.

কুড়

কড

CHARACTER SEGMENTATION (Cont.)

- Sometimes 'Matra' deletion or Base Line removal leads to character partitioning.

গ গ ঙ

গ গ ঙ

- To avoid this problem we have used a function *tryJoin()*, that tries to form the character again by combining the partitions.

FEATURE EXTRACTION

FEATURES USED

- **Structural Features**
 - 'Matra'/Headline
 - Left Sidebar
 - Right Sidebar
- **Statistical Features**
 - Euler Number
 - Horizontal Ratio
 - Vertical Ratio
 - Sum of 'ON' pixels in sixteen 5×5 sub-images

FEATURES USED (Cont.)

- Method of Moments
 - Eccentricity
 - Centroid
 - Horizontal Projection Skewness
 - Vertical Projection Skewness
 - Horizontal Projection Kurtosis
 - Vertical Projection Kurtosis

CLASSIFICATION

PROPOSED METHOD

The classification method implemented in this OCR system uses multiple ANNs rather than a single ANN

WHY MULTIPLE ANNs ?

Problems with Single ANN :

- When there are too many characters to recognize, the neural network becomes unable to detect the characters effectively.
- When there is only one neural network module to recognize a character, there is no other way to identify a character when the neural network module fails. Either the character goes undetected or it gives an erroneous result.

WHY MULTIPLE ANNs ? (Cont.)

Problems with Single ANN :

- When there are too many characters to be recognized and/or if the training pattern is the image of the characters then training the neural network takes a large number of epochs.
- Sometimes it may even be possible that a neural network for this sort of patterns can never be trained.

WHY MULTIPLE ANNs ? (Cont.)

- To ensure that an artificial neural network (ANN) recognizes all the characters effectively it is required that it works only with a small set of characters.
- For this purpose we have designed three classifiers:
 - Classifier Based on Matra
 - Classifier Based on Sidebar
 - Classifier Based on Loops

MATRA BASED CLASSIFIER

ANN	No. of Characters To Be Recognized	Character Set
ANN #1 (characters with 'Matra')	15	L, O, Q, R, S, W, X, Y a, c, e, g, h, i, j
ANN #2 (characters without 'Matra')	10	M, N, P, T, U, Z, b, d, f, -

SIDEBAR BASED CLASSIFIER

ANN	No. of Characters To Be Recognized	Character Set
ANN #3 (characters with 'Left Sidebar')	2	Q, Y
ANN #4 (characters with 'Right Sidebar')	13	L, M, N, O, T, U, Z, b, d, e, f, h, j
ANN #5 (characters without 'Sidebar')	10	P, R, S, W, X, a, c, g, i, -

LOOP BASED CLASSIFIER

ANN	No. of Characters To Be Recognized	Character Set
ANN #6 (characters with 'Loops')	12	L, O, P, Q, R, T, U, W, d, f, h, j
ANN #7 (characters without 'Loops')	13	M, N, S, X, Y, Z, a, b, c, e, g, i, -

SIMULATION PROCESS

Classifier #1

- If the input character has 'Matra' then simulate with ANN #1.
- Else simulate with ANN #2

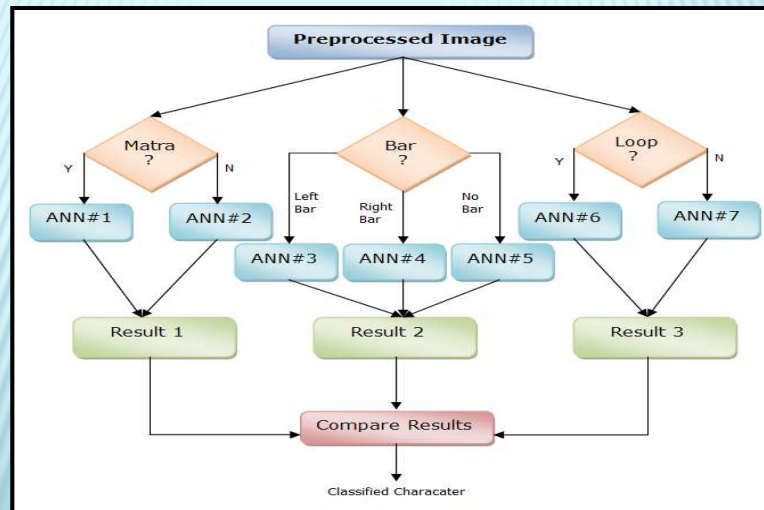
Classifier #2

- If the input character has 'Left Sidebar' then simulate with ANN #3.
- Else if the input character has 'Right Sidebar' then simulate with ANN #4
- Else simulate with ANN #5

Classifier #3

- If the input character has 'Loop' then simulate with ANN #6
- Else simulate with ANN #7

SIMULATION PROCESS (Cont.)



DECISION MAKING PROCESS

The results from all the three classifiers are combined in the following manner:

- If all the three results from the three classifiers are same, then it is considered as the final result.
- If two of the results are same, then it is considered as the final result.
- Lastly, if none of the results are same, then the result from the ANN that has the highest percentage of matching is considered as final result.

ADVANTAGES

- Each ANN becomes much more effective as it has to recognize characters only from a small subset.
- Three out of seven ANNs are simulated from all the three class to recognize a single character, thus the probability that all the ANNs will give a wrong result is reduced greatly.
- Shorter time to train .
- Shorter time to simulate.

TRAINING THE ANNs

Here all the neural networks are trained with features extracted from a resized image of a character, rather than with that resized image only.

FEATURES vs IMAGE

- It takes **several order of magnitude less epochs** to train the neural network using only the features extracted from the resized image than.
- An small 20×20 image needs **400 neurons** in the input layer. Where as it is enough to use only **29 neurons** in our case in the input layer to represent the features.
- When all the features are combined, it carries **more information** about a character than a small 20×20 image.
- Since, there is a drastic reduction of neurons in the input layer, it also means faster simulation or **faster recognition process**.

VOWEL MODIFIERS

- The detection of the vowel modifiers or 'Kars' are treated separately than the basic characters.
- These modifiers are detected based on their aspect ratio.
- When an 'akar' is detected then it is further checked whether it is a 'roshho-ikar' or a 'dirgho-ikar' or an 'oi-kar'
- The detection of the vowel modifiers in the bottom of a character are done in an effective manner.

MERGING SUB CHARACTERS

- In the last stage of the character recognition process, the information about coordinates of bounding box of sub-characters and the context is used to merge some of the sub-characters. The sub-characters are then converted to actual Bengali characters.
- To illustrate this process, let us consider the recognition process of the Bengali letter ৷ . To recognize this character, first we have to recognize the sub-character ৷ . Then we have to figure out that whether there is anything above the letter ৷ . If we can find anything above the ৷ then we can figure out that the recognized character is ৷ .

PROBLEMS

ট টী টে চ চি চি টৌ টৌ

PERFORMANCE ENHANCEMENT

- The performance of the OCR system is remarkably increased by training the ANNs with multiple noisy and noise free images of same character.
- Features are first extracted from all these training images, and then the mean values of all the features are used to train the ANNs.

EXAMPLE : TRAINING IMAGES



OCR PROGRAM

RECOGNIZABLE CHARACTERS

ক খ গ ঘ ঙ

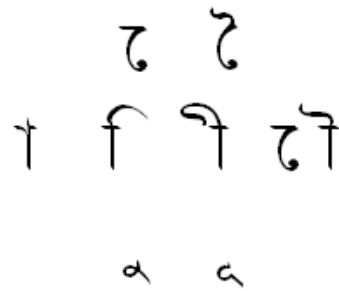
চ ছ জ ঝ ঞ

ট ঠ ড ঢ ণ

ত থ দ ধ ন

প ফ ব ভ ম

RECOGNIZABLE VOWEL MODIFIERS



PERFORMANCE ANALYSIS

EXPERIMENTAL RESULTS

Figure Name	Scan Resolution (dpi)	Number of Characters	Number of Errors	Accuracy Rate(%)	Time (sec)
Test1_300.bmp	300	297	8	97.31	30.39
Test2_300.bmp	300	142	2	98.59	13.39
Test3_150.bmp	150	25	2	92.00	4.38

PERFORMANCE ANALYSIS

- The experimental results shows that the OCR system has an error rate from 5-10%. Some of the reasons are:
 - Some of the characters have very similar structural and statistical features. Such as M and Y.
 - Some of the features may be falsely detected by the feature extraction module because of distorted image or image with very low resolution. For example, there is no loop in M. But sometimes only one pixel connects two disconnected parts that forms a loop.
 - Rotation or translation may lead to recognize a character incorrectly. For example, a rotated or translated character image is likely to have a very different aspect ratio.

CONCLUDING REMARKS

CONCLUSION

- The OCR system has been quite successful in the recognition of characters from Bangla text.
- Input Images of formats like jpg, gif, bmp are supported.
- The accuracy rate is 95 - 100% with test images written in same font.
- The execution time for an text image with 250 characters of size 2500X2500 takes around 40 seconds.
- The output of the OCR is allowed to be saved as text or in MSWord for further processing.
- To reduce the execution time the classifier modules can be executed parallelly.
- Recognition of Bangla compound characters require further research in the field of segmentation operations and feature extraction.

