

BANGLA OPTICAL CHARACTER RECOGNITION

A Thesis

Submitted to the Department of Computer Science and Engineering

of

BRAC University

by

S. M. Murtoza Habib

Student ID: 01201071

In Partial Fulfillment of the

Requirements for the Degree

of

Bachelor of Science in Computer Science

December 2005

DECLARATION

I hereby declare that this thesis is based on the results found by myself and my group partner Nawsher Ahmed Noor. Materials of work found by other researcher are mentioned by reference.

Signature of
Supervisor

Signature of
Author

ACKNOWLEDGMENTS

Special thanks to Dr. Mumit Khan who gave us guidance through our thesis research work and also for his support in the whole process. And also show him our gratitude for considering us as to do the thesis under his supervision.

We thank Sajib Dasgupta for his insight and encouragement for working on this area.

We also thank M. Ashraful Amin for his direction in the area of image processing and neural networks, and also thank Junaed Sattar for his insight in image analysis.

And finally we thank our research institute for providing us with old Bangla books and scanner as a requirement for collection of sample inputs.

ABSTRACT

Optical character recognition (OCR) is a technology to convert a digital image of text to editable text. An OCR system for Bangla language is proposed here. The proposed OCR system scans the digital image and recognizes the character. Before the recognition step, few preprocessing steps are needed. When a document is fed to the optical sensor (scanner) to get the digital image, there may be a few degrees of skew. Skew angle is the angle that the text lines in the digital image makes with the horizontal direction. For this, skew correction is the first step of preprocessing. Others preprocessing steps are noise removal, line separation, word separation, and character separation. After separating the characters from the image, this system recognizes the characters. To recognize a character I use feed forward Neural Network (NN) based recognition scheme.

Table of Contents

	Page
DECLARATION	II
ACKNOWLEDGMENTS	III
ABSTRACT	IV
TABLE OF CONTENTS	V
LIST OF TABLES	VI
LIST OF FIGURES	VII
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: SOME PROPERTIES BANGLA CHARACTER	2
CHAPTER 3: PROCESSING DIGITAL IMAGE	4
CHAPTER 4: SKEW ANGLE DETECTION AND CORRECTION.....	8
CHAPTER 5: LINE AND WORD SEPARATION.....	14
5.1 NOISE REMOVAL	14
5.2 LINE SEGMENTATION	15
5.3 WORD SEGMENTATION	16
5.4 CHARACTER SEGMENTATION.....	17
CHAPTER 6: CHARACTER RECOGNITION.....	25
CHAPTER 7: CONCLUSION.....	26
GLOSSARY	27
REFERENCES	32

List of Tables

Table	Page
TABLE 1: EXAMPLE OF MODIFIED SHAPE OF VOWEL	2
TABLE 2: EXAMPLE OF COMPOUND CHARACTER.....	3

List of Figures

Figure	Page
FIGURE 1: (A) VOWELS; (B) CONSONANT.....	2
FIGURE 2: NON-MODIFIED VOWEL IN A WORD	3
FIGURE 3: AN EXAMPLE OF DIGITAL IMAGE OF BANGLA TEXT	4
FIGURE 4: THE GRAYSCALE IMAGE.....	6
FIGURE 5: THE BINARY IMAGE	7
FIGURE 6: SELECTED CONNECTED COMPONENTS.....	10
FIGURE 7: UPPER ENVELOPE	11
FIGURE 8: UPPER ENVELOPE CONTAINING 2 LINES IN DIFFERENT DIRECTION.....	12
FIGURE 9: RADON TRANSFORM OF UPPER ENVELOPE (FIGURE 7).....	12
FIGURE 10: AFTER CORRECTION THE SKEW ANGLE	13
FIGURE 11: DOTS IN BANGLA CHARACTER	14
FIGURE 12: AFTER REMOVE DOTS HAVING SIZE 4 BY 4 OR LESS	14
FIGURE 13: PLOT OF HORIZONTAL BLACK PIXEL OF THE IMAGE.....	15
FIGURE 14: LINE 1	16
FIGURE 15: PLOT OF VERTICAL BLACK PIXEL OF LINE 1	16
FIGURE 16: FIRST WORD OF LINE 1	17
FIGURE 17: EXAMPLE OF SOME WORDS	17
FIGURE 18: PLOT OF HORIZONTAL BLACK PIXEL OF WORD FIGURE 17(B).....	18
FIGURE 19: PLOT OF HORIZONTAL BLACK PIXEL OF WORD FIGURE 17(D).....	19
FIGURE 20: (A) AND (B) RESPECTIVELY SHOWS SKEL AND THIN OF (C)	20
FIGURE 21: (A) AND (B) RESPECTIVELY SHOWS SKEL AND THIN OF (C).....	21
FIGURE 22: AFTER REMOVAL OF HEADLINE OF THIN WORDS	21
FIGURE 23: EXAMPLE OF CHARACTER NOT CONNECTED	22
FIGURE 24: EXAMPLE OF JOINED CHARACTER.....	23
FIGURE 25: EXAMPLE OF FONT BASED WORD	24

Chapter 1: Introduction

Optical Character Recognition (often abbreviated as OCR) involves reading text from paper and translating the images into a form (say ASCII codes) that the computer can manipulate. Although there has been a significant number of improvements in languages such as English, but recognition of Bengali scripts is still in its preliminary level. This thesis tries to analyze the neural network approach for Bangla Optical Character Recognition. A feed forward network has been used for the recognition process and a back propagation algorithm had been used for training the net. Before the training, some preprocessing steps were involved of course. Preprocessing includes translating scanned image into binary image, skew detection & correction, noise removal, followed by line, word and character separation. In this report, preprocessing steps were discussed from chapter 3 to chapter 5, and chapter 6 explains recognition. Chapter 2 discuss about some properties of Bangla Character.

Translation of scanned image into binary image, skew detection & correction, noise removal, line and word separation of the pre-processing steps were jointly analyzed and done by my group member and myself.

I concentrated on character separation while forming, training and recognition of characters were fully done by my group member.

Chapter 2: Some properties Bangla character

The writing style of Bangla is from left to right comprising of 11 vowels and 39 consonant characters. These characters may be called as the basic characters (figure 1). The concept of upper/lower case is absent in Bangla script. From Fig. 1 it is noted that most of the characters have a horizontal line at the upper level. This horizontal line is called head line. In Bangla language, we call it 'matra'.

<p>A A_v B C D E F G H I J</p>	<p>K L M N O P Q R S T U V W X Y Z _ ` a b c d e f g h i j k l m n o p q r s t u</p>
--	--

(a)

(b)

Figure 1: (a) Vowels; (b) consonant.

In Bangla script sometimes a vowel takes a modified shape depending on the position in a word. If the first character of the word is a vowel then it retains its basic shape. Generally a vowel following by a consonant takes a modified shape and placed at the left or right or both or bottom of the consonant (table 1).

Table 1: Example of modified shape of vowel

Vowel	A _v	B	C	D	E	F	G	H	I	J
Modified shape	v	w	x	y	~	„	‡	% _o	‡v	‡\$
K + vowel	K _v	wK	K _x	K _y	K~	K„	‡K	%K	‡K _v	‡K\$

For two consecutive vowels in a word, the second one remains its basic shape when the first one is in modified shape. (figure 2)

$$L + Av + I \rightarrow LVI$$

Figure 2: non-modified Vowel in a word

Again a consonant or vowel followed by a consonant sometimes takes a compound shape which we call as compound character (table 2). There are about 250 compound characters, where most of them are formed by consonant-consonant combination. Compounding of three consonants is also possible. Most interesting thing is if we change the order of same two consonant, the compound character is changed.

Table 2: Example of compound character

K + K	◦	e + e	eŸ
K + Z	३	e + `	ã
K + b	Kè	P + Q	"Q
K + g	´	P + Q + e	"Q _i
K + l	¶	R + R	¾
K + l + b	¶è	R + R + e	¾ _i
K + l + g	²	R + T	Á
K + j	K ₇	T + R	Ä
j + K	é	0 + K	¼

As we mentioned earlier, most of the basic characters have a head line (matra). Out of 50 basic characters 32 of them have this matra (head line) and most of the compound characters have this matra too.

Chapter 3: Processing digital image

A digital image of text is required which is achieved through scanning a paper or book containing Bangla script; figure 3 is such an image. The quality of the scanned paper in figure 3 was not that good, and thus presence of noise is higher. Now the first task is to convert this image to binary image (black and white). Generally the scanning image is true color (RGB image) and this has to be converted into a binary image, based on a threshold value.

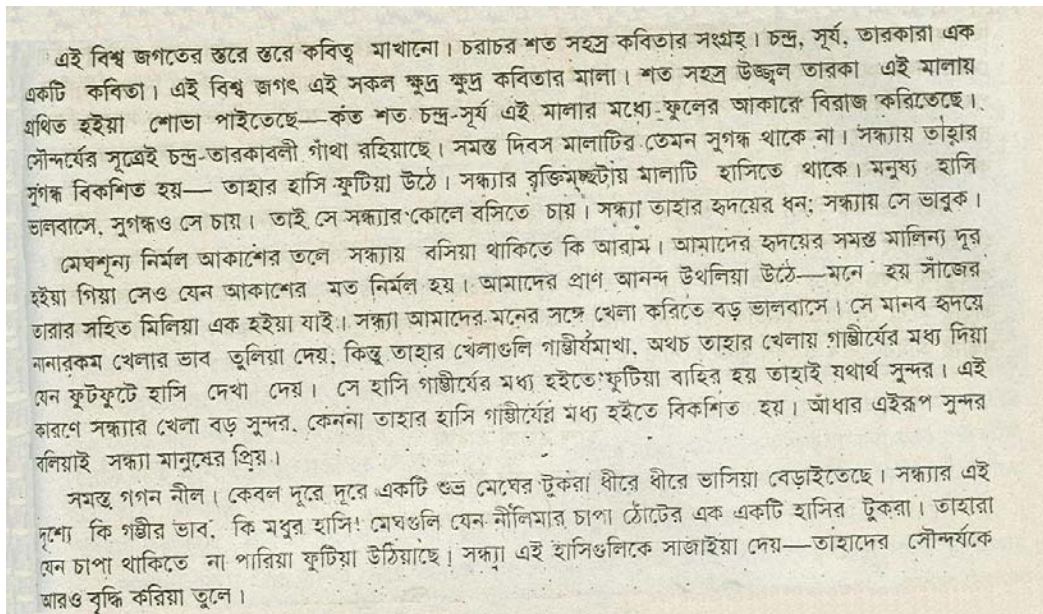


Figure 3: An example of digital image of Bangla text

Here Otsu's Method [1] has been applied to find the threshold value. The method chooses the threshold to minimize the intra class variance of the thresholded black and white pixels. The algorithm is:

Algorithm 1: Otsu's Method

Step 1: count the number of pixel according to color (256 color) and save it to matrix *count*.

- Step 2: calculate probability matrix P of each color, $P_i = \text{count}_i / \text{sum of count}$, where $i = 1, 2, \dots, 256$.
- Step 3: find matrix ω , $\omega_i = \text{cumulative sum of } P_i$, where $i = 1, 2, \dots, 256$.
- Step 4: find matrix μ , $\mu_i = \text{cumulative sum of } P_i * i$, where $i = 1, 2, \dots, 256$ and $\mu_t = \text{cumulative sum of } P_{256} * 256$
- Step 5: calculate matrix σ_b^2 where,
- $$\sigma_b^2_i = \frac{(\mu_t \times \omega_i - \mu_i)^2}{\omega_i - (1 - \omega_i)}$$
- Step 6: Find the location, idx , of the maximum value of σ_b^2 . The maximum may extend over several bins, so average together the locations.
- Step 7: If maximum is not a number, meaning that σ_b^2 is all not a number, and then $threshold$ is 0.
- Step 8: If maximum is a finite number, $threshold = (idx - 1) / (256 - 1)$;

To get a binary image, this RGB format image has to be converted to grayscale format, and then by using the threshold value (found by Otsu's method) this grayscale image is converted to binary image.

A RGB image is converted by eliminating the hue¹ and saturation² information while retaining the luminance³. The easy technique to do obtain this is to multiply 0.2989 with red color, 0.5870 with green color and 0.1140 with blue color of a particular pixel and put the summation of these three values to that particular location.

$$Y = \text{Red} * 0.2989 + \text{Green} * 0.5870 + \text{Blue} * 0.1140$$

¹ A hue refers to the degree of color within the optical spectrum, or visible spectrum, of light. "Hue" may also refer to a particular color within this spectrum.

² Saturation refers to the intensity of a specific hue.

³ Luminance describes the amount of light that passes through or is emitted from a particular area.

The grayscale version of the image (figure 3) is shown in figure 4.

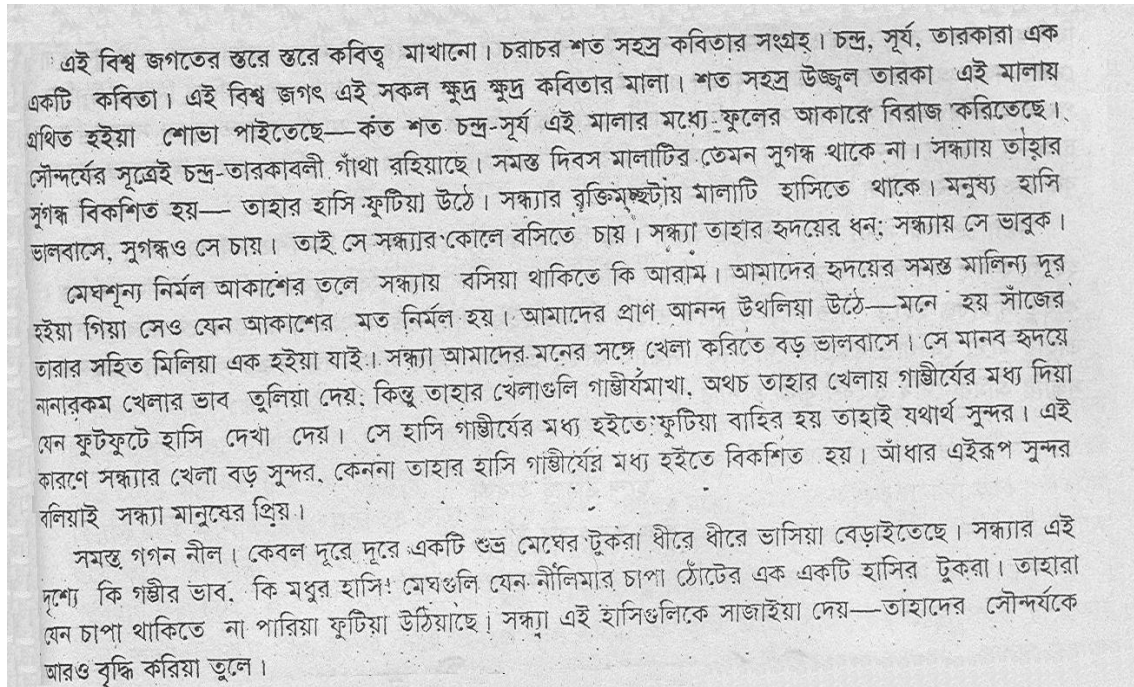


Figure 4: The grayscale image

In a grayscale image there are 256 combinations of black and white colors where 0 means pure black and 255 means pure white. This image is converted to binary image by checking whether or not each pixel value is greater than 255•level (level, found by Otsu's Method). If the pixel value is greater than or equal to 255•level then the value is set to 1 i.e. white otherwise 0 i.e. black. The binary image is shown in figure 5.

এই বিশ্ব জগতের স্তরে স্তরে কবিত্ব মাখানো। চরাচর শত সহস্র কবিতার সংগ্রহ। চন্দ্র, সূর্য, তারকারা এক একটি কবিতা। এই বিশ্ব জগৎ এই সকল ক্ষুদ্র ক্ষুদ্র কবিতার মালা। শত সহস্র উজ্জ্বল তারকা এই মালায় গ্রথিত হইয়া শোভা পাইতেছে—কত শত চন্দ্র-সূর্য এই মালার মধ্যে ফুলের আকারে বিরাজ করিতেছে। সৌন্দর্যের সূত্রেই চন্দ্র-তারকাবলী গাঁথা রহিয়াছে। সমস্ত দিবস মালাটির তেমন সুগন্ধ থাকে না। সন্ধ্যায় তাহার সুগন্ধ বিকশিত হয়— তাহার হাসি ফুটিয়া উঠে। সন্ধ্যার বৃজিমুষ্টিয় মালাটি হাসিতে থাকে। মনুষ্য হাসি ভালবাসে, সুগন্ধও সে চায়। তাই সে সন্ধ্যার কোলে বসিতে চায়। সন্ধ্যা তাহার হৃদয়ের ধন; সন্ধ্যায় সে ভাবুক।

মেঘশূন্য নির্মল আকাশের তলে সন্ধ্যায় বসিয়া থাকিতে কি আরাম। আমাদের হৃদয়ের সমস্ত মালিন্য দূর হইয়া গিয়া সেও যেন আকাশের মত নির্মল হয়। আমাদের প্রাণ আনন্দ উথলিয়া উঠে—মনে হয় সাঁজের তারার সহিত মিলিয়া এক হইয়া যাই। সন্ধ্যা আমাদের মনের সঙ্গে খেলা করিতে বড় ভালবাসে। সে মানব হৃদয়ে নানারকম খেলার ভাব তুলিয়া দেয়; কিন্তু তাহার খেলাগুলি গাঞ্জীরমাখা, অথচ তাহার খেলায় গাঞ্জীরের মধ্য দিয়া যেন ফুটফুটে হাসি দেখা দেয়। সে হাসি গাঞ্জীরের মধ্য হইতে ফুটিয়া বাহির হয় তাহাই যথার্থ সুন্দর। এই কারণে সন্ধ্যার খেলা বড় সুন্দর, কেননা তাহার হাসি গাঞ্জীরের মধ্য হইতে বিকশিত হয়। আধার এইরূপ সুন্দর বলিয়াই সন্ধ্যা মানুষের প্রিয়।

সমস্ত গগন নীল। কেবল দূরে দূরে একটি শুভ্র মেঘের টুকরা ধীরে ধীরে ভাসিয়া বেড়াইতেছে। সন্ধ্যার এই দৃশ্যে কি গঞ্জীর ভাব, কি মধুর হাসি! মেঘগুলি যেন নীলিমার চাপা ঠোঁটের এক একটি হাসির টুকরা। তাহারা যেন চাপা থাকিতে না পারিয়া ফুটিয়া উঠিয়াছে। সন্ধ্যা এই হাসিগুলিকে সাজাইয়া দেয়—তাহাদের সৌন্দর্যকে আরও বৃদ্ধি করিয়া তুলে।

Figure 5: The binary image

Chapter 4: Skew angle detection and correction

As mentioned in chapter 2, most of the Bangla character has headline (matra) and so the skew angle can be detected using this matra. Some important statistics of Bangla language found from [2] and [3] are:

1. The average length of Bangla words is about six characters.
2. About 30%-35% of characters are vowel modifiers which, being small in size, contribute very little to the head line of the word.
3. Most basic characters are consonants, as vowels in basic form can appear at the beginning of the word or when two vowels appear side by side.
4. Compound characters are very infrequent, occurring in about 5% of the cases only.
5. In Bangla 41 characters can appear in the first position of a word. Out of these 41 characters 30 of them have head lines.
6. Probability of getting a character with head line in the first position of a word is: $P_1 = \frac{30}{41}$ and getting a character without head line in the first position is: $p_1' = 1 - P_1 = \frac{11}{41}$.
7. In other positions of a word, there are mostly consonants and 28 out of 39 Bangla consonants have head lines.
8. Probability of getting a consonant with head line for other positions in a word is: $P_2 = \frac{28}{39}$ and probability of getting a character without head line in other positions is: $p_2' = 1 - P_2 = \frac{11}{39}$.
9. Thus, probability of all four characters without head line in a word is $(1 - P_1)(1 - P_2)^3 = 0.00601$ (assuming that all characters are equally likely and independently occurring in a word). Hence, probability that a word will have at least one character with head line is $1 - 0.00601 = 0.99399$. The practical situation is better than these estimates since characters

are not equally likely in a word and most frequently used characters have head lines.

In Bangla, head line connects almost all characters in a word; therefore we can detect a word by the method of connected component labeling [5] (Glossary [a]). As mentioned in [2], for skew angle detection, at first the connected component labeling is done.

At the time of component labeling, for each labeled component its bounding box (minimum upright rectangle containing the component) is defined. The mean b_m and standard deviation b_s of the bounding box width are also computed. Next, components having boundary box width greater than or equal to b_m and less than $b_m + 3b_s$ are preserved. By threshold at b_m the small components like dots, punctuation marks, isolated characters and characters without head line are mostly filtered out while by threshold at $b_m + 3b_s$ big components that may represent graphs and tables are also filtered out (Figure 6). Because of these filtering processes the irrelevant components can not create error in skew estimation.

এই বিশ্ব জগতের স্তবে স্তবে কবিত্ব মাখানো চবাচব শত সহস্র ৩ স গ্রহ চন্দ্র সূর্য এক
 কটি কবিতা এই বিশ্ব জগ এই সকল ক্ষুদ্র ক্ষুদ্র মালা শত সহস্র উজ্জ্বল তাবকা এই মালায়
 গ্রথিত হইয়া শোভা ৩ —কত শত চন্দ্র সূর্য এই মালাব মধ্যে ফুলেব বিবাজ
 সূত্রেই চন্দ্র ত গাঁথা সমস্ত দিবস তেমন সুগন্ধ থাকে না সন্ধ্যায় তাহাব
 গন্ধ ৩ হয়— তাহাব হাসি ফুটিয়া উঠে সন্ধ্যাব ঝঞ্জিমচ্ছটায় মালাটি ৩ থাকে মনুষ্য হাসি
 লবাসে সুগন্ধও সে চায় তাই সে সন্ধ্যাব কোলে বসিতে চায় সন্ধ্যা তাহাব হৃদয়েব ধন সন্ধ্যায় সে ভাবুক
 নিমল ঢলে সন্ধ্যায় বসিয়া াক আবাম আমাদেব হৃদয়েব সমস্ত মালিন্য দূব
 হইয়া গিয়া সেও যেন আকাশেব মত নিমল হয় এ আনন্দ উথলিয়া উঠে—মনে ন্য সাজেব
 বাব সহিত মিলিয়া এক হইয়া যাই সন্ধ্যা আমাদেব মনেব সঙ্গে খল কবিত্তে বড লবাসে সে মানব হৃদয়ে
 খলাব ভাব তলিয় দেয় কিন্তু তাহাব খলাঢ়ি গঞ্জীয়মখ অ তাহাব খলায় গঞ্জীয়েব মধ্য দিয়া
 যন ফুটফুটে হাসি দেখা দেয় সে হাসি গঞ্জীয়ে মধ নই ফুটিয়া বাসিব হয় তাহাই যথার্থ সুন্দব এই
 কাব সন্ধ্যাব খলা বড সুন্দব কেনন তাহাব হাসি গঞ্জীয়েব মধ হইবে বিকশিত ন্য আ বা এইরূপ সুন্দব
 বলিয়াই সন্ধ্যা িষ
 সমস্ত গগন নীল কেবল দূব দূব একটি ত্র শোষণ টকব ধীবে বীবে তস্য বেড ৩ সন্ধ্যাব এই
 শে কি গঞ্জীব ভাব কি মবুব শসি শঘ লি যেন বীণিম চ াটেব এক একটি হাসিব টকবা তাশব
 ঘন চাপা থাকিতে ন পাবয ফুটিয় উঠয়াচ্ সন্ধ্যা এই শাস শিক সাশাহয দেয়—শাদেব
 আবও বৃদ্ধি কবিষা ঢলে

Figure 6: Selected connected components

Then *upper envelope* of the selected components is found as mentioned in [2]. From each pixel of the uppermost row of a bounding box, a vertical scan is performed until a pixel is found. The set of pixels obtained in this way denotes the upper envelope of the component. In figure 7 we show the Upper Envelope of figure 6. Note that in most of the cases the upper envelope contains the head line. In this way irrelevant data could be filtered out for further processing. Here Radon transform (glossary [b]) technique is applied on the upper envelopes for skew estimation.

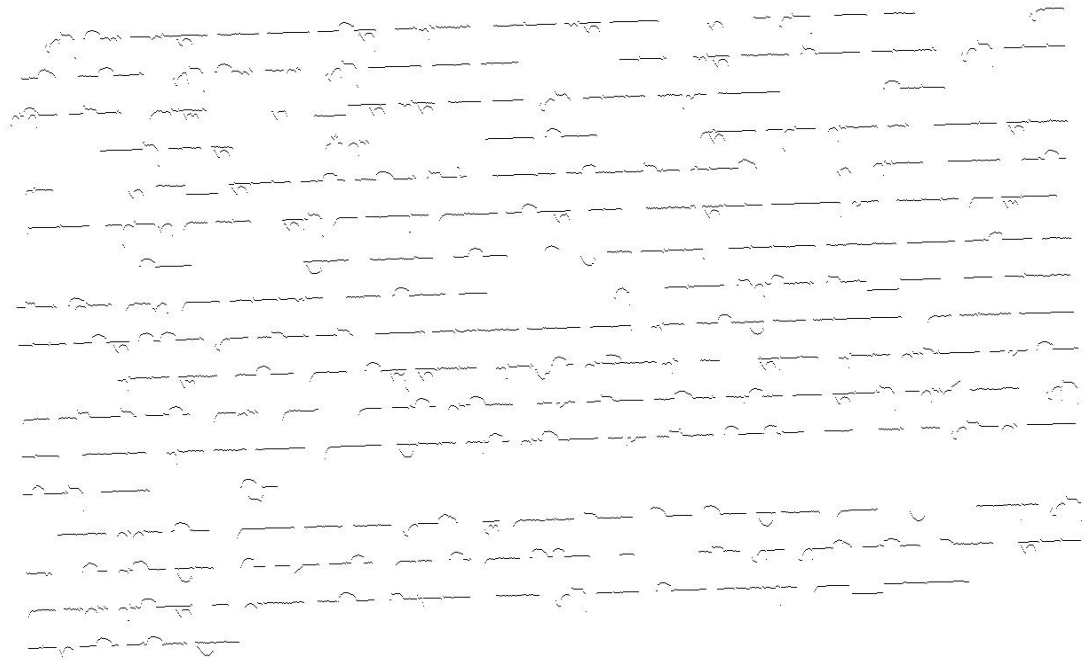


Figure 7: Upper Envelope

The Radon transform of a function $f(x,y)$ is defined as the integral along a straight line defined by its distance P from the origin and its angle of inclination θ , a definition very close to that of the Hough transform (glossary [c]) and requires a lot of processing power in order to be able to do its work in a reasonably finite time. Now-a-days high processing power is not a problem. All processor speed in the market is now more than one GHz and main memory is also very cheap. So, to use Radon Transform will not be a problem.

The Radon transform can detect a line in any angle. It is considered that all the line in a single image has same skew angle and the range of this angle is from -10° to 10° . Here Radon transform will detect the angle from the upper envelope. If the skewed angle is more than 10° or less then -10° the upper envelope will contain 2 lines in different direction. An example is shown in Figure 8 having 50 degree of skewed angle which may create a problem.



Figure 8: Upper envelope containing 2 lines in different direction

Radon transform gives the angles and distance of lines from the origin. The Radon transform of Upper Envelope (Figure 7) is shown in Figure 9.

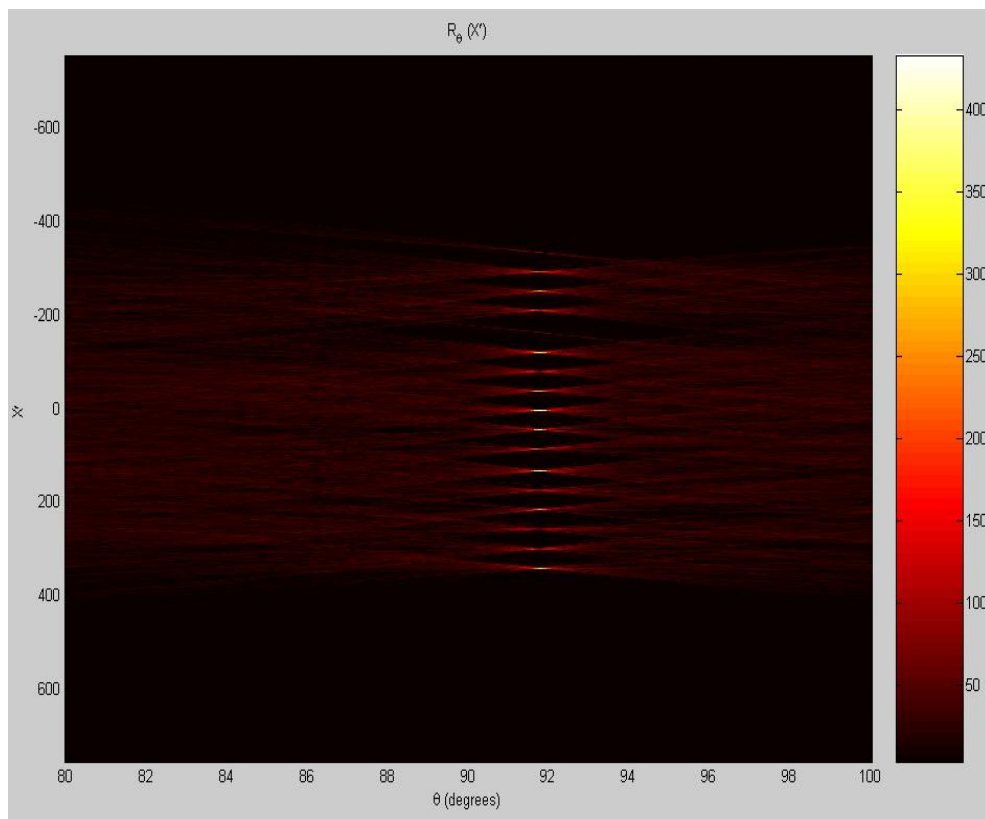


Figure 9: Radon transform of Upper Envelope (Figure 7)

In Figure 9, the point which has a high intensity represents the angle of a straight line. Here one can see that all the straight lines have the same degree of angle which is 91.8. To find the desired angle, this angle is subtracted from 90 degree and thus the desired skewed angle is $90 - 91.8 = -1.8$. So, if the

digital image (Figure 5) is rotated in degree -1.8 , then the desired de-skewed image (Figure 10) is found. It should be mentioned that up to one digit after decimal place has been considered.

এই বিশ্ব জগতের স্তরে স্তরে কবিত্ব মাখানো। চরাচর শত সহস্র কবিতার সংগ্রহ। চন্দ্র, সূর্য, তারকারা এক একটি কবিতা। এই বিশ্ব জগৎ এই সকল ক্ষুদ্র ক্ষুদ্র কবিতার মালা। শত সহস্র উজ্জ্বল তারকা এই মালায় ঋখিত হইয়া শোভা পাইতেছে—কত শত চন্দ্র-সূর্য এই মালার মধ্যে ফুলের আকারে বিরাজ করিতেছে। সৌন্দর্যের সূত্রেই চন্দ্র-তারকাবলী গাঁথা রহিয়াছে। সমস্ত দিবস মালাটির তেমন সুগন্ধ থাকে না। সন্ধ্যায় তাহার সুগন্ধ বিকশিত হয়— তাহার হাসি ফুটিয়া উঠে। সন্ধ্যায় বৃজ্জিমুচ্ছটায় মালাটি হাসিতে থাকে। মনুষ্য হাসি ভালবাসে, সুগন্ধও সে চায়। তাই সে সন্ধ্যায় কোলে বসিতে চায়। সন্ধ্যায় তাহার হৃদয়ের ধন; সন্ধ্যায় সে ভাবুক।

মেঘশূন্য নির্মল আকাশের তলে সন্ধ্যায় বসিয়া থাকিতে কি আরাম। আমাদের হৃদয়ের সমস্ত মালিন্য দূর হইয়া গিয়া সেও যেন আকাশের মত নির্মল হয়। আমাদের প্রাণ আনন্দ উথলিয়া উঠে—মনে হয় সঁজের ভার সহিত মিলিয়া এক হইয়া যাই। সন্ধ্যায় আমাদের মনের সঙ্গে খেলা করিতে বড় ভালবাসে। সে মানব হৃদয়ে নানারকম খেলার ভাব তুলিয়া দেয়; কিন্তু তাহার খেলাগুলি গাভীরমাখা, অথচ তাহার খেলায় গাভীরের মধ্য দিয়া যেন ফুটফুটে হাসি দেখা দেয়। সে হাসি গাভীরের মধ্য হইতে ফুটিয়া বাহির হয় তাহাই যথার্থ সুন্দর। এই কারণে সন্ধ্যায় খেলা বড় সুন্দর, কেননা তাহার হাসি গাভীরের মধ্য হইতে বিকশিত হয়। আঁধার এইরূপ সুন্দর বলিয়াই সন্ধ্যায় মানুষের প্রিয়।

সমস্ত গগন নীল। কেবল দূরে দূরে একটি শুভ্র মেঘের টুকরা ধীরে ধীরে ভাসিয়া বেড়াইতেছে। সন্ধ্যায় এই দৃশ্যে কি গভীর ভাব, কি মধুর হাসি! মেঘগুলি যেন নীলিমার চাপা ঠোঁটের এক একটি হাসির টুকরা। তাহারা যেন চাপা থাকিতে না পারিয়া ফুটিয়া উঠিয়াছে। সন্ধ্যায় এই হাসিগুলিকে সাজাইয়া দেয়—তাহাদের সৌন্দর্যকে আরও বৃদ্ধি করিয়া তুলে।

Figure 10: After correction the skew angle

Chapter 5: Line and word separation

Before separation of lines, some noise must be removed. There are lots of dots in the image (figure10) which may create problem. When removing a dot, one should remember that there are some bangle characters that contain dot. Figure 11 shows such characters. For this it is very difficult to locate the noise dots.

i q o

Figure 11: Dots in bangle character

5.1 Noise removal

Generally a noise dot is very small in size. Here a dot having size 4 by 4 or less is removed. The image after removal of small dots is shown in figure 12.

এই বিশ্ব জগতের স্তরে স্তরে কবিত্ব মাখানো। চবাচর শত সহস্র কবিতাব সংগ্রহ। চন্দ্র, সূর্য, তাবকারা এক একটি কবিতা। এই বিশ্ব জগৎ এই সকল ক্ষুদ্র ক্ষুদ্র কবিতার মালা। শত সহস্র উজ্জ্বল তারকা এই মালায় গ্রথিত হইয়া শোভা পাইতেছে—কত শত চন্দ্র-সূর্য এই মালার মধ্যে ফুলের আকাবে বিরাজ করিতেছে। সৌন্দর্যের সূত্রেই চন্দ্র-তাবকাবলী গাঁথা রহিয়াছে। সমস্ত দিবস মালাটির তেমন সুগন্ধ থাকে না। সন্ধ্যায় তাহার সুগন্ধ বিকশিত হয়— তাহার হাসি ফুটিয়া উঠে। সন্ধ্যাব বৃজিমুচ্ছটায় মালাটি হাসিতে থাকে। মনুষ্য হাসি ভালবাসে, সুগন্ধও সে চায়। তাই সে সন্ধ্যার কোলে বসিতে চায়। সন্ধ্যা তাহাব হৃদয়ের ধন, সন্ধ্যায় সে ভাবুক।

মেঘশূন্য নির্মল আকাশের তলে সন্ধ্যায় বসিয়া থাকিতে কি আবাম। আমাদের হৃদয়ের সমস্ত মালিন্য দূব হইয়া গিয়া সেও যেন আকাশের মত নির্মল হয়। আমাদের প্রাণ আনন্দ উথলিয়া উঠে—মনে হয় সাজের ভাবার সহিত মিলিয়া এক হইয়া যাই। সন্ধ্যা আমাদের মনের সঙ্গে খেলা কবিত্তে বড় ভালবাসে। সে মানব হৃদয়ে নানাবকম খেলার ভাব তুলিয়া দেয়, কিন্তু তাহাব খেলাগুলি গাঞ্জীর্যমাখা, অথচ তাহাব খেলায় গাঞ্জীর্যের মধ্য দিয়া যেন ফুটফুটে হাসি দেখা দেয়। সে হাসি গাঞ্জীর্যের মধ্য হইতে ফুটিয়া বাহির হয় তাহাই যথার্থ সুন্দর। এই কারণে সন্ধ্যার খেলা বড় সুন্দর, কেননা তাহাব হাসি গাঞ্জীর্যের মধ্য হইতে বিকশিত হয়। আধার এইরূপ সুন্দর বলিয়াই সন্ধ্যা মানুষের প্রিয়।

সমস্ত গগন নীল। কেবল দূরে দূবে একটি শুভ্র মেঘের টুকরা ধীবে ধীবে ভাসিয়া বেড়াইতেছে। সন্ধ্যাব এই দৃশ্যে কি গঞ্জীর ভাব, কি মধুর হাসি! মেঘগুলি যেন নীলিমাব চাপা ঠোঁটের এক একটি হাসির টুকরা। তাহারা যেন চাপা থাকিতে না পারিয়া ফুটিয়া উঠিয়াছে। সন্ধ্যা এই হাসিগুলিকে সাজাইয়া দেয়—তাহাদের সৌন্দর্যকে আবও বৃদ্ধি করিয়া তুলে।

Figure 12: After remove dots having size 4 by 4 or less

In this image there are still some noise dots having size more. If these dots are removed, then some character will be also be lost (e.g. coma, colon etc.).

5.2 Line Segmentation

There are free spaces between every two lines. So lines can be separated calculating these spaces. If all the black pixels in a single row (i.e. horizontally) are counted and then plotted (figure 13) one can easily observed it. The peak position is denoted the headline of a line.

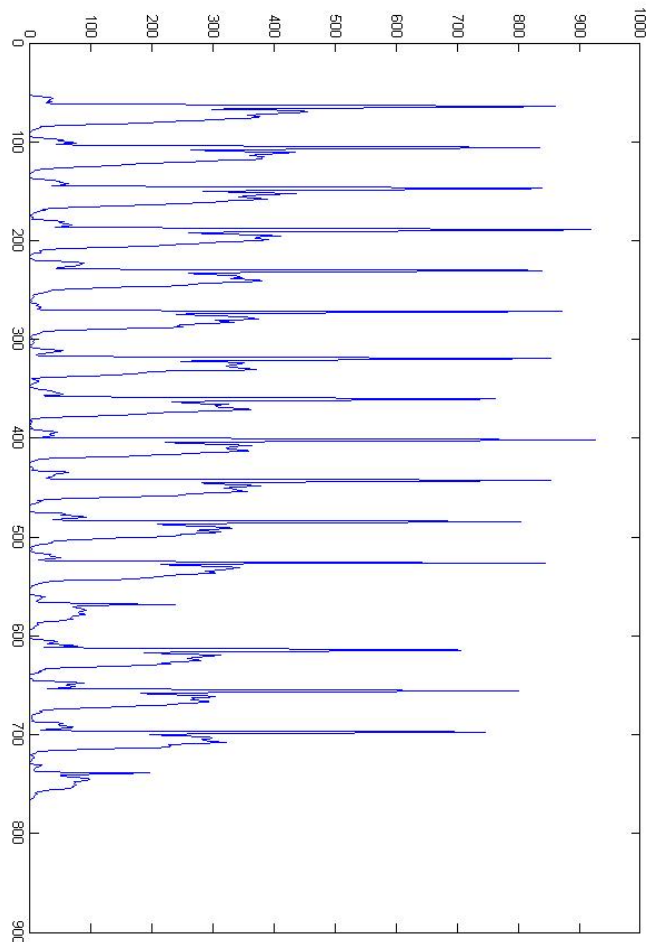


Figure 13: Plot of horizontal Black pixel of the image

After separation of lines, next task is to separate the words. Figure 14 shows 1st line of the page.

এই বিশ্ব জগতের স্তরে স্তবে কবিত্ব মাখানো। চবাচর শত সহস্র কবিতাব সংগ্রহ। চন্দ্র, সূর্য, তাবকারা এক

Figure 14: Line 1

5.3 Word Segmentation

The words are also separated by space. Again there are cases where some spaces exist in a single word. So this has to be kept in mind when words are separated using space. If all the black pixels are counted column wise (i.e. vertically) and plotted (figure 15) one can observe it.

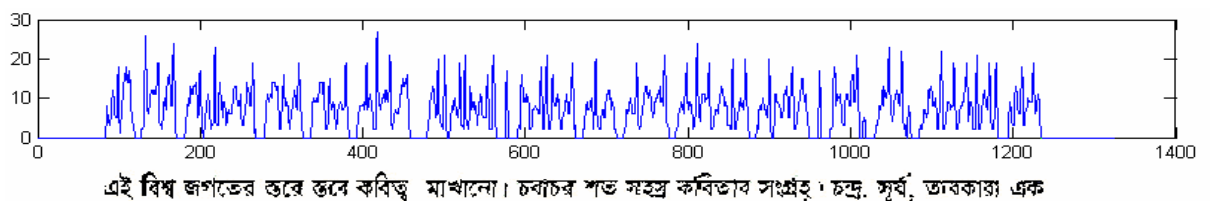


Figure 15: Plot of Vertical Black pixel of line 1

To separate a word there must be at least 6 consecutive columns having 0 black pixels. Here the 1st word is in figure 16.



Figure 16: First word of line 1

5.4 Character segmentation

This is one of the most challenging parts in pre-processing steps. Generally characters are connected in a word by the head line (matra). Figure 17 shows some example of word.



(a)



(b)



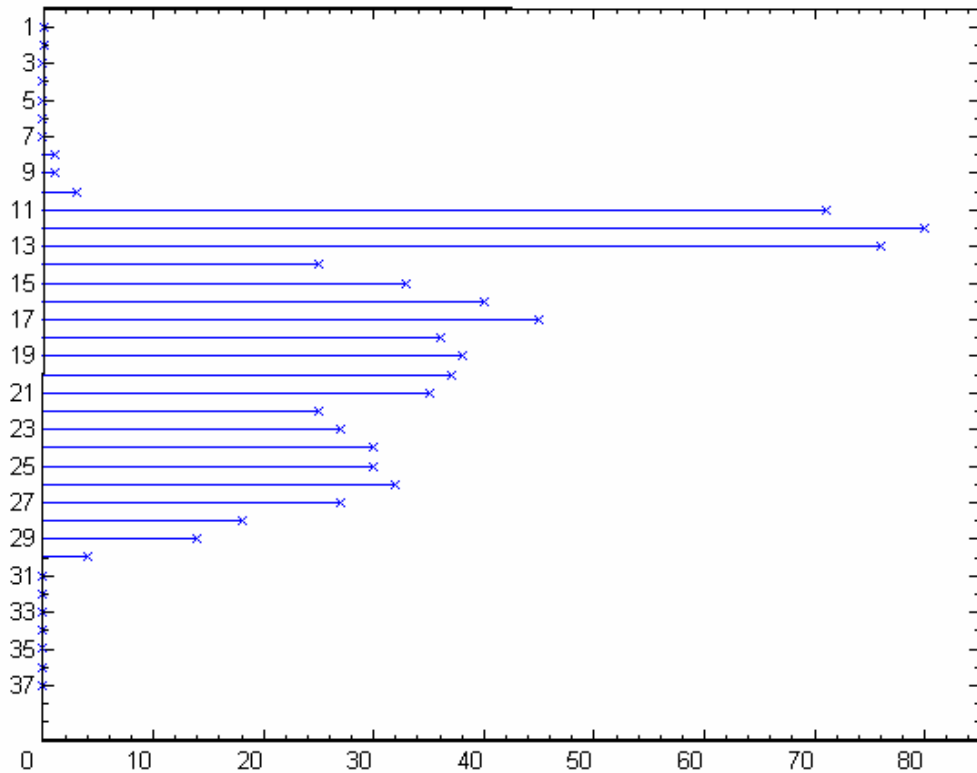
(c)



(d)

Figure 17: example of some words

To segment characters, first the headline has to be removed. If black pixels are counted row wise of Figure 17(b) i.e. horizontally and plotted (Figure 18) then the location of headline can be found.

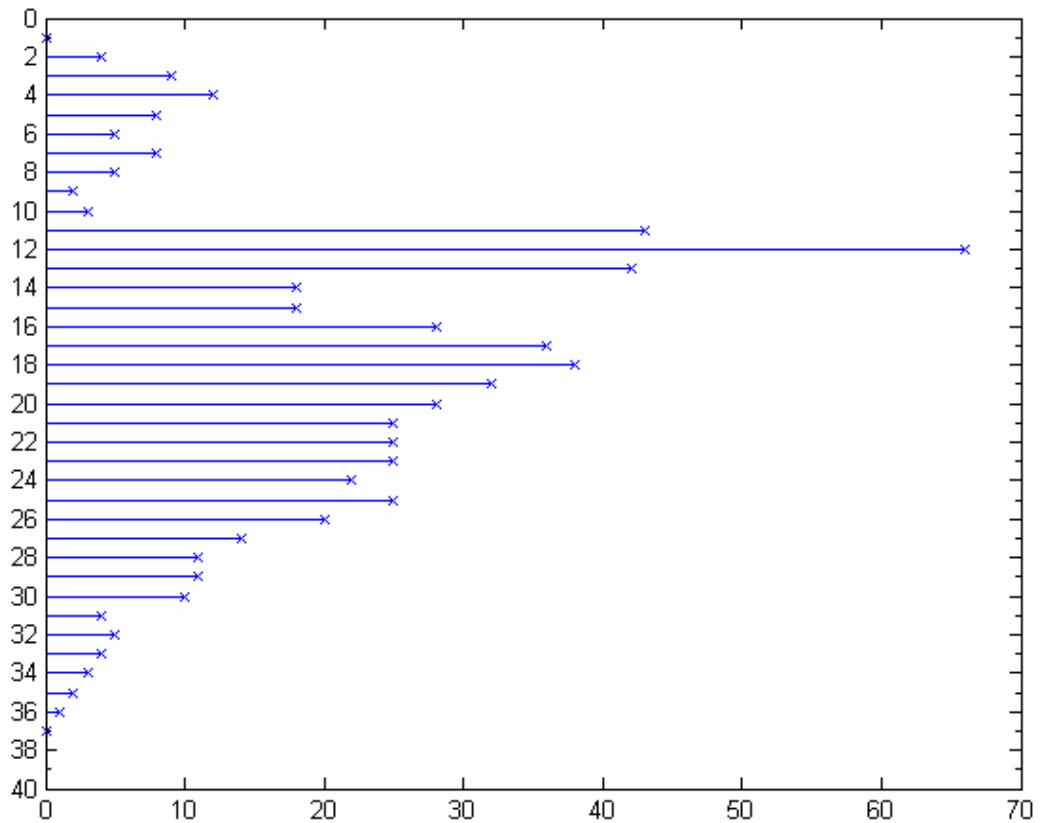


জগতের

Figure 18: Plot of horizontal Black pixel of word figure 17(b)

From this plot the maximum black pixel is found on row 12. Row 11 and 13 have black pixel very near to maximum value. For this one can say row 11 and row 13 are also in the headline. So to remove the headline one has to remove these three rows.

But in figure 19, the situation is different. Figure 19 plots black pixel of Figure 17(d) i.e. horizontally row wise.



কবিত্ত

Figure 19: Plot of horizontal Black pixel of word figure 17(d)

In figure 19, row 12 has the maximum black pixel, but row 11 and 13 neither has black pixel very near to maximum value. For this one cannot say that row 11 and 13 are in the headline. Again we cannot say that the width of headline is three. The width of the headline is variable because of print style (font size).

By using some Morphological operation⁴, one can overcome this problem. Here two morphological operations has been tried, they are thinning and skeletonization operations. Thinning removes pixels so that an object without holes shrinks to a minimally connected stroke, and an object with holes

⁴ Morphological operation is a collection of techniques for digital image processing. Such techniques include closing, shrinking, thinning, thickening, skeletonization, pruning, etc.

shrinks to a connected ring halfway between each hole and the outer boundary. This option preserves the Euler number⁵. And skeletonization removes pixels on the boundaries of objects but does not allow objects to break apart. The pixels remaining make up the image skeleton. This option also preserves the Euler number.

The figure 20 shows the example of thinning (thin) and skeletonization (skel) of word figure 17(a).

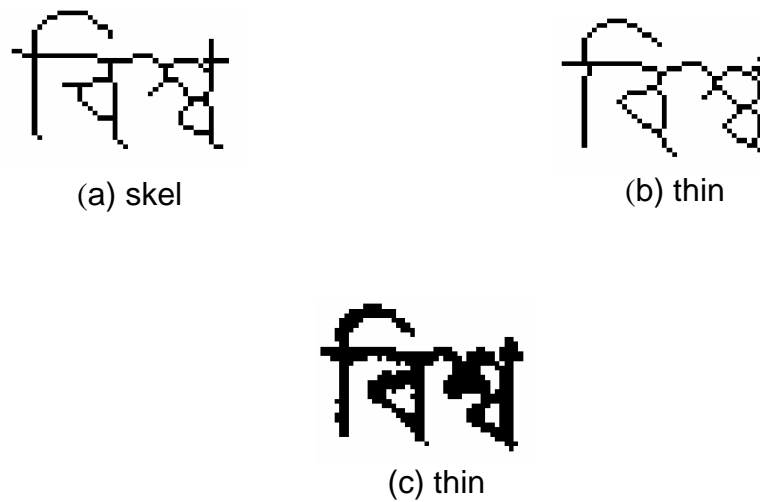


Figure 20: (a) and (b) respectively shows skel and thin of (c)

Here skel detects headline better than thin but character is better in thin. Again if one observes another example in figure 21, which is the example of thin and skel of word figure 17(d).

⁵ Defined the number of components minus the number of holes.

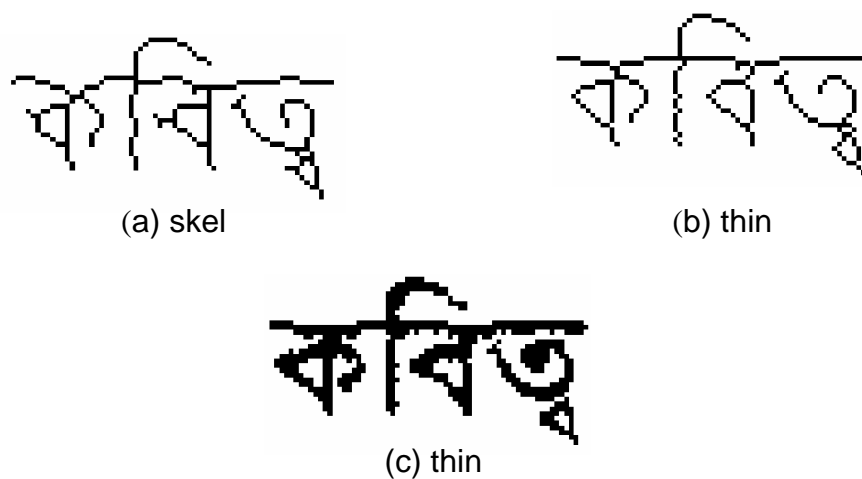


Figure 21: (a) and (b) respectively shows skel and thin of (c)

Here thin detects headline better than skel and also character is better in thin. Figure 22 shows the image after removing the headline of thin image.

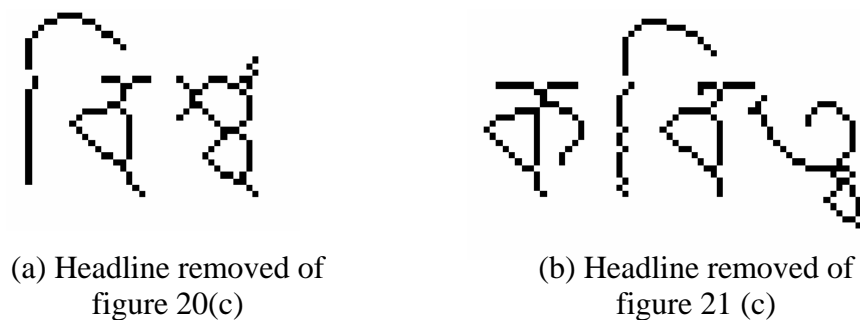


Figure 22: After removal of headline of thin words

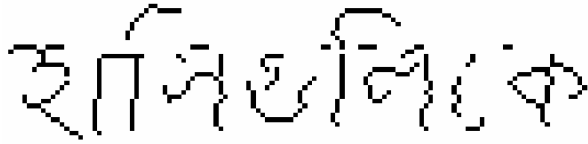
Character separation can be done by connected component labeling because all the pixel of a character is connected. But in cases the character itself broke. Figure 23 shows an example where we cannot use connected component labeling.



(a) Word from Image



(b) Thin word of (a)



(c) Headline removed of thin word

Figure 23: Example of character not connected

Again there are some cases where two characters are connected. Figure 24 shows an example.



(a) Word from Image



(b) Thin word of (a)



(c) Headline removed of thin word

Figure 24: Example of joined character

It is very difficult to perfectly segment the character of figure 23 and 24. These problems are absent for font base computer compose image. Figure 25 shows an example where the font name is SutonnyMJ.



Figure 25: Example of Font based word

From font base word character separation is easier than print based. Character can be separated by using connected components which is considered as input of recognition step.

Chapter 6: Character Recognition

The whole character itself is used as the neural network input. The size (25 by 25) of the input image is kept fixed.

Architecture of the network is as follows:

- Input layer: 625 neurons
- 1st Hidden layer: 100 neurons
- 2nd Hidden layer: 50 neurons
- Output layer: 9 neurons (for each character)
- Transfer function: log-sigmoid
- Incremental training algorithm, standard back-propagation method

Details of this part can be found on the thesis report of my group member.

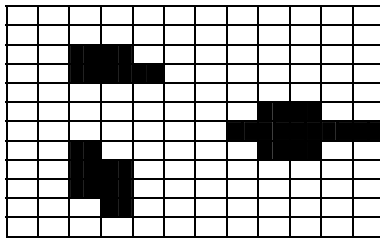
Chapter 7: Conclusion

This thesis tries to suggest an approach for the recognition of Bangla scripts. Approaches suggested from the beginning of scanning a document to converting it to binary image, skew detection and correction, line separation and word segmentation has been successfully stated. One of the challenges faced in the character segmentation part is that two characters are sometimes joined together. There are even cases where a single character breaks apart. Solutions to these challenges are likely to be presented in future. In our current approach, the whole character itself was used as a feature. In future implementation feature extraction will be more comprehensive.

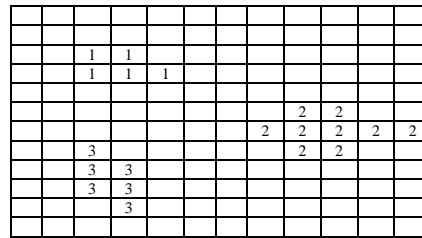
Glossary

[a] Connected Components Labeling

A set of pixels in which each pixel is connected to all other pixel is called a connected component. A component labeling algorithm finds all connected components in an image and assign a unique label to all points in the same component.



(a)



(b)

Sequential Connected Components Algorithm 8-connectivity

Step 1: Scan image left to right, top to bottom.

Step 2: If the pixel is black

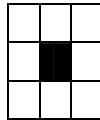
- (a) If only one of its upper-left or upper or upper-right or left neighbors has a label, then copy the label.
- (b) If both have the same label, then copy the label
- (c) If both have labels and they are not same, then copy the lowest labels of its neighbors and enter the labels in the equivalence table as equivalent labels.
- (d) Otherwise assign a new label to this pixel and enter this label in the equivalence table.

Step 3: If there are more pixels to consider, then go to step 2.

Step 4: Find the lowest label for each equivalent set in the equivalence table.

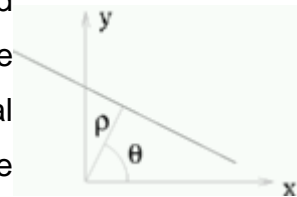
Step 5: Scan the picture. Replace each label by the lowest label in its equivalent set.

Here 8-connectivity means there are 8 neighbored pixel of a particular pixel shown as image below:



[b] The Radon Transform

In recent years the Hough transform and the related Radon transform have received much attention. These two transforms are able to transform two dimensional images with lines into a domain of possible line parameters, where each line in the image will give a peak positioned at the corresponding line parameters. This has lead to many line detection applications within image processing, computer vision, and seismic.



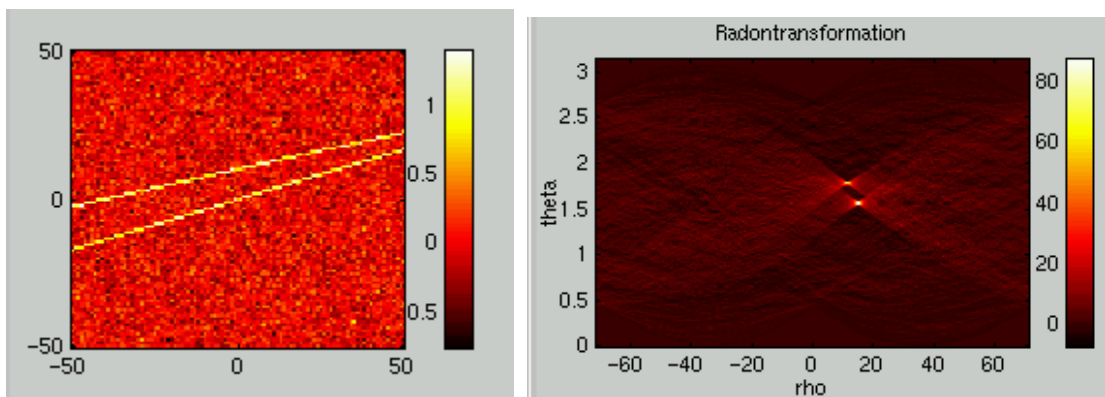
Several definitions of the Radon transform exists, but the are related, and a very popular form expresses lines in the form $\rho = x \cos \theta + y \sin \theta$, where θ is the angle and ρ the smallest distance to the origin of the coordinate system. As shown in the two following definitions (which are identical), the Radon transform for a set of parameters (ρ, θ) is the line integral through the image $g(x, y)$, where the line is positioned corresponding to the value of (ρ, θ) . The $\delta(\cdot)$ is the delta function which is infinite for argument 0 and zero for all other arguments (it integrates to one).

$$\check{g}(\rho, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \delta(\rho - x \cos \theta - y \sin \theta) dx dy$$

or the identical expression

$$\check{g}(\rho, \theta) = \int_{-\infty}^{\infty} g(\rho \cos \theta - s \sin \theta, \rho \sin \theta + s \cos \theta) ds$$

Using this definition an image containing two lines are transformed into the Radon transform shown to the right



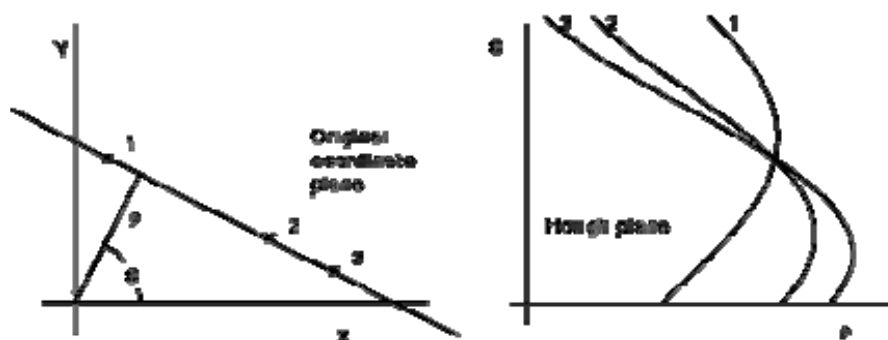
It can be seen that two very bright spots are found in the Radon transform, and the positions shown the parameters of the lines in the original image. A simple threshold algorithm could then be used to pick out the line parameters, and given that the transform is linear many lines will just give rise to a set of distinct point in the Radon domain. In Ph.D. thesis [3], Peter Toft investigated the relationship of Radon transform with the Hough transform, and it is shown that the Radon transform and the Hough transform are related but NOT the same.

[c] Hough transform:

The Hough transform is a standard tool in image analysis that allows recognition of global patterns in an image space by recognition of local patterns (ideally a point) in a transformed parameter space. It is particularly useful when the patterns one is looking for are sparsely digitized, have "holes" and/or the pictures are noisy.

The basic idea of this technique is to find curves that can be parameterized like straight lines, polynomials, circles, etc., in a suitable parameter space. Although the transform can be used in higher dimensions the main use is in two dimensions to find, e.g. straight lines, centers of circles with a fixed radius, parabolas $y = ax^2 + bx + c$ with constant c , etc.

As an example consider the detection of straight lines in an image. We assume them parameterized in the form: $\rho = x \cos \theta + y \sin \theta$, where ρ is the perpendicular distance from the origin and θ the angle with the normal. Collinear points (x_i, y_i) , with $i=1, \dots, N$, are transformed into N sinusoidal curves $\rho = x_i \cos \theta + y_i \sin \theta$ in the (ρ, θ) plane, which intersect in the point (ρ, θ) .



Care has to be taken when one quantizes the parameter space (ρ, θ) . When the bins of the (ρ, θ) space (it is easy to visualize the transform as a two-

dimensional histogram) are chosen too fine, each intersection of two sinusoidal curves can be in a different bin. When the quantization is not fine enough, on the other hand, nearly parallel lines which are close together will lie in the same bin.

For a certain range of quantized values of parameters ρ and θ , each (x_i, y_i) is mapped into the (ρ, θ) space and the points that map into the locations (ρ_m, θ_m) are accumulated in the two-dimensional histogram, IHIST (ρ_m, θ_m) , i.e. $\text{IHIST}(\rho_m, \theta_m) = \text{IHIST}(\rho_m, \theta_m) + 1$.

If a grey level image $g(x, y)$ is given, and g_i is the grey value at the point (x_i, y_i) the grey values are accumulated: $\text{IHIST}(\rho_m, \theta_m) = \text{IHIST}(\rho_m, \theta_m) + g_i$.

References

Papers presented at conferences (Published):

- [1] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979.

- [2] B.B. Chaudhuri and U. Pal, "Skew Angle Detection of Digitized Indian Script Documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, February 1997

- [3] B.B. Chaudhuri and U. Pal, "Relational Studies Between Phoneme and Grapheme Statistics in Modern *Bangla* Language," *J. Acoustical Society of India*, vol. 23, pp. 67-77, 1995.

Thesis:

- [4] Peter Toft: "The Radon Transform - Theory and Implementation", Ph.D. thesis. Department of Mathematical Modelling, Technical University of Denmark, June 1996. 326 pages.

Books:

- [5] Ramesh Jain and Rangachar Kasturi and Brian G. Schunck, Machine Vision, International Edition 1995 McGraw-Hill, Inc.