Shahjalal University of Science and Technology
Department of Computer Science and Engineering

Document Decomposition System for Bangla OCR

Biswajit Debnath, Reg. No: 2011331009, Dept. of CSE
Md. Sajid Shahriar, Reg. No: 2011331048, Dept. of CSE

Adviser:
Dr. Muhammed Zafar Iqbal, Professor, Dept. of CSE

Co-Adviser:
Mr. Sabir Ismail, Assistant Professor, Dept. of CSE
Mr. Md. Saiful Islam, Assistant Professor, Dept. of CSE

18th October, 2016

Document Decomposition System for Bangla OCR

A Thesis submitted to the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

Name: Biswajit Debnath, Reg. No: 2011331009, Dept. of CSE
Name: Md. Sajid Shahriar, Reg. No: 2011331048, Dept. of CSE

Adviser:
Dr. Muhammed Zafar Iqbal, Professor, Dept. of CSE

Co-Adviser:
Mr. Sabir Ismail, Assistant Professor, Dept. of CSE
Mr. Md. Saiful Islam, Assistant Professor, Dept. of CSE

18th October, 2016

Recommendation Letter from Thesis Adviser

The Thesis

entitled "**Document Decomposition System for Bangla OCR**"

submitted by the students

1. **Biswajit Debnath**
2. **Md. Sajid Shahriar**

is a record of research work carried out under my supervision and I, hereby, approve that the report be submitted in partial fulfillment of the requirements for the award of their Bachelor Degrees.

_____

Adviser : Dr. Muhammed Zafar Iqbal, Professor, Dept. of CSE
Date : / 10 / 2016

_____

Co-Adviser : Mr. Sabir Ismail, Assistant Professor, Dept. of CSE
Date : / 10 / 2016

_____

Co-Adviser : Mr. Md. Saiful Islam, Assistant Professor, Dept. of CSE
Date : / 10 / 2016

Certificate of Acceptance of the Thesis

The Thesis

entitled "**Document Decomposition System for Bangla OCR**"

submitted by the students

1. **Biswajit Debnath**
2. **Md. Sajid Shahriar**

on _____

is, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

_____	_____	_____
Head of the Department	Chairman, Thesis Committee	Adviser
Dr. Md. Reza Selim	Dr. Md. Reza Selim	Dr. Muhammed Zafar Iqbal
Professor, Dept. of CSE	Professor, Dept. of CSE	Professor, Dept. of CSE

ABSTRACT

Resources like newspaper or printed documents contain important contents. Content Digitizing is a possible solution of protecting and retrieving the endangered documents. Optical Character Recognition or in short OCR is the best option or process for digitizing these printed documents. Some pre-processing procedures like Segmentation or Decomposition help in Optical Character Recognition methodology. By following these pre-processing methodology, textual part of the document can be identified. These textual parts are the main component to document digitize. Our work is based on segmenting the Bangla Documents. Our goal was to decomposing the contents from a document along with de-skewing it, retrieving the tabular data and rebuilding it. The methodology we implemented work 96.88% perfectly.

Keyword: Bangla Document, Digitizing, OCR, Segmentation, Decomposition, DE skew, Tabular Data.

Acknowledgements

This thesis work and paper is made possible through the help and support from all sentient beings. Especially, please allow us to dedicate my acknowledgment of gratitude toward the following significant advisers, co-advisers and contributors who was with us throughout the whole time:

First and foremost, we would like to thank to our Adviser Dr. Muhammed Zafar Iqbal for his most support and encouragement. His comments, remarks, engagement, guidance and constant supervision as well as for providing necessary information regarding the thesis was really helpful to us to do the analysis and finding out what to do.

Second, we would like to thank our Co-Adviser Mr. Sabir Ismail and Mr. Md. Saiful Saif to advise us about the thesis and to provide valuable information which was really help to accelerate our work.

Finally, we sincerely thank to our seniors Fahad Hasan and Tasmin Afroz Tannee. Who previously contributed with this work and after that help us to catch up with the work.

Contents

List of Figures

Chapter 1

Introduction

Humans like to dream big and their dreams are persistently something critical like machine replication of human limits, including reading, composing and so forth. From these dreams, machine reading has created into reality all through the latest five decades. Optical Character Recognition has wound up a champion among the best applications of advancement in the field of Computer Vision[1] [1].

1.1 Meaning of Optical Character Recognition

Optical character recognition (optical character reader) (OCR) is the mechanical or electronic transformation of images/pictures which may be as typed, handwritten or printed text into machine-encoded text [2]. For data digitization OCR is now the most common method for printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as Machine Translation[2], Speech Synthesis, Key Data and Text Mining [2]. It is a field of research in Pattern Recognition[3], Artificial Intelligence[4] and Computer Vision. [2].

[1] Computer Vision - A field that includes methods for acquiring, processing, analyzing, and understanding images.

[2] Machine Translation - Performs simple substitution of words in one language for words in another.

[3] Pattern Recognition - Focuses on the recognition of patterns and regularities in data.

[4] Artificial Intelligence - It is the intelligence exhibited by machines or software.

It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation.

1.2 Process of an Optical Character Recognition System

Optical Character Recognition is mainly an "offline" process, in which is it performs some analysis on a static document. The offline work is performed after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents.

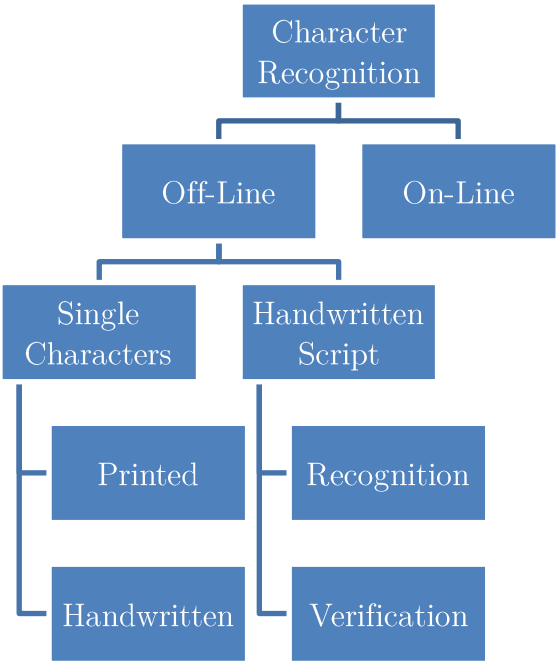


Figure 1 The different areas of character recognition *[1]*.

The software made for OCR often follows three techniques or methods before giving the final output. It deals with the problem of recognizing optically processed characters.

1.2.1 Preprocessing

- **De-skew** - Aligning the document properly.
- **Binarization** - Convert the image from color or grayscale to black and white.
- **Layout Analysis** - Identifies columns, paragraph, captions, images etc.
- **Segmentation** - Multiple characters that are connected in the image artifacts must be separated.

1.2.2 Character Recognition

- An image is stored on pixel by pixel basis.
- Input glyph being isolated from the rest of the image then used pattern matching techniques for recognition.

1.2.3 Post Processing

- **Grouping** - Merging the symbols output to string.
- **Error Analysis** - Error detection and correction

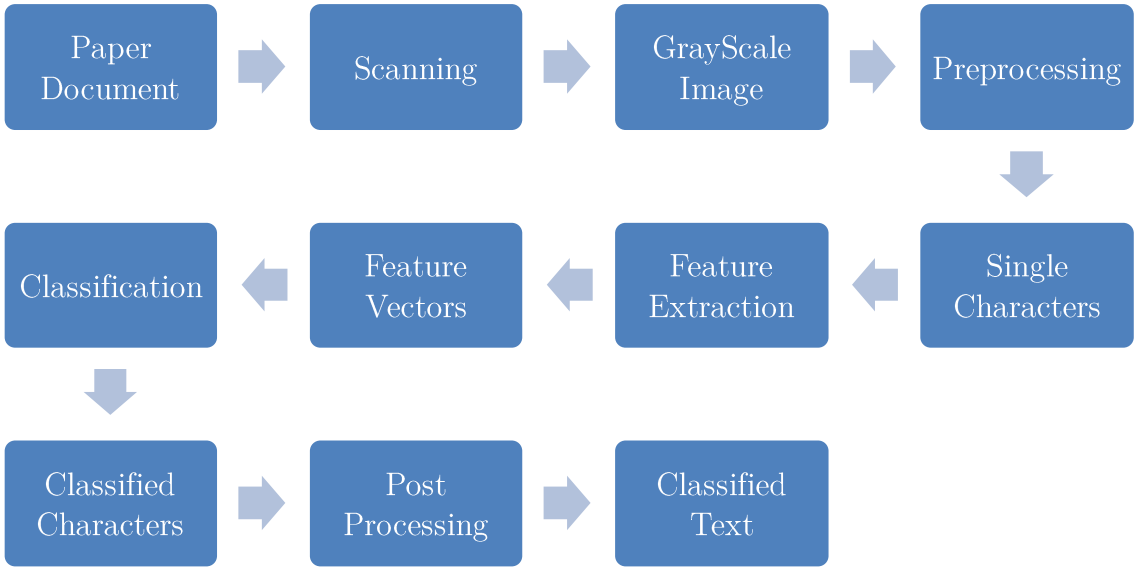


Figure 2 Steps of Optical Character Recognition.

1.3 Importance of Optical Character Recognition

If we follow the very traditional way for digitizing the data, it will need a huge manpower, because we have to create each digital document by giving the inputs manually. For cases like these, automatic identification may be an alternative and OCR is the best alternative of that process. OCR converts normal scanned documents to text-searchable, so it allow us to search content on the same. A scanned file is usually stored as an image if not converted into a text-searchable file with the help of OCR technology. Once that is done, the document allows the search engine to perform content search inside the file and retrieve appropriate results. OCR also can identify any kind of text inside the document is wrong.

Chapter 2

Preprocessing

Preprocessing the most crucial part to start the digitizing the scanned or hand written document data. Process like De-skew, Binarization, Layout-Analysis are included in preprocessing.

2.1 What is preprocessing in OCR

Technically documents like newspaper not only contains just some textual data but also it is populated with figures, tabular datasheets, and images may contain many other contents. Along with this kind of component variation scanned document could be skewed. Sometimes document may contain unnecessary points which could create a messy result.

যশোরের পাখির বাসা যায় ইউরোপে

মনিরুল ইসলাম, যশোর

যশোরে তৈরি শৌখিন পাখির বাসা ইউরোপের ছয় দেশে রপ্তানি হচ্ছে। এতে বছরে কোটি টাকার সমপরিমাণ বৈদেশিক মুদ্রা আসছে দেশে। এর মাধ্যমে সরাসরী হয়েছে এ অঞ্চলের হাজারো নিম্ন আয়ের মানুষ।

ইউরোপের জার্মানি, ফ্রান্স, বেলজিয়াম, নেদারল্যান্ডস, স্পেন ও পর্তুগালের বিভিন্ন শহরে শৌখিন পাখি উৎপাদনের খামারে যাচ্ছে ৪০ ধরনের পাখির বাসা। তবে পুঁজির সংকটের কারণে এ পণ্যের রপ্তানি বাণিজ্যে যতটা প্রসার ঘটার কথা ছিল, ততটা ঘটেনি বলে দাবি উদ্যোক্তাদের।

এ পণ্য রপ্তানির অন্যতম উদ্যোক্তা ঢাকার সিদ্ধেশ্বরী এলাকার বাসিন্দা খায়রুল আলম। তিনি যশোরের কয়েকটি গ্রামের নিম্ন আয়ের মানুষের মাধ্যমে পাখির বাসা তৈরি করান। পরে এগুলো রপ্তানি উপযোগী করে নৌপথে বিদেশে পাঠান।

জানতে চাইলে খায়রুল আলম বলেন, যশোর সদর উপজেলার আবাদ কুড়া, সীতারামপুর ও বাহাদুরপুর গ্রামে তৈরি শৌখিন পাখির বাসার ইউরোপের বাজারে ব্যাপক চাহিদা। অন্তত ৪০ ধরনের পাখির বাসা ছয়টি দেশে এখন রপ্তানি করা হচ্ছে। বর্তমানে বছরে প্রায় ১ কোটি টাকা মূল্যের পাখির বাসা রপ্তানি হচ্ছে।'

সম্প্রতি আবাদ কুড়া ও সীতারামপুর গ্রামে গিয়ে দেখা গেছে, ঘরে ঘরে পাখির বাসা তৈরির কাজ চলছে। ঘরের বারান্দা ও আঙিনায় বসে নারী-পুরুষ মিলে পাখির বাসা বুননের কাজ করছেন। পুরুষেরা বাঁশের চাটাই দিয়ে বুননের মূল উপকরণ তৈরি করছেন। নারীরা বাসা তৈরির জো (কাজের শুরু) তুলছেন। আরেকজন বাসা তৈরির কাজ শেষ করছেন।

আবাদ কুড়া গ্রামের দাসপাড়ার কিনারাম দাসের বাড়ির উঠানে গিয়ে দেখা গেল, তিনি তলতা বাঁশের চাটাই চিরে দা দিয়ে সূচারু করছেন। পাশে বসে তাঁর স্ত্রী কাঞ্চন দাস ও মা রঙ্গিতা দাস বাসা বুননের জো তুলে দিচ্ছেন।

জার্মানি, ফ্রান্স, বেলজিয়াম, নেদারল্যান্ডস, স্পেন ও পর্তুগালে যাচ্ছে ৪০ ধরনের পাখির বাসা

পুঁজির সংকটের কারণে এ পণ্যের রপ্তানি বাণিজ্যে যতটা প্রসার ঘটার কথা ছিল, ততটা ঘটেনি বলে দাবি উদ্যোক্তাদের

কাজের ফাঁকে কিনারাম দাস বলেন, '৩০ বছর ধরে পাখির বাসা তৈরির কাজ করছি। ৪০ ধরনের বাসা বানাতে পারি। বাঁশ, পাট, কাতা (নারকেলের ছোবড়া), খড়, বিচালি, বাঁশের পাতা, কাঠ—এ-জাতীয় নানা উপাদান দিয়ে এ বাসা তৈরি করা হয়। যখন যে ধরনের অর্ডার থাকে, তখন সেই ধরনের বাসা তৈরি করা হয়।'

কিনারাম বলেন, 'বড় ধরনের বাসার জন্য ৭ টাকা, মাঝারি ও ছোট বাসার ক্ষেত্রে ৩ টাকা করে মজুরি দেওয়া হয়। তিনজন মিলে কাজ করে মাসে ১ হাজার ৫০০ থেকে ২ হাজার ৬০০ খণ্ড বাসা তৈরি করা যায়। মজুরি কম হলেও কাজটি বাড়িতে বসেই করা যায়। নারীরাও তা করতে পারে। তবে বাঁশের দাম এখন বেড়েছে। আমরা মজুরি বাড়ানোর জন্য দাবি তুলেছি।'

এ কাজ করে কিনারাম সংসার চালিয়েও পাকা বাড়ি নির্মাণ করেছেন। আগে তাঁরা ছাপরাঘরে থাকতেন। এখন পাকা বাড়িতে থাকেন, এটাই তাঁদের আনন্দ। এ শিল্পের সঙ্গে সম্পৃক্ত প্রায় সবাই টিনের ছাউনির পাকা বাড়ি করেছেন।

পার্শ্ববর্তী সীতারামপুর গ্রামের ঋষিপাড়ার চণ্ডীদাস ৩০ বছর আগে প্রথম পাখির বাসা তৈরির কাজ শুরু করেছিলেন।

সীতারামপুর গ্রামে চণ্ডীদাসের বাড়িতে গিয়ে দেখা গেল, চণ্ডীদাস ও তাঁর মেয়ে জামাই আরনল্ড বিশ্বাস নারকেল ও পাট দিয়ে বারুই পাখির বাসার মতো বাসা তৈরি করছেন। আরনল্ড বিশ্বাস বলেন, 'উচ্চমাধ্যমিক পাস করে অন্য চাকরিতে যাইনি। ছোটবেলা থেকে পাখির বাসা তৈরির কাজ শিখেছি। এখন এ কাজ করেই জীবিকা নির্বাহ করি।'

চণ্ডীদাস বলেন, শুরুর দিকে ঢাকার সেলিম সাহেব একটি নমুনা বাসা এনে বলেন, এ ধরনের পাখির বাসা তৈরি করতে পারো কি না দেখো। সঙ্গে সঙ্গে আমরা কয়েকজন ওই রকম বাসা তৈরি করে দিলাম। সেই থেকে সেলিম সাহেব আমাদের পাখির বাসা তৈরির কাজ দেন। প্রথমে আমরা ৩০ জনকে বাসা তৈরির প্রশিক্ষণ দিই। সেই থেকে কাজ করে যাচ্ছি।'

সুখলাল বলেন, গ্রাম থেকে পাখির বাসা সংগ্রহ করে এ কেন্দ্রে এনে ধুয়ে রোদে শুকিয়ে প্রতিটি বাসার গায়ে লেবেল দিয়ে ১২টা করে প্যাকেট করা হয়। পরে কার্টনে করে ট্রাকের মাধ্যমে ঢাকার অফিসে পাঠানো হয়। সেখান থেকে বিদেশে যায়।

এ শিল্পের মূল উদ্যোক্তা খায়রুল আলম ঢাকা বিশ্ববিদ্যালয় থেকে আন্তর্জাতিক সম্পর্ক বিষয়ে স্নাতকোত্তর শেষ করে চাকরির চেষ্টা করেননি। উদ্যোক্তা হওয়ার স্বপ্নে তিনি পাখির বাসাসহ পোষ্য প্রাণীর নানা উপকরণ বিদেশে রপ্তানি করেন।

খায়রুল আলম বলেন, '৩০ বছর আগে যশোরের বেসরকারি সংস্থা 'থাঁচতে শেখাই নির্বাহী পরিচালক অঞ্জেলা গোমেজের মাধ্যমে যশোরে গিয়ে পাখির বাসা তৈরি শুরু করেছিলাম। এখনো সে কাজ চলছে।' তিনি বলেন, 'বিদেশ ট্রিডিং বার্ড বা পোষা পাখির চাষ হয়। আমাদের দেশের তৈরি পাখির বাসায় ওই পাখি ডিম পাড়ে, বাচ্চা ফুটায়। যে কারণে বিদেশে এ বাসার অনেক চাহিদা। আগে চীনের দখলে ছিল ইউরোপের পাখির বাসার বাজার। সেখানে শ্রমের দাম বৃদ্ধি ও বাসা তৈরির উপকরণের সহজলভ্যতা না থাকায় বাজারটি আমরা পেয়েছি। পুঁজির অভাবে আমরা ঠিকমতো প্রসার ঘটাতে পারছি না। সরকার এ খাতে সহজ শর্তে ঋণ দিলে ইউরোপ-আমেরিকার পোষ্য পাখির উপকরণ রপ্তানির বাজার আমাদের দখলে আসবে।'

Figure 3: A Sample Scanned Document.

To improve the chances of successful digitizing process of the documents images need to be de-skewed, noise removed and segmented. These steps are called the preprocessing in OCR. A brief description is given below about these and the steps are done for it:

2.1.1 De-Skew

While scanning a document, there is always a chance it to be skewed. So a problem could be arise. In later steps of preprocessing each components need to be separated, which is much needed for corrected analysis the document. Part of image or part of a textual area could be cut off because of a skewed document. As a result the output we will get may not the match with the document. Checking and de-skewing the skewed image will lower the error rate at a great rate.

Figure 4 An Example of Image De-skew *[3]*

De-skewing a textual image is done by casting rays from left to right and find how many blackish pixels the rays intersect with. Then, by rotating the angle of the rays by a small amount and cast the rays again, and doing the same process again and again we can finally de-skew our document [4].

2.1.2 Noise Removal

Newspaper like documents are printed in Newsprint paper. It contains a lot of noise like positive or negative spots. Old documents also have a high rate of having noise in it. In the steps of recognition these positive or negative spots increase the chances of bad impact of analysis. Documents should be and needed to be preprocessed by removing those noise and have smoothed edges.

Image Before Noise Filter Image After Noise Filter

Figure 5: Image Noise Removal Example

2.1.3 Binarization

Binarization means creating a “binary image” from the colored or grey scale image document. The reason it is called binary image is mainly it contains only black and white color. Like 0 and 1. The task of Binarization is performed as a simple way of separating the text (or any other desired image component) from the background. Binarization influences the OCR stage to a significant extent the quality [2, 5].

Figure 6: Binarization Example

2.1.4 Layout Analysis & Segmentation

After de-skewing, removal of the noise and the Binarization, identifying the columns, paragraphs, captions, images etc. need to be detected as distinct blocks. This identification process is done by analyzing the document.

As OCR is the system through which any document is transformed into digital data. Now it is a fact that OCR can process texts only. If it is given an input, which contains non-textual components, garbage texts will be produced as output. So, if we want OCR to work perfectly is must be ensured that it is not fed any kind of non-textual input.

Chapter 3

Document Decomposition

In one of the previous discussion we have said that, Documents aren't always contain a simple layout. It may be as complex as having table, columns, images etc. So avoiding error on character detection stage, decomposing is the most needed pre-processing step. Automation of decomposition for a document plays a vital role when the system meets with a huge archive of documents. At that time automated decomposition process with higher accuracy ensures the most time saving decomposing.

3.1 What is Document Decomposition

Document decomposition is mainly a part of Layout Analysis. As it said that documents could have various contents, and not all the contents are needed in the step of recognition. So if these components are separated through e process than we can only use the textual part for the recognition stage.

At the document decomposition phase, the document image is decomposed in its basic principal components, such as text areas, titles, articles and the close contact of different segments hinders the application of many standard page segmentation algorithms. [6, 7].

Chapter 4

Background Study

To drive the study or action of the breaking down, research is a verifiable necessity required in order to progress further. For separating the reports and in system of making digitize data center, various examination is so far going on to breakdown the records. Many of these research work was done for English documents. As there are always structural difference between documents to documents, working or contemplating on decomposing complex Newspaper archives may come convenient as far as future work. In this section we discussed about the necessity of studying about other research works what type of research is done in the past and what we observed from various research.

4.1 Reason of Studying Previous Research Works

If we can find out that there has been some of the work was done before or if the work fits with some of the basic points with any research work, studying those research works will help a lot to get on the track of own research. If there is any kind of similarity, then it is obvious that studying those work will surely give a concept to develop own work methodology. Understanding on the work also depends on studying about the topic.

4.2 Research on Document Analysis and Recognition

To get a brief knowledge about our work, we started to look out others work on decomposing scanned documents. We find out a various number of literature for working in order to learn and work. These papers are focused on decomposing an English based document. On Document analysis and decomposition an international academic conference in the banner of "International Conference on Document Analysis and Recognition" is held in every two years of gap [8]. Our target was to study on segmenting documents like newspaper, as it is compiled with a lot of complex contents. In September, 1999 this conference was held on Bangalore of India [8]. The specialty of this conference was arranging the First International Newspaper Segmentation contest. So this paper came out handy for us to start or study.

4.2.1 Decomposing Newspaper using Image Processing & Document Analysis

To recognize a newspaper articles and segmenting its pages, an approach was introduced by B. Gatos, S.L. Mantzaris, K.V. Chandrinos, A. Tsigris and S.J. Perantonis whom are from Lambrakis Press, Athens Greece and Institute of Informatics and Telecommunications, National Research Center "Demokritos", Athens Greece. They introduced their work on the Fifth International Conference of ICDAR'99 held on Bangalore [9].

They proposed a rule-based approach to article tracking and identification. This proposal was mainly focused on newspaper images. The most significant findings of them is the separating the complex layout of newspaper pages, in particular the oldest ones, where text columns are located very close to each other in a haphazard way, as well as the poor scanning results obtained because of low print quality or deterioration through time. The rules they set to decompose the newspaper is described below:

- **Page Segmentation:** This division is centered around decomposing the essential segments of the daily paper, for example, segments, features pictures, lines and so on [9]. In a newspaper as the layout is organized in a haphazard way, the contact of different segments are so close that the application of many standard page segmentation algorithms fails [10, 11].So they proposed smearing and labeling of regions and on gradual extraction of image components in the following order [9]:
    1. Lines.
    2. Images and Drawings
    3. Background Lines
    4. Text and headline Blocks.

  They find out that vertical background need to preserve as a vertical border if the text columns nearly in touch of each other.
- **Line Extraction:** In their proposal they identified that horizontal and vertical foreground lines are usually too close to other kinds of segments, they are to be the first elements to be identified then all other region categories will be more effectively extracted. The approach they took for line identification is based on Hough Transform as well as on Morphological transformations [12].This technique gives the accurate results combining with fast implementation and can be summarized in the following steps:

  They subsampled the input image with respect to foreground pixel and then find out the vertical and horizontal lines of a newspaper. Another sub sampling was done with respect to background pixel. From the result the image need to be extract into two grayscale images which will be Foreground Horizontal (F_H) and Foreground Vertical (F_V). For each line in vertical and horizontal the pixel length needed to be assigned by creating a property of minimum length and maximum

width line with interesting characteristics will be identified. The segment whose height are greater than the largest expected headline letter was chosen then.

Again with the respect to background pixel another sub-sampling is done. With this sub-sampled image is the horizontal projections are examined with the FFT. If there are no dominant frequencies found then the block was classified as image.

Figure 7: Newspaper Image Decomposition: (a) Original, (b) Segmented Image

Then for the recognition of the text blocks, they have taken 4 threshold values.

$$T_s = \begin{cases} 0, if\ H_s\ x,y\ < t_1 \\ 1, if\ H_s\ x,y\ \geq t_1\ \widehat{}\ H_s\ x,y\ < t_2 \\ 2, if\ H_s\ x,y\ \geq t_2\ \widehat{}\ H_s\ x,y\ < t_3 \\ 3, H_s\ x,y\ \geq t_3 \end{cases} \quad (1)$$

[9]

Here t_1, t_2 and t_3 are respectively minimum expected letter height of text segment, a title and a title region. Depending on those values they decided whether a text is from headline, sub headline or columns.

4.2.2 Decomposing using a Split and Merge Approach

On January 2001, another conference was held in Seattle of USA on the banner of ICDAR. In this Conference Karim Hadjar, Oliver Hitz and Rolf Ingold from DIUF, University of Fribourg presented their work on newspaper documentation. They proposed decomposing the documents on using the Split and Merge Approach [13]. Their research mainly follows the work describe below:

In their research work, they proposed a system for daily paper page division in view of part and consolidating zones. They slice the info picture as per the even and vertical line. At that point to distinguish pictures they ascertained associated part. Each associated component is depicted by its rectangular jumping box. [13] For each letter, they got isolated rectangular box. They picked a limit estimation and if a rectangular box is greater than this esteem then that was chosen to be picture. Further they concluded that in a picture, the bouncing boxes of the encased associated parts are probably going to cover, where as in writings areas there are not very much covering it.

২০০৮ সালের ফাইল ছবিতে প্রাচীন পালমিরা থিয়েটারে একদল পর্যটক ● রয়টার্স

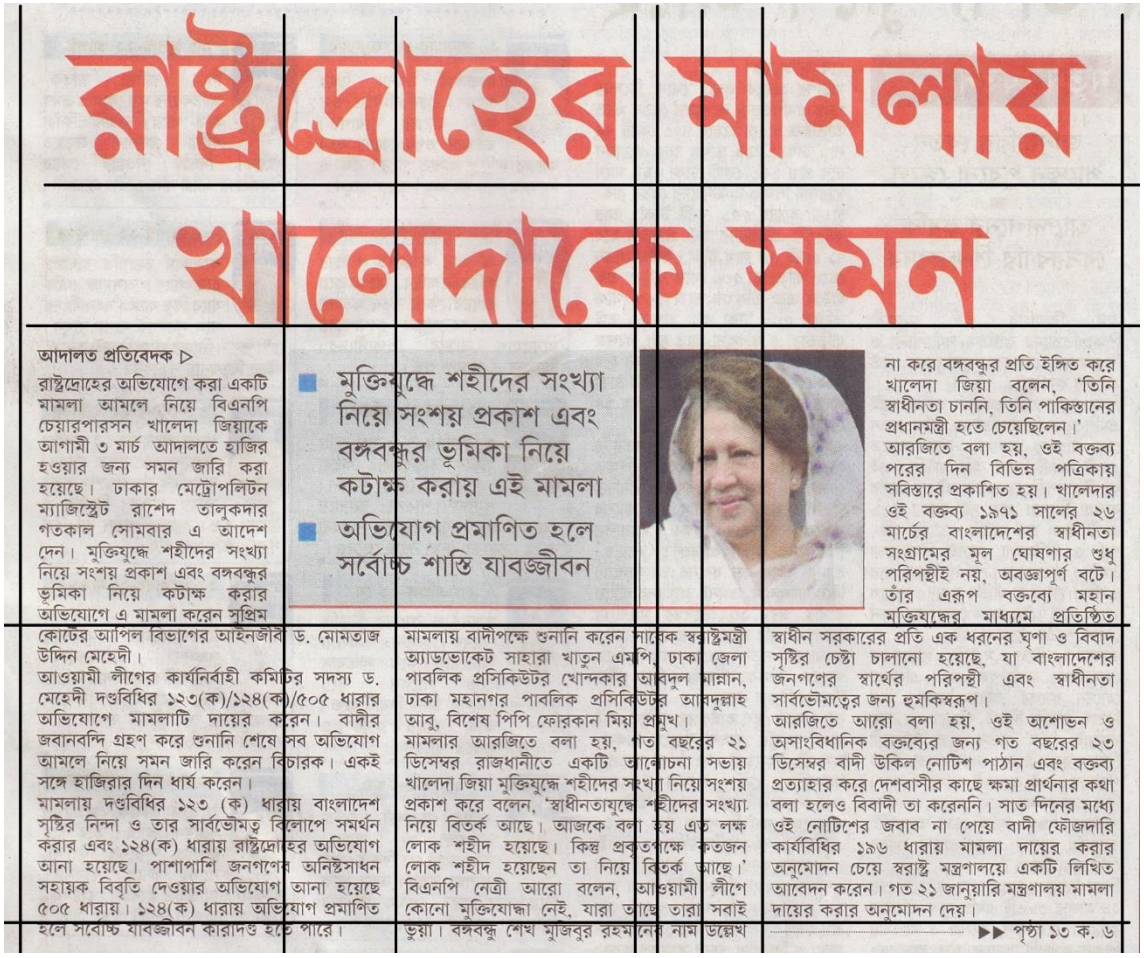
প্রাচীন শহর পালমিরা

পালমিরা শহরের পত্তন কবে হয়েছিল, তার নিশ্চিত সঠিক সময় কারও জানা নেই। তবে খ্রিষ্টপূর্ব প্রথম শতক থেকে পরবর্তী ৪০০ বছর এ মরূদ্যান শহরটি উন্নতির শিখরে পৌঁছেছিল বলে ঐতিহাসিকেরা মনে করেন। মসলা, সুগন্ধি, রেশম ও হাতির দাঁতের শিল্পের ওপর ভিত্তি করে এক সম্পদশালী শহরে পরিণত হয় পালমিরা।

'মরুর মুক্তা' নামে পরিচিত এ শহরে আছে দুই হাজার বছরেরও বেশি সময়ের পুরোনো মন্দির, কারুকার্যমণ্ডিত স্তম্ভসহ নানা নিদর্শন। জাতিসংঘের শিক্ষা, বিজ্ঞান ও সংস্কৃতি সংস্থা ইউনেসকো ১৯৮০ সালে পালমিরাকে বিশ্ব ঐতিহ্য ঘোষণা করে। ২০১১ সালে সিরিয়ায় গৃহযুদ্ধ শুরু হওয়ার আগে প্রতিবছর গড়ে দেড় লাখ পর্যটক পালমিরা ভ্রমণ করতেন।

Figure 8: In a content locale, there are no overlapping boxes of associated components.

To distinguish between text and title blocks. Text blocks are blocks, which contain text that is smaller or equal to the dominant letter height of the whole image. Blocks with text greater than the dominant letter height are title blocks.


রাষ্ট্রদ্রোহের মামলায় খালেদাকে সমন

আদালত প্রতিবেদক ▷

রাষ্ট্রদ্রোহের অভিযোগে করা একটি মামলা আমলে নিয়ে বিএনপি চেয়ারপারসন খালেদা জিয়াকে আগামী ৩ মার্চ আদালতে হাজির হওয়ার জন্য সমন জারি করা হয়েছে। ঢাকার মেট্রোপলিটন ম্যাজিস্ট্রেট রাশেদ তালুকদার গতকাল সোমবার এ আদেশ দেন। মুক্তিযুদ্ধে শহীদের সংখ্যা নিয়ে সংশয় প্রকাশ এবং বঙ্গবন্ধুর ভূমিকা নিয়ে কটাক্ষ করার অভিযোগে এ মামলা করেন সুপ্রিম কোর্টের আপিল বিভাগের আইনজীবী ড. মোমতাজ উদ্দিন মেহেদী।

আওয়ামী লীগের কার্যনির্বাহী কমিটির সদস্য ড. মেহেদী দণ্ডবিধির ১২৩(ক)/১২৪(ক)/৫০৫ ধারার অভিযোগে মামলাটি দায়ের করেন। বাদীর জবানবন্দি গ্রহণ করে শুনানি শেষে সব অভিযোগ আমলে নিয়ে সমন জারি করেন বিচারক। একই সঙ্গে হাজিরার দিন ধার্য করেন।

মামলায় দণ্ডবিধির ১২৩ (ক) ধারায় বাংলাদেশ সৃষ্টির নিন্দা ও তার সার্বভৌমত্ব বিলোপে সমর্থন করার এবং ১২৪(ক) ধারায় রাষ্ট্রদ্রোহের অভিযোগ আনা হয়েছে। পাশাপাশি জনগণের অনিষ্টসাধন সহায়ক বিবৃতি দেওয়ার অভিযোগ আনা হয়েছে ৫০৫ ধারায়। ১২৪(ক) ধারায় অভিযোগ প্রমাণিত হলে সর্বোচ্চ যাবজ্জীবন কারাদণ্ড হতে পারে।

- মুক্তিযুদ্ধে শহীদের সংখ্যা নিয়ে সংশয় প্রকাশ এবং বঙ্গবন্ধুর ভূমিকা নিয়ে কটাক্ষ করায় এই মামলা
- অভিযোগ প্রমাণিত হলে সর্বোচ্চ শাস্তি যাবজ্জীবন

মামলায় বাদীপক্ষে শুনানি করেন সাবেক স্বরাষ্ট্রমন্ত্রী অ্যাডভোকেট সাহারা খাতুন এমপি, ঢাকা জেলা পাবলিক প্রসিকিউটর খোন্দকার আবদুল মান্নান, ঢাকা মহানগর পাবলিক প্রসিকিউটর আবদুল্লাহ আবু, বিশেষ পিপি ফোরকান মিয়া প্রমুখ।

মামলার আরজিতে বলা হয়, গত বছরের ২১ ডিসেম্বর রাজধানীতে একটি আলোচনা সভায় খালেদা জিয়া মুক্তিযুদ্ধে শহীদের সংখ্যা নিয়ে সংশয় প্রকাশ করে বলেন, 'স্বাধীনতাযুদ্ধে শহীদের সংখ্যা নিয়ে বিতর্ক আছে। আজকে বলা হয় এক লক্ষ লোক শহীদ হয়েছে। কিন্তু প্রকৃতপক্ষে কতজন লোক শহীদ হয়েছেন তা নিয়ে বিতর্ক আছে।' বিএনপি নেত্রী আরো বলেন, আওয়ামী লীগে কোনো মুক্তিযোদ্ধা নেই, যারা তাছে তারা সবাই ভুয়া। বঙ্গবন্ধু শেখ মুজিবুর রহমানের নাম উল্লেখ না করে বঙ্গবন্ধুর প্রতি ইঙ্গিত করে খালেদা জিয়া বলেন, 'তিনি স্বাধীনতা চাননি, তিনি পাকিস্তানের প্রধানমন্ত্রী হতে চেয়েছিলেন।'

আরজিতে বলা হয়, ওই বক্তব্য পরের দিন বিভিন্ন পত্রিকায় সবিস্তারে প্রকাশিত হয়। খালেদার ওই বক্তব্য ১৯৭১ সালের ২৬ মার্চের বাংলাদেশের স্বাধীনতা সংগ্রামের মূল ঘোষণার শুধু পরিপন্থীই নয়, অবজ্ঞাপূর্ণ বটে। তাঁর এরূপ বক্তব্যে মহান মুক্তিযুদ্ধের মাধ্যমে প্রতিষ্ঠিত স্বাধীন সরকারের প্রতি এক ধরনের ঘৃণা ও বিবাদ সৃষ্টির চেষ্টা চালানো হয়েছে, যা বাংলাদেশের জনগণের স্বার্থের পরিপন্থী এবং স্বাধীনতা সার্বভৌমত্বের জন্য হুমকিস্বরূপ।

আরজিতে আরো বলা হয়, ওই অশোভন ও অসাংবিধানিক বক্তব্যের জন্য গত বছরের ২৩ ডিসেম্বর বাদী উকিল নোটিশ পাঠান এবং বক্তব্য প্রত্যাহার করে দেশবাসীর কাছে ক্ষমা প্রার্থনার কথা বলা হলেও বিবাদী তা করেননি। সাত দিনের মধ্যে ওই নোটিশের জবাব না পেয়ে বাদী ফৌজদারি কার্যবিধির ১৯৬ ধারায় মামলা দায়ের করার অনুমোদন চেয়ে স্বরাষ্ট্র মন্ত্রণালয়ে একটি লিখিত আবেদন করেন। গত ২১ জানুয়ারি মন্ত্রণালয় মামলা দায়ের করার অনুমোদন দেয়।

▶▶ পৃষ্ঠা ১৩ ক. ৬

Figure 9: Splitting Newspaper image into small zones.

To achieve the labeling, their algorithm computes the height of all lines inside the blocks to classify as title or text. They have said that their proposed methods will work flawlessly if and only if every input image was noiseless and was not skewed.

২০০৮ সালের ফাইল ছবিতে প্রাচীন পালমিরা থিয়েটারে একদল পর্যটক ● রয়টার্স

প্রাচীন শহর পালমিরা

পালমিরা শহরের পত্তন কবে হয়েছিল, তার নিশ্চিত সঠিক সময় কারও জানা নেই। তবে খ্রিষ্টপূর্ব প্রথম শতক থেকে পরবর্তী ৪০০ বছর এ মরূদ্যান শহরটি উন্নতির শিখরে পৌঁছেছিল বলে ঐতিহাসিকেরা মনে করেন। মসলা, সুগন্ধি, রেশম ও হাতির দাঁতের শিল্পের ওপর ভিত্তি করে এক সম্পদশালী শহরে পরিণত হয় পালমিরা।

'মরুর মুক্তা' নামে পরিচিত এ শহরে আছে দুই হাজার বছরেরও বেশি সময়ের পুরোনো মন্দির, কারুকার্যমণ্ডিত স্তম্ভসহ নানা নিদর্শন। জাতিসংঘের শিক্ষা, বিজ্ঞান ও সংস্কৃতি সংস্থা ইউনেসকো ১৯৮০ সালে পালমিরাকে বিশ্ব ঐতিহ্য ঘোষণা করে। ২০১১ সালে সিরিয়ায় গৃহযুদ্ধ শুরু হওয়ার আগে প্রতিবছর গড়ে দেড় লাখ পর্যটক পালমিরা ভ্রমণ করতেন।

Figure 10: A part of correctly segmented. Each part is each labeled with different color.

At the end they proposed that the result they obtained is an automated newspaper decomposition but can create problems in some other cases.

4.2.3 Automated Text-Line Height Calculation for Font Size Detection

In this segment, we have to study for the crucial part of the document reconstruction. While reconstructing the document, we have to work it size of the image used in the document, different font size for labeling headline, sub-headline and paragraph. OCR intelligence needs to detect the font size in text documents in order to understand the document image and extract knowledge from the document texts [14] [15].

Working with the scanned documents is doing the processing with an uncompressed image document. D. S. B. F. R. Chen at el. [15] Proposed a technique which use the Line Segmentation process in order to detect the font size.

On the other hand, compressed document gives us the high performance and storage adaptability, so that machines like fax machines [16], Photocopier machines and many others uses the compressed form of the documents [17].If the documents are like run length compressed TIFF documents detection of the font size come handy [17] in terms of complexity like time and space [18], [19], [20]. Mohammed Javed at el. [17] proposed an approach to automate the font size detection straight from the run length compressed TIFF documents by working with the binary documents. These methodology worked without doing any decompression of the document [17].

Chapter 5

Research on Bangla Document Decomposition

In the year of 1994 the work on OCR was done by U. Pal and B.B. Chaudhuri in title associated with "OCR in Bangla: an Indo-Bangladeshi Language" [21]. In later they again worked with the title "A complete printed Bangla OCR system" [22]. But both of their research work was mainly based on recognition system. And many of the published research work also focused just for the recognition system.

In our study we find out that there is lack of digitize Bangla document, many research like Machine Translation, Text Mining work in Bangla didn't advanced and for many special cases the decomposition process done in other language. It is a matter of fact that as Bangla letters are not separated from each other. Therefore, in Bangla newspaper rectangular boxes bounding texts are likely to overlap. This is the main reason documents in Bangla needs to be decomposed uniquely.

5.1 Research on Bangla Document Decomposition in SUST

At Shahjalal University of Science and Technology, Sylhet Department of Computer Science and Technology, development of Bangla OCR has been started since 1998. But they work of decomposition mainly started from the Batch 2006. Batch from 2006 and 2007 has worked on headline and column detection. Batch from 2008 Syed Rezwanul Haque Rubel, Arifuzzaman Sohel started working with "Tesseract" an OCR engine developed by Hewlett Packard Lab and sponsored by Google. [23]. From the Batch 2010 which was done by Fahad Hasan and Tasmin Afroz Tannee, They borrowed some of the steps research work of the document decomposition and added extra rules which was set by them.

5.2 Decomposition Research from Batch 2006

This batch proposed their methodology by setting a Fixed Space Width, which means the minimum space between headlines and news body. Then fixed space height which means the minimum space between columns. From their proposalat first Space width set as 0 and starting point is (0, 0).At starting point it check the pixel vertically and find the white line. If this process find white line then it increase Space Width. At the point (x, y) where Space Width is greater than Fixed Space Width, they cut the image in a rectangular shape from (x, y) to starting point. Then the process save this image as a sub image and set starting point as (x, y). If it found a line, which is not fully white, then they set Space Width to 0 and continued this process for whole image. If number of sub images from previous step is greater than 1, they considered the first sub image as headline image and other sub images as input for last step. They load each sub images for checking the pixel horizontally. To do so, set Space Height as 0 and starting point (0, 0). If we find the white line, we increase Space Height. At the point (x1, y1) where Space Height is greater than Fixed Space Height, we cut the image in a rectangular shape from stating point to (x1, y1). After that, they saved this image as a final child image and set starting point as (x1, y1). If we find a line which is not fully white, we set Space Height as 0 and continue this process for whole image.

Figure 11: Input and Output for using the work of 2006 Batch.

We can clearly see that their process is just separating the headline and the column portion of a newspaper document. But from the criteria of recognition part images should be removed must.

5.3 Decomposition Research from Batch 2007

Batch from 2007 followed the work from the batch 2006 but with setting some different rules to adjust the output. They took the width of first couple of text lines from the top of the image. It compared the widths of this line. If there is a rapid decrease in the width then they decided that headline ends on the top of current line. Afterward to detect columns, they proposed a system to count the number of black pixels and keep the result sequentially. The result would be a sequence of zero and non-zero value. Zero sequence is deduced to be column and nonzero sequence are column space.

It is a matter to notice that to work perfectly with this method, documents supposed to have the headline on the top of the article. So this approach will work for paper images like the below one. Because, here the headline is on the top.

Figure 12: Input and Output for work from 2007 Batch.

Many the time it can happen that the headline or sub headline is not on the top of the article but in betwixt the columns.

Figure 13: Sub Headline between Columns

Moreover, most importantly, newspaper page contains a lot of images along with the texts. Sometimes images are at the top of the columns, and sometimes they are in between two consecutive columns.

যশোরের পাখির বাসা যায় ইউরোপে

মনিরুল ইসলাম, যশোর

জার্মানি, ফ্রান্স, বেলজিয়াম, নেদারল্যান্ডস, স্পেন ও পর্তুগালে যাচ্ছে ৪০ ধরনের পাখির বাসা

যশোরে তৈরি শৌখিন পাখির বাসা ইউরোপের ছয় দেশে রপ্তানি হচ্ছে। এতে বছরে কোটি টাকার সমপরিমাণ বৈদেশিক মুদ্রা আসছে দেশে। এর মাধ্যমে স্বাবলম্বী হয়েছে এ অঞ্চলের হাজারো নিম্ন আয়ের মানুষ।

ইউরোপের জার্মানি, ফ্রান্স, বেলজিয়াম, নেদারল্যান্ডস, স্পেন ও পর্তুগালের বিভিন্ন শহরে শৌখিন পাখি উৎপাদনের খামারে যাচ্ছে ৪০ ধরনের পাখির বাসা। তবে পুঁজির সংকটের কারণে এ পণ্যের রপ্তানি বাণিজ্যে যতটা প্রসার ঘটার কথা ছিল, ততটা ঘটেনি বলে দাবি উদ্যোক্তাদের।

এ পণ্য রপ্তানির অন্যতম উদ্যোক্তা ঢাকার সিদ্ধেশ্বরী এলাকার বাসিন্দা খায়রুল আলম। তিনি যশোরের কয়েকটি গ্রামের নিম্ন আয়ের মানুষের মাধ্যমে পাখির বাসা তৈরি করান। পরে এগুলো রপ্তানি উপযোগী করে নৌপথে বিদেশে পাঠান।

জানতে চাইলে খায়রুল আলম বলেন, 'যশোর সদর উপজেলার আবাদ কচুয়া, সীতারামপুর ও বাহাদুরপুর গ্রামে তৈরি শৌখিন পাখির বাসার ইউরোপের বাজারে ব্যাপক চাহিদা। অন্তত ৪০ ধরনের পাখির বাসা ছয়টি দেশে এখন রপ্তানি করা হচ্ছে। বর্তমানে বছরে প্রায় ১ কোটি টাকা মূল্যের পাখির বাসা রপ্তানি হচ্ছে।'

সম্প্রতি আবাদ কচুয়া ও সীতারামপুর গ্রামে গিয়ে দেখা গেছে, ঘরে ঘরে পাখির বাসা তৈরির কাজ চলছে। ঘরের বারান্দা ও আঙিনায় বসে নারী-পুরুষ মিলে পাখির বাসা বুননের কাজ করছেন। পুরুষেরা বাঁশের চাটাই দিয়ে বুননের মূল উপকরণ তৈরি করছেন। নারীরা বাসা তৈরির জো (কাজের শুরু) তুলছেন। আরেকজন বাসা তৈরির কাজ শেষ করছেন।

আবাদ কচুয়া গ্রামের দাসপাড়ার কিনারাম দাসের বাড়ির উঠানে গিয়ে দেখা গেল, তিনি তলতা বাঁশের চাটাই চিরে দা দিয়ে সুচারু করছেন। পাশে বসে তাঁর স্ত্রী কাঞ্চন দাস ও মা রঞ্জিতা দাস বাসা বুননের জো তুলে দিচ্ছেন।

কাজের ফাঁকে কিনারাম দাস বলেন, '৩০ বছর ধরে পাখির বাসা তৈরির কাজ করছি। ৪০ ধরনের বাসা বানাতে পারি। বাঁশ, পাট, কাতা (নারকেলের ছোবড়া), খড়, বিচালি, বাঁশের পাতা, কাঠ—এ-জাতীয় নানা উপাদান দিয়ে এ বাসা তৈরি করা হয়। যখন যে ধরনের অর্ডার থাকে, তখন সেই ধরনের বাসা তৈরি করা হয়।'

কিনারাম বলেন, 'বড় ধরনের বাসার জন্য ৭ টাকা, মাঝারি ৫ ও ছোট বাসার ক্ষেত্রে ৩ টাকা করে মজুরি দেওয়া হয়। তিনজন মিলে কাজ করে মাসে ১ হাজার ৫০০ থেকে ২ হাজার ৬০০ বড় বাসা তৈরি করা যায়। মজুরি কম হলেও কাজটি বাড়িতে বসেই করা যায়। নারীরাও তা করতে পারে। তবে বাঁশের দাম এখন বেড়েছে। আমরা মজুরি বাড়ানোর জন্য দাবি তুলেছি।'

এ কাজ করে কিনারাম সংসার চালিয়েও পাকা বাড়ি নির্মাণ করেছেন। আগে তাঁরা ছাপরাঘরে থাকতেন। এখন পাকা বাড়িতে থাকেন, এটাই তাঁদের আনন্দ। এ শিল্পের সঙ্গে সম্পৃক্ত প্রায় সবাই টিনের ছাউনির পাকা বাড়ি করেছেন।

পার্শ্ববর্তী সীতারামপুর গ্রামের ঋষিপাড়ার চণ্ডীদাস ৩০ বছর আগে প্রথম পাখির বাসা তৈরির কাজ শুরু করেছিলেন।

সীতারামপুর গ্রামে চণ্ডীদাসের বাড়িতে গিয়ে দেখা গেল, চণ্ডীদাস ও তাঁর মেয়ের জামাই আরনল্ড বিশ্বাস নারকেল ও পাট দিয়ে বাবুই পাখির বাসার মতো বাসা তৈরি করছেন। আরনল্ড বিশ্বাস বলেন, 'উচ্চমাধ্যমিক পাস করে অন্য চাকরিতে যাইনি। ছোটবেলা থেকে পাখির বাসা তৈরির কাজ শিখেছি। এখন এ কাজ করেই জীবিকা নির্বাহ করি।'

চণ্ডীদাস বলেন, শুরুর দিকে ঢাকার সেলিম সাহেব একটি নমুনা বাসা এনে বলেন, এ ধরনের পাখির বাসা তৈরি করতে পারো কি না দেখো। সঙ্গে সঙ্গে আমরা কয়েকজন ওই রকম বাসা তৈরি করে দিলাম। সেই থেকে সেলিম সাহেব আমাদের পাখির বাসা তৈরির কাজ দেন। প্রথমে আমরা ৩০ জনকে বাসা তৈরির প্রশিক্ষণ দিই। সেই থেকে কাজ করে যাচ্ছি।'

সুখলাল বলেন, গ্রাম থেকে পাখির বাসা সংগ্রহ করে এ কেন্দ্রে এনে ধুয়ে রোদে শুকিয়ে প্রতিটি বাসার গায়ে লেবেল দিয়ে ১২টা করে প্যাকেট করা হয়। পরে কার্টনে করে ট্রাকের মাধ্যমে ঢাকার অফিসে পাঠানো হয়। সেখান থেকে বিদেশে যায়।

এ শিল্পের মূল উদ্যোক্তা খায়রুল আলম ঢাকা বিশ্ববিদ্যালয় থেকে আন্তর্জাতিক সম্পর্ক বিষয়ে স্নাতকোত্তর শেষ করে চাকরির চেষ্টা করেননি। উদ্যোক্তা হওয়ার স্বপ্নে তিনি পাখির বাসাসহ পোষ্য প্রাণীর নানা উপকরণ বিদেশে রপ্তানি করেন।

খায়রুল আলম বলেন, '৩০ বছর আগে যশোরের বেসরকারি সংস্থা 'বাঁচতে শেখা'র নির্বাহী পরিচালক আঞ্জেলা গোমেজের মাধ্যমে যশোরে গিয়ে পাখির বাসা তৈরি শুরু করেছিলাম। এখনো সে কাজ চলছে।' তিনি বলেন, 'বিদেশ ব্রিডিং বার্ড বা পোষা পাখির চাষ হয়। আমাদের দেশের তৈরি পাখির বাসায় ওই পাখি ডিম পাড়ে, বাচ্চা ফুটায়। যে কারণে বিদেশে এ বাসার অনেক চাহিদা। আগে চীনের দখলে ছিল ইউরোপের পাখির বাসার বাজার। সেখানে শ্রমের দাম বৃদ্ধি ও বাসা তৈরির উপকরণের সহজলভ্যতা না থাকায় বাজারটি আমরা পেয়েছি। পুঁজির অভাবে আমরা ঠিকমতো প্রসার ঘটাতে পারছি না। সরকার এ খাতে সহজ শর্তে ঋণ দিলে ইউরোপ-আমেরিকার পোষ্য পাখির উপকরণ রপ্তানির বাজার আমাদের দখলে আসবে।'

পুঁজির সংকটের কারণে এ পণ্যের রপ্তানি বাণিজ্যে যতটা প্রসার ঘটার কথা ছিল, ততটা ঘটেনি বলে দাবি উদ্যোক্তাদের

Figure 14: A Complex Document.

5.4 Decomposition Research from Batch 2010

As there was many demerits on decomposing the documents, Batch of 2010 started to work from scratch and set a methodology to work on a complex document structure. The working proposal they gave is like below:

- Edge Detection.
- Conversion of input image.
- Finding the blank spaced area (element separators) among several elements.
- Finding components adjacent to the separators and segmenting them.
- Using these portions to detect which category they belong to.

So our findings is that, they mainly focused on the work of B. Gatos at el. [9] And Karim Hadjar at el. [13] In this process the result of the input document images started to give a very good result. So this methodology was quite satisfactory. They used some threshold filter to detect the Images, Headlines and columns from the component they separated.

এক যুবকের চোখে শত মানুষের কান্না

নিজস্ব প্রতিবেদক ●

বুকে 'ন্যায়বিচার চাই' প্ল্যাকার্ড বা লেখা নিয়ে প্রধানমন্ত্রীর বাসভবনের কাছে দাঁড়িয়ে থাকা অবুল কালাম আজাদের দাবি শেষ পর্যন্ত শত মানুষের ন্যায়বিচার পাওয়ার আকাঙ্ক্ষার প্রকাশ ঘটিয়েছে। রাস্তায় একা দাঁড়িয়ে পড়া এই যুবক কার্যত চোখে করে নিয়ে এসেছিলেন নারায়ণগঞ্জের সোনারগাঁ উপজেলার নানাখী গ্রামের শত মানুষের কান্না।

শেষ পর্যন্ত আজাদের প্রতিবাদ কানে তুলেছে রাষ্ট্র। গতকাল তাঁকে ডেকে নিয়েছে স্থানীয় প্রশাসন। প্রধানমন্ত্রী শেখ হাসিনার মুখ্য সচিব অবুল কালাম আজাদ তাঁর কথা শুনতে চেয়েছেন।

গ্রামে জমিজমার 'তুচ্ছ সমস্যা' নিয়ে আজাদ কেন প্রধানমন্ত্রীর দুয়ারে কড়া নাড়ার সাহস দেখিয়েছেন, তা খুঁজতে গিয়ে বেরিয়ে এসেছে ভয়ংকর সব অভিযোগ। উপজেলার সাদিপুর ইউনিয়নের পঞ্চমীঘাট বাজারের পাশে নানাখী মৌজায় মেট্রো গার্মেন্টস অ্যান্ড ডাইং নামের একটি কারখানা করছেন নারায়ণগঞ্জের ব্যবসায়ী অমল পোদ্দার। অভিযোগ উঠেছে, তিনি ওই কারখানার জন্য এক বছর ধরে এলাকার নিরীহ মানুষের ফসলি জমি জোর করে কিনে নিচ্ছেন। বিক্রি করতে রাজি না হলে বাল্ ফেলে দখল করছেন।

গ্রামবাসী অভিযোগ করেছেন, ইতিমধ্যে ২১ জন বাসিন্দার ৩৪ বিঘা জমি দখল করা হয়েছে। ৩৫ জনকে কিছু টাকা হাতে ধরিয়ে জমি নিয়ে এলাকা থেকে বিদায় করে দেওয়া হয়েছে। এখনো অনেককে জমি, বসতবাড়ি ছেড়ে চলে যাওয়ার হুমকি দেওয়া হচ্ছে।

ভুক্তভোগীদের অভিযোগ, এ কাজে কারখানার মালিককে সহায়তা করছেন স্থানীয় আওয়ামী লীগের অঙ্গসংগঠনের কিছু নেতা-কর্মী এবং র‍্যাবের কয়েকজন সদস্য। জমি বিক্রি করতে রাজি হননি এমন কয়েকটি পরিবারের সদস্যকে র‍্যাব গ্রেপ্তার করে মামলা দায়ের করেছে। স্থানীয় প্রশাসনের কাছে গিয়ে ভুক্তভোগীরা কোনো প্রতিকার পাননি। শেষ পর্যন্ত গত রোববার আজাদ নামের যুবকটি এসে দাঁড়িয়েছেন প্রধানমন্ত্রীর দুয়ারে। সেটা করতে গিয়েও হয়রানির শিকার হয়েছে আজাদের পরিবার।

এরপর পৃষ্ঠা ২ কলাম ৫

নারায়ণগঞ্জের সোনারগাঁ উপজেলার পঞ্চমীঘাট এলাকায় জমি দখলের প্রতিবাদ করায় র‍্যাবের সাজানো মামলায় গ্রেপ্তার সোহেল মাহমুদের মা আমেনা বেগমের আহাজারি। ছবিটি গত বুধবার সকালে তোলা ● প্রথম আলো

অবুল কালাম আজাদ

Figure 15: An Input Image we used in 2010 Batch Methodology.

Image Detection:

To decide whether a block is image or not they have proposed three filters.

It checks whether the length and width of each block is greater than a threshold value or not. Here the threshold value is calculated as 100. If it finds out the pixel ratio of the blocks. If the ratio is greater than 1500 then they anticipate that this block may be an image. Finally they used to determine the ratio of the sum of the total length of each block and total width of each block. If a block is image then the ratio will be greater than 1.25. If any block satisfies all these three filters then they are decided to be an image.

Figure 16: Image detected from Document using 2010 Batch methodology

Headline Detection:

For filtering out the Headlines they set up three threshold rules:

Here they check the height and width of a block. In addition, if the process have found out that for being a headline a block should have a height should be in between 30 to 99. Here they find out the horizontal histogram of first 30% pixels towards the height of any box. If, the highest value of the horizontal histogram equal or greater than 80% of the width of the block, as well as, the horizontal histogram of last 5 pixels of the box will be 10% less than the width of the box. If any block fulfills these two characteristics then they have decided them to be a headline.

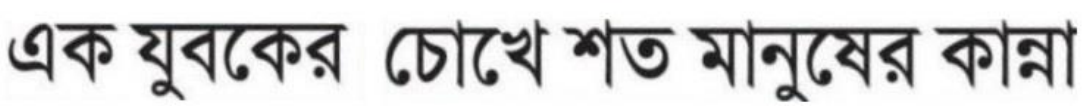


Figure 17: Headline detected from Document using 2010 Batch methodology

Column Detection:

For A newspaper contains headlines, images, columns mainly. Therefore, they marked all the other blocks as columns.

Figure 18: Column detected from Document using 2010 Batch methodology.

Chapter 6

Findings on Previous Methodology

The Bangla Document Decomposition work done by the Batch 2010 from the Department of Computer Science and Engineering of Shahjalal University of Science and Technology was the starting point for our research work. They started their work from the scratch and build a new system. So we look into the deep of this system, tried to find the errors how make a decision how to resolve them.

6.1 Findings on 2011 batch Methodology

Our first finding was that they worked on the document "e-Prothom Alo" and "e-Jonokontho". Which is a big fact on further research work. Our goal is to create a decomposition system where it have work with scanned documents. Now by studying scanned document we find out that, there is will be always some noise in it by default. And for aging, the ratio of the noise increases exponentially. This high rate of noise creates a huge impact on finding out the connected component on segmentation step. It will make the whole document one connected component and the whole segmentation process will fail on the start.


গ্যালারি কায়ায় 'ড্রি

নিজস্ব প্রতিবেদক

গ্যালারি কায়ার এবারের চিত্র প্রদর্শনীটি কয়েক প্রজন্মের নয় শিল্পীর ছবি নিয়ে। এঁদের সবার পেইন্টার হিসেবে খ্যাতি রয়েছে, কিন্তু এবারের প্রদর্শনীটি তাঁদের ড্রয়িং নিয়ে। তাই প্রদর্শনীর শিরোনামও 'ড্রয়িংস'।

অংশগ্রহণকারী শিল্পীরা হলেন মুর্তজা বশীর, চন্দ্র শেখর দে, রতন মজুমদার, জামাল আহমেদ, কাজী রকিব, রণজিৎ দাস, শিশির ভট্টাচার্য্য, গৌতম চক্রবর্তী ও মোহাম্মদ ইকবাল।

Figure 19: A Scanned Image is zoomed to preview the noise.

As we said previously that, the images are not scanned directly but downloaded from the e-prothom-alo site. These image documents are in 96dpi. Which is a really bad input for the recognition part. Each of the font will be barely recognized. Also from the binarized image we can see that a lot of bits are missing from the lines of the columns due to low resolution. These missing bits will create a very big factor for the recognition part. So for a perfect recognition with a higher rate, higher dpi images needed to be use. So we used higher dpi scanned pictures to check their methodology performances. For higher resolution images, the methodology fails to segment the documents.

কেন্দ্রীয় ব্যাংকের ৭-৮ কর্মকর্তার দায়িত্বে অবহেলা

বিশেষ প্রতিনিধি

রিজার্ভ চুরির ঘটনায় বাংলাদেশ ব্যাংকের সাত-আটজন কর্মকর্তাকে সন্দেহের তালিকায় রেখেছে পুলিশের অপরাধ তদন্ত বিভাগ (সিআইডি)। এসব কর্মকর্তার ব্যক্তিগত ব্যাংক হিসাব, মোবাইল আদান-প্রদান ও মোবাইল ফোনের কথোপকথনের তথ্য খতিয়ে দেখা হচ্ছে। তাঁদের দফায় দফায় জিজ্ঞাসাবাদও করেছে সিআইডি।

তদন্তকারী সংস্থার কর্মকর্তারা অবশ্য এখনই কারও নাম জানাতে রাজি হননি। এমনকি তাঁরা কেন্দ্রীয় ব্যাংকের কোন পর্যায়ের কর্মকর্তা—সে ইঙ্গিতও দিতে চাইছেন না সিআইডি।

সিআইডির অতিরিক্ত ডিআইজি

রিজার্ভের অর্থ চুরি

সন্দেহভাজনের তালিকা আমরা ছোট করে এনেছি

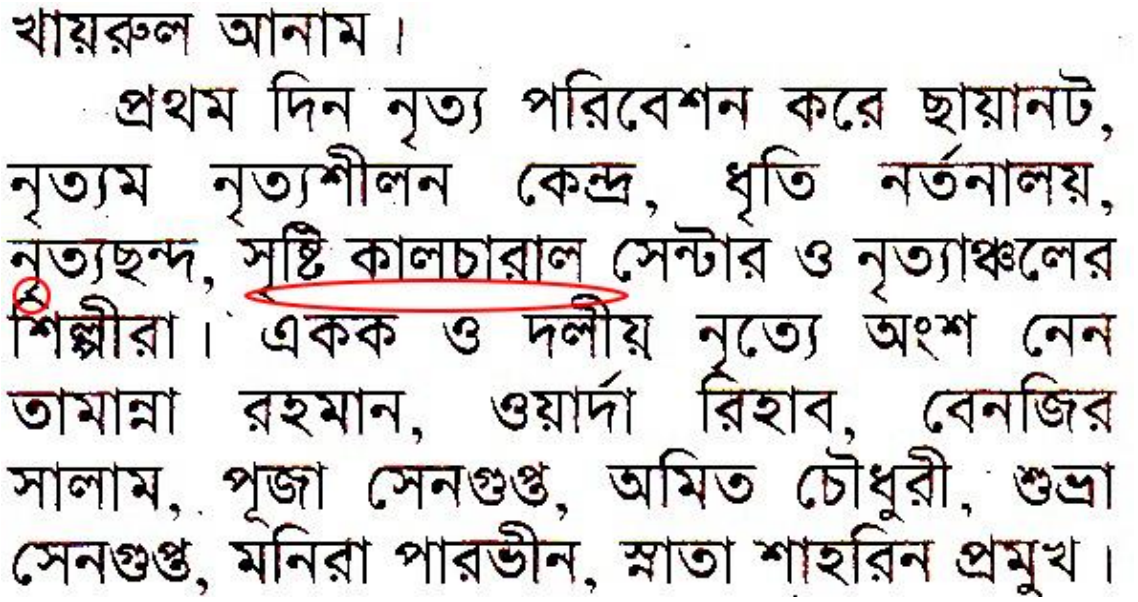
শাহ আলম
অতিরিক্ত ডিআইজি, সিআইডি

বিভিন্ন সময়ে একাধিক কর্মকর্তার ব্যক্তিগত ও সাংবিধিক কম্পিউটার জব্দ করেছে সিআইডি। অনেক কম্পিউটার জব্দ করা হয় কেন্দ্রীয়

Figure 20: 96dpi image Binarized

Because some important threshold values they determined for finding out the connected components only works with the 96 dpi image. Higher DPI like 300, it failed to segment properly.

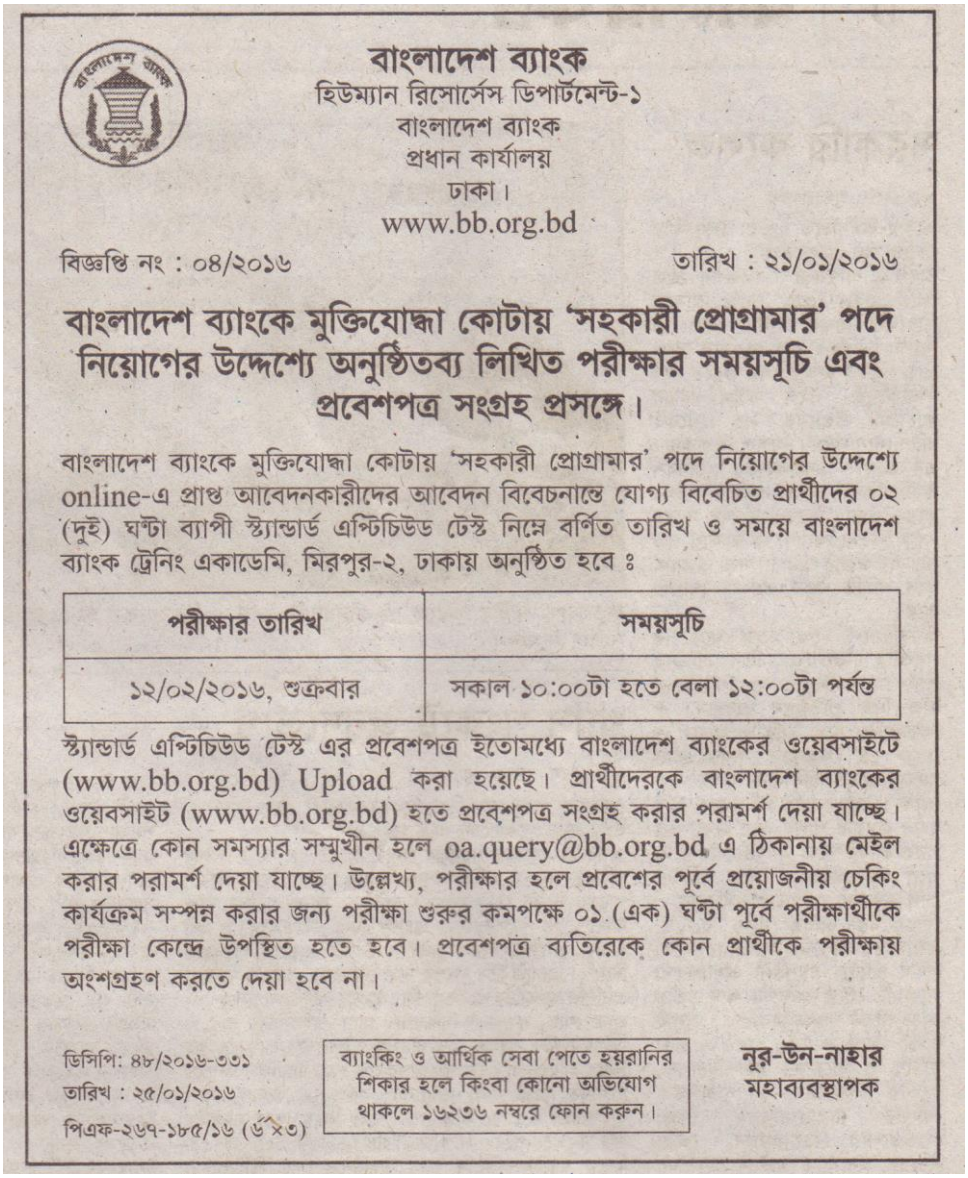
Another problem is on determining the line height. The method they use to count the line height was to start a track if a black bit is found in a row and then start find is there any row which has no black bit and the height from the tracking point must be greater than 5. But if any kind of overlapping case arise this process fails to find out the line height. A more than 1 line height will be shown as a line height.


খায়রুল আনাম।

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ।

Figure 21: Overlapping case are highlighted in red area.

On next point we find out that it fails to detect a table components. The image filtering case checks if it has a border or not. If anything with border is found as a component it labels it as an image. So tables with borders are detected as an image.



বাংলাদেশ ব্যাংক
হিউম্যান রিসোর্সেস ডিপার্টমেন্ট-১
বাংলাদেশ ব্যাংক
প্রধান কার্যালয়
ঢাকা।
www.bb.org.bd

বিজ্ঞপ্তি নং : ০৪/২০১৬ তারিখ : ২১/০১/২০১৬

বাংলাদেশ ব্যাংকে মুক্তিযোদ্ধা কোটায় 'সহকারী প্রোগ্রামার' পদে নিয়োগের উদ্দেশ্যে অনুষ্ঠিতব্য লিখিত পরীক্ষার সময়সূচি এবং প্রবেশপত্র সংগ্রহ প্রসঙ্গে।

বাংলাদেশ ব্যাংকে মুক্তিযোদ্ধা কোটায় 'সহকারী প্রোগ্রামার' পদে নিয়োগের উদ্দেশ্যে online-এ প্রাপ্ত আবেদনকারীদের আবেদন বিবেচনান্তে যোগ্য বিবেচিত প্রার্থীদের ০২ (দুই) ঘন্টা ব্যাপী স্ট্যান্ডার্ড এপ্টিচিউড টেস্ট নিম্নে বর্ণিত তারিখ ও সময়ে বাংলাদেশ ব্যাংক ট্রেনিং একাডেমি, মিরপুর-২, ঢাকায় অনুষ্ঠিত হবে ঃ

পরীক্ষার তারিখ	সময়সূচি
১২/০২/২০১৬, শুক্রবার	সকাল ১০:০০টা হতে বেলা ১২:০০টা পর্যন্ত

স্ট্যান্ডার্ড এপ্টিচিউড টেস্ট এর প্রবেশপত্র ইতোমধ্যে বাংলাদেশ ব্যাংকের ওয়েবসাইটে (www.bb.org.bd) Upload করা হয়েছে। প্রার্থীদেরকে বাংলাদেশ ব্যাংকের ওয়েবসাইট (www.bb.org.bd) হতে প্রবেশপত্র সংগ্রহ করার পরামর্শ দেয়া যাচ্ছে। এক্ষেত্রে কোন সমস্যার সম্মুখীন হলে oa.query@bb.org.bd এ ঠিকানায় মেইল করার পরামর্শ দেয়া যাচ্ছে। উল্লেখ্য, পরীক্ষার হলে প্রবেশের পূর্বে প্রয়োজনীয় চেকিং কার্যক্রম সম্পন্ন করার জন্য পরীক্ষা শুরুর কমপক্ষে ০১ (এক) ঘন্টা পূর্বে পরীক্ষার্থীকে পরীক্ষা কেন্দ্রে উপস্থিত হতে হবে। প্রবেশপত্র ব্যতিরেকে কোন প্রার্থীকে পরীক্ষায় অংশগ্রহণ করতে দেয়া হবে না।

ডিসিপি: ৪৮/২০১৬-৩৩১
তারিখ : ২৫/০১/২০১৬
পিএফ-২৬৭-১৮৫/১৬ (৬×৩)

ব্যাংকিং ও আর্থিক সেবা পেতে হয়রানির শিকার হলে কিংবা কোনো অভিযোগ থাকলে ১৬২৩৬ নম্বরে ফোন করুন।

নূর-উন-নাহার
মহাব্যবস্থাপক

Figure 22: A Sample of a Table Component

The last error we find out that, there is no case for handling the skew case. During scanning an image documents, sometimes there are some skew. These skewness is

a problem both for segmentation and recognition. So we need to DE Skew the image before implementing the segmentation process.

The very last thing we will discuss in this topic is about reconstructing the document. We find out that, there was not any implementation or research was done for reconstructing the document. So we also tried to work on this topic to resolve the issue.

Chapter 7

Current Methodology: Part 1

Rather than start over the work from the scratch, we took the previous work and changed some set of rules. Also we took the 300dpi as a standard scanned document input for research work o document decomposition. In this section we discussed about the work we have done for our course CSE 400. On later part of this chapter we discussed what we did to improve the performance of our previous work.

7.1 Noise Reduction

At first, the whole image is taken as a form of matrix. In order to create a perfect result, the noises should be removed from the scanned document. In this process Bilateral Filter Method is used. Gaussian Blur also removes noise pixel by smoothing the image, but in Bilateral Filter it preserves the edges and smooth out the noise pixels.

গ্যালারি কায়ায় 'ড্রয়িংস'

নিজস্ব প্রতিবেদক

গ্যালারি কায়ার এবারের চিত্র প্রদর্শনীটি কয়েক প্রজন্মের নয় শিল্পীর ছবি নিয়ে। এঁদের সবার পেইন্টার হিসেবে খ্যাতি রয়েছে, কিন্তু এবারের প্রদর্শনীটি তাঁদের ড্রয়িং নিয়ে। তাই প্রদর্শনীর শিরোনামও 'ড্রয়িংস'।

অংশগ্রহণকারী শিল্পীরা হলেন মুর্তজা বশীর, চন্দ্র শেখর দে, রতন মজুমদার, জামাল আহমেদ, কাজী রকিব, রণজিৎ দাস, শিশির ভট্টাচার্য্য, গৌতম চক্রবর্তী ও মোহাম্মদ ইকবাল।

প্রদর্শনীর উদ্বোধন হয়েছে গতকাল বিকেলে। অনুষ্ঠানটি পরিচালনা করেন গ্যালারি কায়ার পরিচালক শিল্পী গৌতম চক্রবর্তী।

অনুষ্ঠানে প্রধান অতিথি ছিলেন বিদ্যুৎ, জ্বালানি ও খনিজ সম্পদ প্রতিমন্ত্রী নসরুল হামিদ। তিনি বলেন, 'এ ধরনের প্রদর্শনী আরও হওয়া দরকার। একসময় সরকার শিল্পকর্ম সংগ্রহ করে শিল্পীদের উৎসাহ দিত। এখনো সেটা করা উচিত। আমরা এ ব্যাপারে দৃষ্টি দিচ্ছি।'

অতিথি হিসেবে আরও ছিলেন স্ট্যান্ডার্ড চার্টার্ড বাংলাদেশের কান্ট্রি হেড অব করপোরেট অ্যাফেয়ার্স বিটপী দাশ চৌধুরী। দর্শকদের মধ্যে উপস্থিত ছিলেন জাতীয় জাদুঘরের সাবেক মহাপরিচালক এনামুল হক ও শিল্প সমালোচক সাদেক খান।

প্রদর্শনীটি চলবে ২৭ নভেম্বর পর্যন্ত। প্রতিদিন বেলা ১১টা থেকে রাত ৮টার মধ্যে একটু সময় নিয়ে গ্যালারিতে টুঁ মারলেই ছবিগুলো দেখা যাবে।

হুমায়ূনের জন্মদিনে

বাংলা সাহিত্যের অন্যতম জনপ্রিয় লেখক, নাট্যকার ও চলচ্চিত্রকার হুমায়ূন আহমেদের ৬৭তম জন্মদিন ছিল শুক্রবার। এই শিল্পস্রষ্টা না থাকলেও তাঁর জন্মদিন ঘিরে গতকাল ছিল নানা আয়োজন। হুমায়ূনের বইয়ের প্রকাশক, পাঠক, অনুরাগী, বন্ধু ও সান্নিধ্যপ্রাপ্ত ব্যক্তিদের আনন্দ-

ছায়ানট মিলনায়তনে নৃত্য উৎসবের প্রথম দিনে দলীয় পরিবেশনা ● প্রথম আলো

উচ্ছ্বাস ও ভালোবাসায় বর্ণাঢ্যভাবে উদ্‌যাপিত হলো তাঁর জন্মদিন।

রাজধানীর সুফিয়া কামাল জাতীয় গণগ্রন্থাগার প্রাঙ্গণে শুরু হয়েছে হুমায়ূন আহমেদ একক বইমেলা। চ্যানেল আই প্রাঙ্গণে ছিল হুমায়ূন মেলা। গাজীপুরের নুহাশপল্লীতেও ছিল নানা আয়োজন। নেত্রকোনার কেন্দুয়া উপজেলার কুতুবপুর ও মোহনগঞ্জেও ছিল বর্ণাঢ্য অনুষ্ঠানমালা। এসব আয়োজনে সকাল থেকে রাত অবধি ছিল ভক্তদের পদচারণ।

শুক্রবার প্রথম প্রহরে ধানমন্ডির 'দখিন হাওয়া'য় হুমায়ূনপত্নী মেহের আফরোজ শাওন তাঁর দুই পুত্র নিষাদ ও নিনিতকে সঙ্গে নিয়ে কেক কাটেন।

সুফিয়া কামাল জাতীয় গণগ্রন্থাগারের খোলা প্রাঙ্গণে ১৫টি প্রকাশনা সংস্থার উদ্যোগে একক বইমেলার উদ্বোধনের আগেই দুপুর থেকে পাঠকেরা ভিড় জমাতে শুরু করেন। ছুটির দিন হওয়ার মেলা ছিল নানা বয়সী মানুষের পদচারণে মুখর। ৬৭টি বর্ণিল বেলুন আকাশে অবমুক্ত করার মধ্য দিয়ে নয় দিনের হুমায়ূন আহমেদ বইমেলার উদ্বোধন করেন প্রয়াত লেখকের অনুজ প্রখ্যাত শিশুসাহিত্যিক অধ্যাপক মুহম্মদ জাফর ইকবাল। এ সময় ছিলেন প্রধান অতিথি অর্থমন্ত্রী আবুল মাল আবদুল মুহিত, মেহের আফরোজ শাওন প্রমুখ। মেলা উদ্বোধনের পর অন্যপ্রকাশ থেকে প্রকাশিত সালেহ চৌধুরী সম্পাদিত *শিশু-কিশোর হুমায়ূন রচনাবলি*-২-এর মোড়ক উন্মোচন করা হয়। ২১ নভেম্বর পর্যন্ত বইমেলা প্রতিদিন বেলা ১১টা থেকে রাত ৮টা পর্যন্ত সবার জন্য খোলা।

চ্যানেল আইতে মেলা : চ্যানেল আই প্রাঙ্গণে হুমায়ূন মেলায় ছিল গান, নাচ, শিশু চিত্রাঙ্কন, স্মৃতিচারণা। জনপ্রিয় চরিত্র হিমুর পছন্দের পোশাক হলুদ পাঞ্জাবি পরে দল বেঁধে মেলায় আসে তরুণেরা। তরুণীদের পরনে ছিল হলুদ রঙের শাড়ি। সকালে ৬৭টি হলুদ রঙের বেলুন উড়িয়ে শুরু হয় মেলার আনুষ্ঠানিকতা। উদ্বোধনী অনুষ্ঠানে স্বাগত বক্তব্য দেন চ্যানেল আইয়ের ব্যবস্থাপনা পরিচালক ফরিদুর রেজা সাগর।

নুহাশপল্লীর মূল ফটকে হুমায়ূন আহমেদের ম্যুরাল স্থাপন অনুষ্ঠানসহ 'হিমু মেলা' সরাসরি সম্প্রচার করেছে চ্যানেল আই।

ছায়ানটে দুই দিনের নৃত্য উৎসব শুরু

টানা কয়েক বছরের ধারাবাহিকতায় এবারও ছায়ানট আয়োজন করেছে দুই দিনের নৃত্য উৎসব। গতকাল শুক্রবার ছায়ানট মিলনায়তনে এ উৎসব শুরু হয়েছে। উদ্বোধনের শুরুতে স্বাগত বক্তব্য দেন ছায়ানট সাধারণ সম্পাদক খায়রুল আনাম।

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ। আজ উৎসবের শেষ দিনের অনুষ্ঠান শুরু হবে সন্ধ্যা সাড়ে ছয়টায়।

মুর্তজা বশীরের একটি ড্রয়িং

হুমায়ূন মেলায় গান করেন সাদি মহম্মদ

গ্যালারি কায়ায় 'ড্রয়িংস'

নিজস্ব প্রতিবেদক

গ্যালারি কায়ার এবারের চিত্র প্রদর্শনীটি কয়েক প্রজন্মের নয় শিল্পীর ছবি নিয়ে। এঁদের সবার পেইন্টার হিসেবে খ্যাতি রয়েছে, কিন্তু এবারের প্রদর্শনীটি তাঁদের ড্রয়িং নিয়ে। তাই প্রদর্শনীর শিরোনামও 'ড্রয়িংস'।

অংশগ্রহণকারী শিল্পীরা হলেন মুর্তজা বশীর, চন্দ্র শেখর দে, রতন মজুমদার, জামাল আহমেদ, কাজী রকিব, রণজিৎ দাস, শিশির ভট্টাচার্য্য, গৌতম চক্রবর্তী ও মোহাম্মদ ইকবাল।

প্রদর্শনীর উদ্বোধন হয়েছে গতকাল বিকেলে। অনুষ্ঠানটি পরিচালনা করেন গ্যালারি কায়ার পরিচালক শিল্পী গৌতম চক্রবর্তী।

অনুষ্ঠানে প্রধান অতিথি ছিলেন বিদ্যুৎ, জ্বালানি ও খনিজ সম্পদ প্রতিমন্ত্রী নসরুল হামিদ। তিনি বলেন, 'এ ধরনের প্রদর্শনী আরও হওয়া দরকার। একসময় সরকার শিল্পকর্ম সংগ্রহ করে শিল্পীদের উৎসাহ দিত। এখনো সেটা করা উচিত। আমরা এ ব্যাপারে দৃষ্টি দিচ্ছি।'

অতিথি হিসেবে আরও ছিলেন স্ট্যান্ডার্ড চার্টার্ড বাংলাদেশের কান্ট্রি হেড অব করপোরেট অ্যাফেয়ার্স বিটপী দাশ চৌধুরী। দর্শকদের মধ্যে উপস্থিত ছিলেন জাতীয় জাদুঘরের সাবেক মহাপরিচালক এনামুল হক ও শিল্প সমালোচক সাদেক খান।

প্রদর্শনীটি চলবে ২৭ নভেম্বর পর্যন্ত। প্রতিদিন বেলা ১১টা থেকে রাত ৮টার মধ্যে একটু সময় নিয়ে গ্যালারিতে ঢুঁ মারলেই ছবিগুলো দেখা যাবে।

হুমায়ূনের জন্মদিনে

বাংলা সাহিত্যের অন্যতম জনপ্রিয় লেখক, নাট্যকার ও চলচ্চিত্রকার হুমায়ূন আহমেদের ৬৭তম জন্মদিন ছিল শুক্রবার। এই শিল্পস্রষ্টা না থাকলেও তাঁর জন্মদিন ঘিরে গতকাল ছিল নানা আয়োজন। হুমায়ূনের বইয়ের প্রকাশক, পাঠক, অনুরাগী, বন্ধু ও সান্নিধ্যপ্রাপ্ত ব্যক্তিদের আনন্দ-

ছায়ানট মিলনায়তনে নৃত্য উৎসবের প্রথম দিনে দলীয় পরিবেশনা ● প্রথম আলো

উচ্ছ্বাস ও ভালোবাসায় বর্ণাঢ্যভাবে উদ্‌যাপিত হলো তাঁর জন্মদিন।

রাজধানীর সুফিয়া কামাল জাতীয় গণগ্রন্থাগার প্রাঙ্গণে শুরু হয়েছে হুমায়ূন আহমেদ একক বইমেলা। চ্যানেল আই প্রাঙ্গণে ছিল হুমায়ূন মেলা। গাজীপুরের নুহাশপল্লীতেও ছিল নানা আয়োজন। নেত্রকোনার কেন্দুয়া উপজেলার কুতুবপুর ও মোহনগঞ্জেও ছিল বর্ণাঢ্য অনুষ্ঠানমালা। এসব আয়োজনে সকাল থেকে রাত অবধি ছিল ভক্তদের পদচারণ।

শুক্রবার প্রথম প্রহরে ধানমন্ডির 'দখিন হাওয়া'য় হুমায়ূনপত্নী মেহের আফরোজ শাওন তাঁর দুই পুত্র নিষাদ ও নিনিতকে সঙ্গে নিয়ে কেক কাটেন।

সুফিয়া কামাল জাতীয় গণগ্রন্থাগারের খোলা প্রাঙ্গণে ১৫টি প্রকাশনা সংস্থার উদ্যোগে একক বইমেলার উদ্বোধনের আগেই দুপুর থেকে পাঠকেরা ভিড় জমাতে শুরু করেন। ছুটির দিন হওয়ার মেলা ছিল নানা বয়সী মানুষের পদচারণে মুখর। ৬৭টি বর্ণিল বেলুন আকাশে অবমুক্ত করার মধ্য দিয়ে নয় দিনের হুমায়ূন আহমেদ বইমেলার উদ্বোধন করেন প্রয়াত লেখকের অনুজ প্রখ্যাত শিশুসাহিত্যিক অধ্যাপক মুহম্মদ জাফর ইকবাল। এ সময় ছিলেন প্রধান অতিথি অর্থমন্ত্রী আবুল মাল আবদুল মুহিত, মেহের আফরোজ শাওন প্রমুখ। মেলা উদ্বোধনের পর অন্যপ্রকাশ থেকে প্রকাশিত সালেহ চৌধুরী সম্পাদিত *শিশু-কিশোর হুমায়ূন রচনাবলি*-২-এর মোড়ক উন্মোচন করা হয়। ২১ নভেম্বর পর্যন্ত বইমেলা প্রতিদিন বেলা ১১টা থেকে রাত ৮টা পর্যন্ত সবার জন্য খোলা।

চ্যানেল আইতে মেলা : চ্যানেল আই প্রাঙ্গণে হুমায়ূন মেলায় ছিল গান, নাচ, শিশু চিত্রাঙ্কন, স্মৃতিচারণা। জনপ্রিয় চরিত্র হিমুর পছন্দের পোশাক হলুদ পাঞ্জাবি পরে দল বেঁধে মেলায় আসে তরুণেরা। তরুণীদের পরনে ছিল হলুদ রঙের শাড়ি। সকালে ৬৭টি হলুদ রঙের বেলুন উড়িয়ে শুরু হয় মেলার আনুষ্ঠানিকতা। উদ্বোধনী অনুষ্ঠানে স্বাগত বক্তব্য দেন চ্যানেল আইয়ের ব্যবস্থাপনা পরিচালক ফরিদুর রেজা সাগর।

নুহাশপল্লীর মূল ফটকে হুমায়ূন আহমেদের ম্যুরাল স্থাপন অনুষ্ঠানসহ 'হিমু মেলা' সরাসরি সম্প্রচার করেছে চ্যানেল আই।

ছায়ানটে দুই দিনের নৃত্য উৎসব শুরু

টানা কয়েক বছরের ধারাবাহিকতায় এবারও ছায়ানট আয়োজন করেছে দুই দিনের নৃত্য উৎসব। গতকাল শুক্রবার ছায়ানট মিলনায়তনে এ উৎসব শুরু হয়েছে। উদ্বোধনের শুরুতে স্বাগত বক্তব্য দেন ছায়ানট সাধারণ সম্পাদক খায়রুল আনাম।

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ।

Figure 7.1.2: Noise Filtered Document

7.2 Shaded pixel removed

From the noise free image we have created a threshed Matrix of the image. This image doesn't contain any shaded pixel. For Removing shaded pixels, getting chance of correct components increases.

গ্যালারি কায়ায় 'ড্রয়িংস'

নিজস্ব প্রতিবেদক ●

গ্যালারি কায়ার এবারের চিত্র প্রদর্শনীটি কয়েক প্রজন্মের নয় শিল্পীর ছবি নিয়ে। এঁদের সবার পেইন্টার হিসেবে খ্যাতি রয়েছে, কিন্তু এবারের প্রদর্শনীটি তাঁদের ড্রয়িং নিয়ে। তাই প্রদর্শনীর শিরোনামও 'ড্রয়িংস'।

অংশগ্রহণকারী শিল্পীরা হলেন মূর্তজা বশীর, চন্দ্র শেখর দে, রতন মজুমদার, জামাল আহমেদ, কাজী রকিব, রণজিৎ দাস, শিশির ভট্টাচার্য্য, গৌতম চক্রবর্তী ও মোহাম্মদ ইকবাল।

প্রদর্শনীর উদ্বোধন হয়েছে গতকাল বিকেলে। অনুষ্ঠানটি পরিচালনা করেন গ্যালারি কায়ার পরিচালক শিল্পী গৌতম চক্রবর্তী।

অনুষ্ঠানে প্রধান অতিথি ছিলেন বিদ্যুৎ, জ্বালানি ও খনিজ সম্পদ প্রতিমন্ত্রী নসরুল হামিদ। তিনি বলেন, 'এ ধরনের প্রদর্শনী আরও হওয়া দরকার। একসময় সরকার শিল্পকর্ম সংগ্রহ করে শিল্পীদের উৎসাহ দিত। এখনো সেটা করা উচিত। আমরা এ ব্যাপারে দৃষ্টি দিচ্ছি।'

অতিথি হিসেবে আরও ছিলেন স্ট্যান্ডার্ড চার্টার্ড বাংলাদেশের কান্ট্রি হেড অব করপোরেট অ্যাফেয়ার্স বিটপী দাশ চৌধুরী। দর্শকদের মধ্যে উপস্থিত ছিলেন জাতীয় জাদুঘরের সাবেক মহাপরিচালক এনামুল হক ও শিল্প সমালোচক সাদেক খান।

প্রদর্শনীটি চলবে ২৭ নভেম্বর পর্যন্ত। প্রতিদিন বেলা ১১টা থেকে রাত ৮টার মধ্যে একটু সময় নিয়ে গ্যালারিতে টুঁ মারলেই ছবিগুলো দেখা যাবে।

হুমায়ূনের জন্মদিনে

বাংলা সাহিত্যের অন্যতম জনপ্রিয় লেখক, নাট্যকার ও চলচ্চিত্রকার হুমায়ূন আহমেদের ৬৭তম জন্মদিন ছিল শুক্রবার। এই শিল্পস্রষ্টা না থাকলেও তাঁর জন্মদিন ঘিরে গতকাল ছিল নানা আয়োজন। হুমায়ূনের বইয়ের প্রকাশক, পাঠক, অনুরাগী, বন্ধু ও সান্নিধ্যপ্রাপ্ত ব্যক্তিদের আনন্দ-

ছায়ানট মিলনায়তনে নৃত্য উৎসবের প্রথম দিনে দলীয় পরিবেশনা ● প্রথম আলো

উচ্ছ্বাস ও ভালোবাসায় বর্ণাঢ্যভাবে উদ্‌যাপিত হলো তাঁর জন্মদিন।

রাজধানীর সুফিয়া কামাল জাতীয় গণগ্রন্থাগার প্রাঙ্গণে শুরু হয়েছে হুমায়ূন আহমেদ একক বইমেলা। চ্যানেল আই প্রাঙ্গণে ছিল হুমায়ূন মেলা। গাজীপুরের নুহাশপল্লীতেও ছিল নানা আয়োজন। নেত্রকোনার কেন্দুয়া উপজেলার কুতুবপুর ও মোহনগঞ্জেও ছিল বর্ণাঢ্য অনুষ্ঠানমালা। এসব আয়োজনে সকাল থেকে রাত অবধি ছিল ভক্তদের পদচারণ।

শুক্রবার প্রথম প্রহরে ধানমন্ডির 'দখিন হাওয়া'য় হুমায়ূনপত্নী মেহের আফরোজ শাওন তাঁর দুই পুত্র নিষাদ ও নিনিতকে সঙ্গে নিয়ে কেক কাটেন।

সুফিয়া কামাল জাতীয় গণগ্রন্থাগারের খোলা প্রাঙ্গণে ১৫টি প্রকাশনা সংস্থার উদ্যোগে একক বইমেলার উদ্বোধনের আগেই দুপুর থেকে পাঠকেরা ভিড় জমাতে শুরু করেন। ছুটির দিন হওয়ার মেলা ছিল নানা বয়সী মানুষের পদচারণে মুখর। ৬৭টি বর্ণিল বেলুন আকাশে অবমুক্ত করার মধ্য দিয়ে নয় দিনের হুমায়ূন আহমেদ বইমেলার উদ্বোধন করেন প্রয়াত লেখকের অনুজ প্রখ্যাত শিশুসাহিত্যিক অধ্যাপক মুহম্মদ জাফর ইকবাল। এ সময় ছিলেন প্রধান অতিথি অর্থমন্ত্রী আবুল মাল আবদুল মুহিত, মেহের আফরোজ শাওন প্রমুখ। মেলা উদ্বোধনের পর অন্যপ্রকাশ থেকে প্রকাশিত সালেহ চৌধুরী সম্পাদিত *শিশু-কিশোর হুমায়ূন রচনাবলি*-২-এর মোড়ক উন্মোচন করা হয়। ২১ নভেম্বর পর্যন্ত বইমেলা প্রতিদিন বেলা ১১টা থেকে রাত ৮টা পর্যন্ত সবার জন্য খোলা।

চ্যানেল আইতে মেলা : চ্যানেল আই প্রাঙ্গণে হুমায়ূন মেলায় ছিল গান, নাচ, শিশু চিত্রাঙ্কন, স্মৃতিচারণা। জনপ্রিয় চরিত্র হিমুর পছন্দের পোশাক হলুদ পাঞ্জাবি পরে দল বেঁধে মেলায় আসে তরুণেরা। তরুণীদের পরনে ছিল হলুদ রঙের শাড়ি। সকালে ৬৭টি হলুদ রঙের বেলুন উড়িয়ে শুরু হয় মেলার আনুষ্ঠানিকতা। উদ্বোধনী অনুষ্ঠানে স্বাগত বক্তব্য দেন চ্যানেল আইয়ের ব্যবস্থাপনা পরিচালক ফরিদুর রেজা সাগর।

নুহাশপল্লীর মূল ফটকে হুমায়ূন আহমেদের ম্যুরাল স্থাপন অনুষ্ঠানসহ 'হিমু মেলা' সরাসরি সম্প্রচার করেছে চ্যানেল আই।

ছায়ানটে দুই দিনের নৃত্য উৎসব শুরু

টানা কয়েক বছরের ধারাবাহিকতায় এবারও ছায়ানট আয়োজন করেছে দুই দিনের নৃত্য উৎসব। গতকাল শুক্রবার ছায়ানট মিলনায়তনে এ উৎসব শুরু হয়েছে। উদ্বোধনের শুরুতে স্বাগত বক্তব্য দেন ছায়ানট সাধারণ সম্পাদক খায়রুল আনাম।

প্রথম দিন নৃত্য পরিবেশন করে ছায়ানট, নৃত্যম নৃত্যশীলন কেন্দ্র, ধৃতি নর্তনালয়, নৃত্যছন্দ, সৃষ্টি কালচারাল সেন্টার ও নৃত্যাঞ্চলের শিল্পীরা। একক ও দলীয় নৃত্যে অংশ নেন তামান্না রহমান, ওয়ার্দা রিহাব, বেনজির সালাম, পূজা সেনগুপ্ত, অমিত চৌধুরী, শুভ্রা সেনগুপ্ত, মনিরা পারভীন, স্নাতা শাহরিন প্রমুখ।

Figure 24: Image Matrix of Non Shaded Pixel

7.3 Edge Detection

From the noise free image we find edges in the input image and mark them in the output map using the 'Canny' algorithm. The working method of 'Canny' edge detector is described below.

7.3.1 Canny Edge Detector

The *Canny Edge detector* was developed by John F. Canny in 1986. Also known to many as the *optimal detector*, canny algorithm aims to satisfy three main criteria:

- **Low error rate:** Meaning a good detection of only existent edges.
- **Good localization:** The distance between edge pixels detected and real edge pixels have to be minimized.
- **Minimal response:** Only one detector response per edge.

Steps:

1. Filter out any noise. The Gaussian filter is used for this purpose. An example of a Gaussian kernel of size = 5 that might be used is shown below:

$$k = \frac{1}{159}\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} \tag{2}$$

2. Find the intensity gradient of the image. For this, we follow a procedure analogous to Sobel: a. Apply a pair of convolution masks (in x and y directions):

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{3}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \tag{4}$$

b. Find the gradient strength and direction with:

$$G = \sqrt{G_x^2 + G_y^2} \tag{5}$$

$$\theta = \arctan(\frac{G_y}{G_x}) \tag{6}$$

3. Non-maximum suppression is applied. This removes pixels that are not considered to be part of an edge. Hence, only thin lines (candidate edges) will remain.
4. **Hysteresis:** The final step. Canny does use two thresholds (upper and lower): a. if a pixel gradient is higher than the upper threshold, the pixel is accepted as an edge b. If a pixel gradient value is below the lower threshold, then it is rejected. c. If the pixel gradient is between the two thresholds, then it will be accepted only if it is connected to a pixel that is above the upper threshold. Canny recommended an upper: lower ratio between 2:1 and 3:1.

গ্যালারি কায়ায় 'ড্রয়িংস'

হায়নাটে মিলনায়তনে নৃত্য উৎসবের প্রথম দিনে দলীয় পরিবেশনা ও প্রথম আলো

হুমায়ূনের জন্মদিনে

ছায়ানটে দুই দিনের নৃত্য উৎসব শুরু

7.4 Conversion of input image

This process is done for creating an array of connected components. To process this step an array will contain values for every pixel according to the length of the vertical and horizontal line it belongs to. That is every pixel of same line will be given a same value, which is the total length or height of that black line. Here we assigned 0 (black) for pixel to the array if the pixel color is not black (>0). Thus we will get an image where, the area which contains texts, images etc. will be black and the area which are not textual area (white spaced, mainly the separating space among articles) will be white.

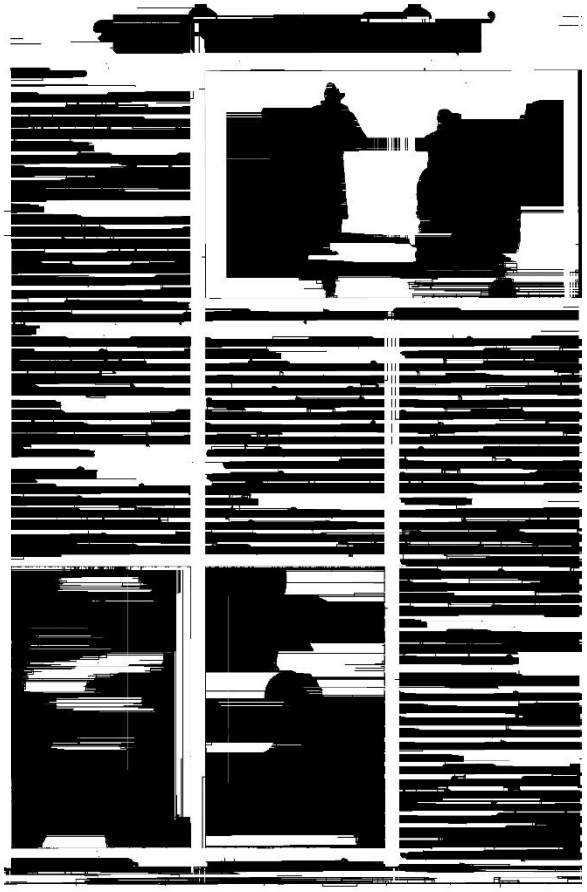


Figure 26: Conversion Image

7.5 Discerning the Element Separators

Again a binary image of the input image is need to be created for this step. As we know, usually all the textual elements of any page starts after maintaining a certain gap from the page borders [13]. We pick pixels and compare them with some threshold values to determine whether they belong to non-text area (white spaced area of converted image) or not. If any pixel, which is, we compare the height as well as comparing the width, we can anticipate that it lies on the horizontal non-text area. Again when a pixel is found which is, we compare the height as well as comparing the width, we can anticipate that it lies on the vertical non-text area. For comparing this height and we have set a static value. These static values are assigned by observing the document structures.

7.6 Segmenting the Black Boxes

In this stage, we store the top-left and bottom-right coordinates of each black box neighboring the horizontal or vertical element separators from the converted image. Then we cut our image according to coordinates of the black blocks. This cutting process is done by a using the BFS Algorithm.

7.7 Distinguishing the Elements

A newspaper mainly contains three kinds of elements like image, headline and columns. In this stage, we detected the category of each segmented block.

Image Detection:

From the previous methodology, we just change some of the threshold value to work on higher resolution image detection. Because the number of pixels is much more than a low res documents in a high res document.

1st filter

It checks whether the length and width of each block is greater than a threshold value or not. The threshold value was set by observing the document structure.

2nd filter

It finds out the pixel ratio of the blocks. If the ratio is at much higher rate, then we anticipate that this block may be an image.

3rd filter

It is used to determine the ratio of the sum of the total length of each block and total width of each block. If a block is image then the ratio will be greater than 1.25. If any block satisfies all these three filters then they are decided to be an image.

Figure 27: Detected Images

Table Detection:

In last methodology, in every case table was detected as an image. So it was easy filter for us. After detecting it as an image. Another filter process is done with the sub matrix. In this filter it is checked that is there any line present which takes more than 70% in a row and is the percentage of those line is below 5%.This set of rule is checked for both vertical histogram and horizontal histogram of that sub matrix. The whole process is like below:

1. HH_{max} = Max(Horizontal Histogram)
2. VH_{max} = Max(Vertical Histogram)
3. HH_{Range} = Threshold Percentage of HH_{max}
4. VH_{Range} = Threshold Percentage of VH_{max}
5. HHF = Frequency (All HH in HH_{Range})
6. VHF = Frequncey (All VH in VH_{Range})
7. SHH = (HHF/Frequency (HH)) * 100.0
8. SVH = (VHF/Frequency (VH)) * 100.0

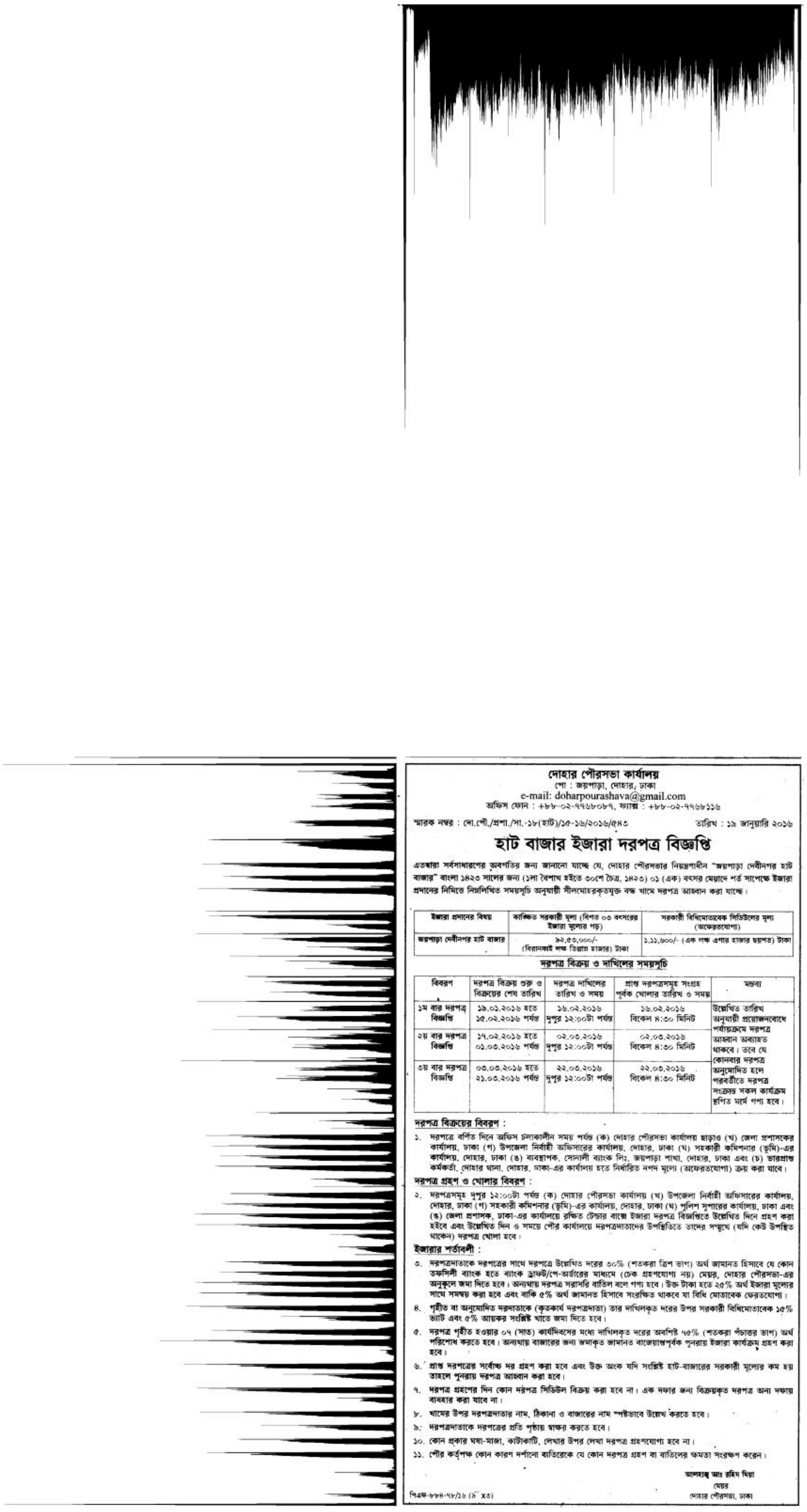


Figure 28: A Table with Vertical and Horizontal Histogram

Headlines Detection:

From the previous methodology we have just change some of the threshold value to work on the case of detecting headlines. To ensure whether a block is headline or not, at first, we checked the height and width of the block and then we checked a specific ratio through the histogram of that block. Only the component left after detecting the images are used in detecting the headlines. Here we check the height and width of a block. In addition, we have found out that for being a headline a block should have a height should be greater than other textual components. The highest range is detected as headline.

Figure 29: Detected Headline

Column Detection:

A newspaper contains headlines, images, columns mainly. Therefore, other blocks are marked as columns. We can see there are 6 columns in the given image.

7.8 Line Height Calculation

From the methodology of previous work, height of a line was calculated in the process like below:

1. Start from row 1.
2. If Black Bit found start= row position.
3. Go on until row with no black bit found.
4. If case 3 is True then position of that row - start is the height of the line

5. Else Height of the total sub matrix is the line height.

This Methodology has a negative side if an overlapping case like below occurs.

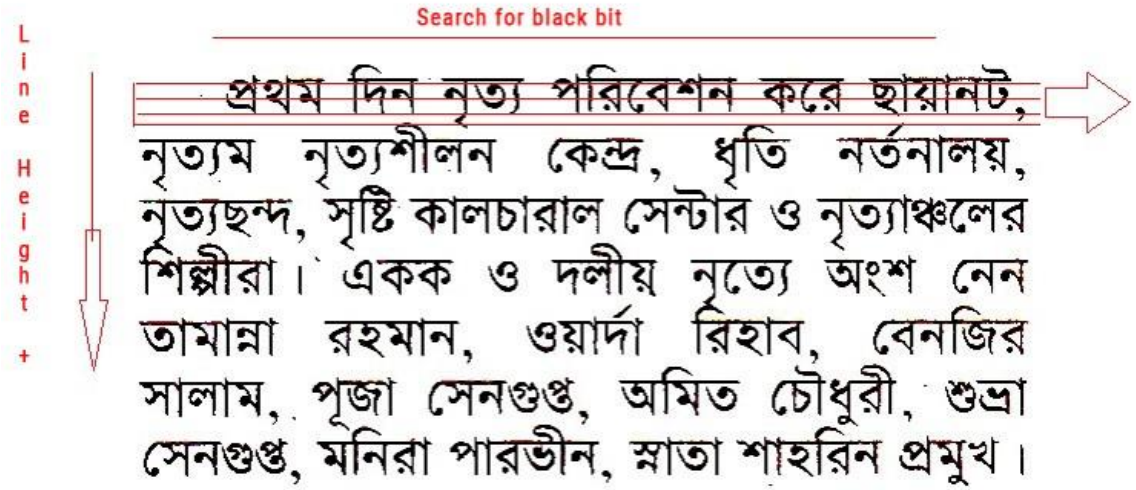


Figure 30: Line Overlapping Case

So we have proposed a new line height calculation technique.

1. Take the height of all connected component which is similar to a word.
2. H_{Filter} = Remove Garbage Height (H, Threshold)
3. Finding H_{Avg} from H_{Filter}
4. $H_{Standard\ daviation}$ of H_{Filter}
5. Set $H_{Max} = H_{Avg} + H_{Standard\ Deviation}$

   $H_{Min} = H_{Avg} - H_{Standard\ Deviation}$
6. $H_{Candidate} = H_{Min} < H_{Filter} < H_{Max}$
7. $H_{Line\ Height}$ = Max of $H_{Candidate}$

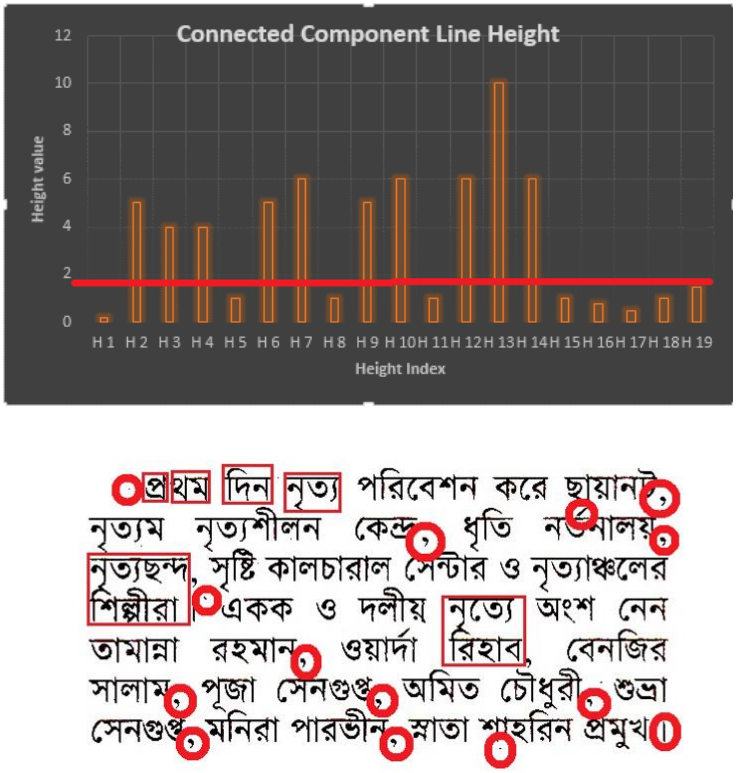


Figure 7.8.2: Removing garbage line height

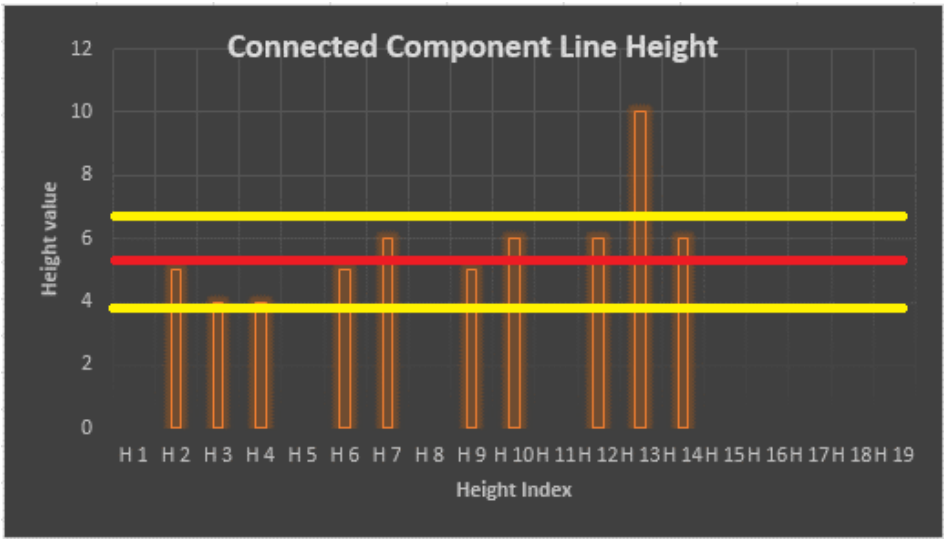


Figure 7.8.3: Line Height Standard Deviation

In this process we get a result which much more correct than previous system.

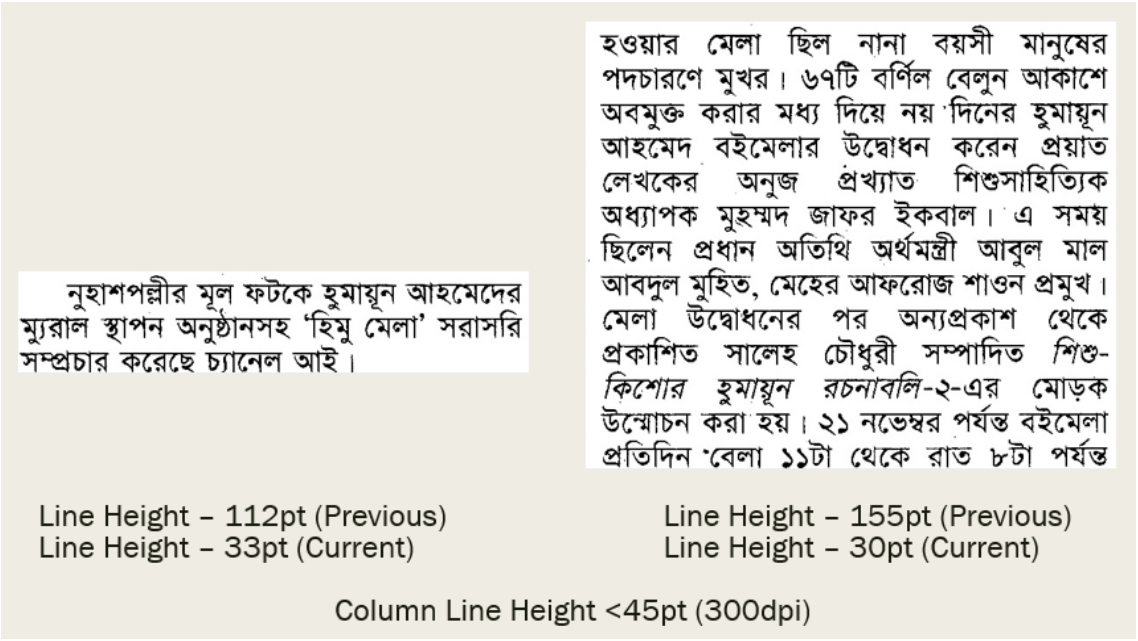


Figure 7.8.4: Line Height in new Methodology

7.9 Comparing Threshold Values with Previous Methodology

	Line Height for 96dpi	Line Height for 300dpi
Headline	h>50px	h>100px
Sub-Headline	h>23px	h>45px
Column	h>8px	h>18px

Figure 7.9.1: Line height threshold for categorization.

	Space gap for 96dpi	Space gap for 300dpi
HORIZONTAL NON TEXT AREA VERTICAL VALUE	6	8
HORIZONTAL NON TEXT AREA HORIZONTAL VALUE	150	180
VERTICAL NON TEXT AREA HORIZONTAL VALUE	7	9
VERTICAL NON TEXT AREA VERTICAL VALUE	200	270

Figure 7.9.2: Conversion space threshold.

Chapter 8

Current Methodology: Part 2

In this chapter we discussed how we improved our performances by solving the skew problem. Also we discussed about the document reconstruction system after recognition.

8.1 Skew Detection and De-skew

During scanning an image, if the document wasn't alignment correctly, we have to face the skew problem. So if we start to decompose a skew image, then our methodology will not work perfectly. To going on with our methodology, we must DE Skew it first.

8.2 Using Hough Transform for De-skew

Hough transform is very handful when we need to extract shape features form the image. In most of the cases, it is used for detecting lines but we can also use it to find circles and ellipses by extending the features. We also can extend this methodology to detect the skew of a scan image, then calculate the skew angle. Using this property, we have to rotate the image to remove the skew. The Interline Cross-Correlation concept and the Hough Transform was used by Hong Yan to perform the skew correction option [24]. In the year 1997 Huei-Fen Jiang, Chin chuan Han, Kuo-Chin Fan also proposed a fast approach to detect the skew document using the Hough Transformation [25]. In recent work Chandan Singh, Nitin Bhatia and Amandeep Kaur researched on accurate skew correction also based on the Hough Transform [26]. So we implemented their work to detect any kind of skew. Using this methodology we saw that, we were able to perform correctly if there are no image. But in case of an image, the methodology fails for detecting any kind of line. Rather than it detects others shape.

দুজনের দেশে ফেরা

Figure 8.2.1: Detected line property using Hough Transform.

Figure 8.2.2: Rotated by the angle got from Hough Transform.

8.3 Detecting Skew Contour and rotating by its slop

Contours are useful in terms of analyzing shape an object along with detecting and recognize it. Using this system we detect the border of the image. The border of the document a rectangle. So we will say any other detect contours are not useful for. If there is any image, it's contours will also be an rectangle and it will be parallel to it's parent. We then

detect how much the rectangle is rotated perspective to (0, 0) point. The result angle will help us to correct the skew.

সুনামগঞ্জে বয়স্ক ভাতা প্রকল্প

জীবিত ব্যক্তিদের মৃত দেখিয়ে তালিকা থেকে বাদ

সুনামগঞ্জ প্রতিনিধি

সুনামগঞ্জের সদর উপজেলায় বয়স্ক ভাতা পেতেন, এমন চার ব্যক্তিকে 'মৃত দেখিয়ে' তাঁদের ভাতা পাওয়ার সুযোগ থেকে বঞ্চিত করা হয়েছে। তাঁদের বয়স্ক ভাতার বই অন্য চারজনের নামে বরাদ্দ করা হয়েছে। তাঁরাই এখন ভাতা পাচ্ছেন। স্থানীয় একটি ইউনিয়ন পরিষদের দুই সদস্য এই অনিয়মের সঙ্গে জড়িত বলে অভিযোগ পাওয়া গেছে। তবে এ অভিযোগ অস্বীকার করে তাঁরা উপজেলা সমাজসেবা কার্যালয়কে এ ঘটনার জন্য দায়ী করেছেন।

সুনামগঞ্জ সদর উপজেলার রঙ্গারচর ইউনিয়নে এ ঘটনা ঘটেছে। এর মধ্যে দুজন দুই বছর এবং অন্য দুজন এক বছর ভাতা পেয়েছেন।

ভুক্তভোগী ব্যক্তিদের সূত্রে জানা গেছে, ২০১৩ সালে সদর উপজেলার হাসাউড়া গ্রামের মঞ্জিলা খাতুন ও আবদুল মতিন বয়স্ক ভাতার জন্য মনোনীত হন। এরপর তাঁরা টানা দুই বছর ভাতা পান। কিন্তু গত বছর তাঁরা কোনো ভাতা পাননি। এ ছাড়া ২০১৪ সালে রঙ্গারচর গ্রামের আবদুল ছত্তার ও করফুল নেছা বয়স্ক ভাতার জন্য মনোনীত হন। তাঁরা ২০১৪ সালের জানুয়ারি থেকে ডিসেম্বর মাস পর্যন্ত ভাতা পেয়েছেন।

গত সোমবার জানতে চাইলে আবদুল ছাত্তার *প্রথম আলোকে* বলেন, গত বছরের জুন মাসে রঙ্গারচর ইউনিয়ন পরিষদের (ইউপি) সংরক্ষিত ওয়ার্ডের সদস্য ছালেহা বেগম তাঁর ও করফুল নেছার কাছ থেকে বয়স্ক ভাতার বই দুটি নেন। এরপর আর বইগুলো ফেরত দেননি। গত সপ্তাহে এক আত্মীয়ের মাধ্যমে তিনি সমাজসেবা কার্যালয়ে খোঁজ নিয়ে জানতে পারেন, তাঁদের 'মৃত' দেখিয়ে দুজনের নাম কেটে দেওয়া হয়েছে। তাঁদের বদলে অন্য দুজনের নাম প্রতিস্থাপন করা হয়েছে।

সদর উপজেলা সমাজসেবা কার্যালয় সূত্রে জানা গেছে, খাতাপত্রে আবদুল ছত্তারের বইটি দেওয়া হয়েছে উপজেলার নৈগাঙ গ্রামের জাহাঙ্গীর মিয়া এবং করফুল নেছার বইটি দেওয়া হয়েছে একই গ্রামের গয়না বিবিকে।

হাসাউড়া গ্রামের দুই ব্যক্তির কাছ থেকেও একইভাবে ৭ নম্বর ওয়ার্ডের সদস্য আঞ্জব আলী বই নিয়ে আর ফেরত দেননি। উপজেলা সমাজসেবা কার্যালয় সূত্রে জানা গেছে, হাসাউড়া গ্রামের মঞ্জিলা খাতুনের বই সরবরাহ করা হয়েছে একই গ্রামের মফিলা খাতুনের নামে। আবদুল মতিনের বইটি সরবরাহ করা হয়েছে পাশের জাহাঙ্গীরনগর ইউনিয়নের আরেক ব্যক্তির নামে।

রঙ্গারচর ইউপির সংরক্ষিত ওয়ার্ডের সদস্য ছালেহা বেগম বলেন, তিনি বইগুলো ফেরত পেতে ছয় মাস ধরে সমাজসেবা কার্যালয়ে ঘোরাঘুরি করছেন। কিন্তু তাঁকে বলা হচ্ছে, এগুলো অবৈধ, বই দুটি আর ফেরত দেওয়া যাবে না। তিনি বলেন, 'এখন শুনছি, তাঁদের নাকি মৃত দেখিয়ে অন্যদের নামে বই দুটি দেওয়া হয়েছে।'

রঙ্গারচর ইউপির ৭ নম্বর ওয়ার্ডের সদস্য আঞ্জব আলী বলেন, 'দুই বছর ভাতা পাওয়ার পর এখন বলা হচ্ছে, বইয়ে সমস্যা আছে। অথচ এসব বিষয়ে আমরা কোনো কিছুই জানি না। মঞ্জিলা খাতুন প্রায় প্রতিদিন বইয়ের জন্য আমার কাছে এসে কান্নাকাটি করেন। আবদুল মতিনকেও কোনো জবাব দিতে পারছি না।'

এসব বিষয়ে ভারপ্রাপ্ত জেলা সমাজসেবা কর্মকর্তা নুরুল ইসলাম পাটোয়ারী বলেন, বয়স্ক ভাতার কোনো টাকা যাতে অবিতরণকৃত না থাকে, এ বিষয়ে কর্মকর্তাদের ওপর চাপ থাকে। এ কারণে কাউকে না পাওয়া গেলে নাম প্রতিস্থাপন করতে হয়। তাড়াহুড়া করতে গিয়ে অনেক সময় ভুল হয়। তিনি বলেন, 'আশা করি, ভবিষ্যতে আর এ রকম ভুল হবে না। আমি বলেছি, যাঁরা আগে ভাতা পেতেন, তাঁরাই পাবেন। এই চারজনের স্থলে যাঁদের নাম প্রতিস্থাপন করা হয়েছে, প্রয়োজনে পরবর্তী সময়ে তাঁদের ভাতা দেওয়ার ব্যবস্থা করা হবে।'

Figure 8.3.1: Detected Contour (Green Box) and rotated by the angle.

8.4 Reconstructing the Document

After segmentation, The Textual part will be sent to the recognition, it will return an output as text. This text will be used for reconstruction the document again. When the document was segmented, we saved the coordinate points to recognize from where it was segmented and where it was located. We then sorted the position of the component and added to a document one by one.

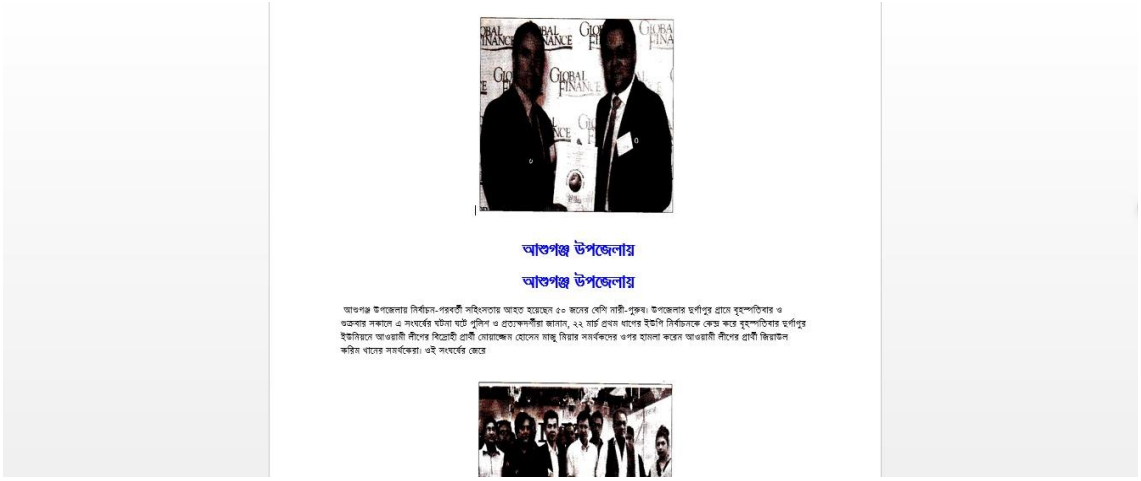


Figure 8.4.1: Screenshot of a portion of a reconstructed Document.

Chapter 9

Result Analysis

The methodology we improved for the decomposition part was mainly tested on various pages from newspaper. A big portion of those testing dataset was taken from "Prothom-Alo". All the scan copy was in 300dpi. We also increased our dataset to test the result. The basic of our analysis is, we took a document count how many images there are, how many paragraphs also how many headlines. Then we compare is the output is correct or not.

9.1 Accuracy for Current Decomposition Methodology

	CONFUSION MATRIX	P R E D I C T E D				
A C T U A L		**Image**	**Table**	**Headline**	**Sub-Headline**	**Column**
	Image	112	1	1	0	0
	Table	2	27	0	0	0
	Headline	1	0	122	2	0
	Sub-Headline	0	0	1	61	1
	Column	0	0	0	4	300

Figure 8.1: Result analysis of current methodology

	TP	TN	FP	FN	Accuracy
Image	112	518	3	2	98.25%
Table	27	615	1	2	93.10%
Headline	122	508	2	3	97.6%
Sub-Headline	61	566	6	2	96.83%
Column	300	334	1	4	98.68%

9.2 Accuracy Comparison with Previous (Batch 2010) Decomposition Methodology

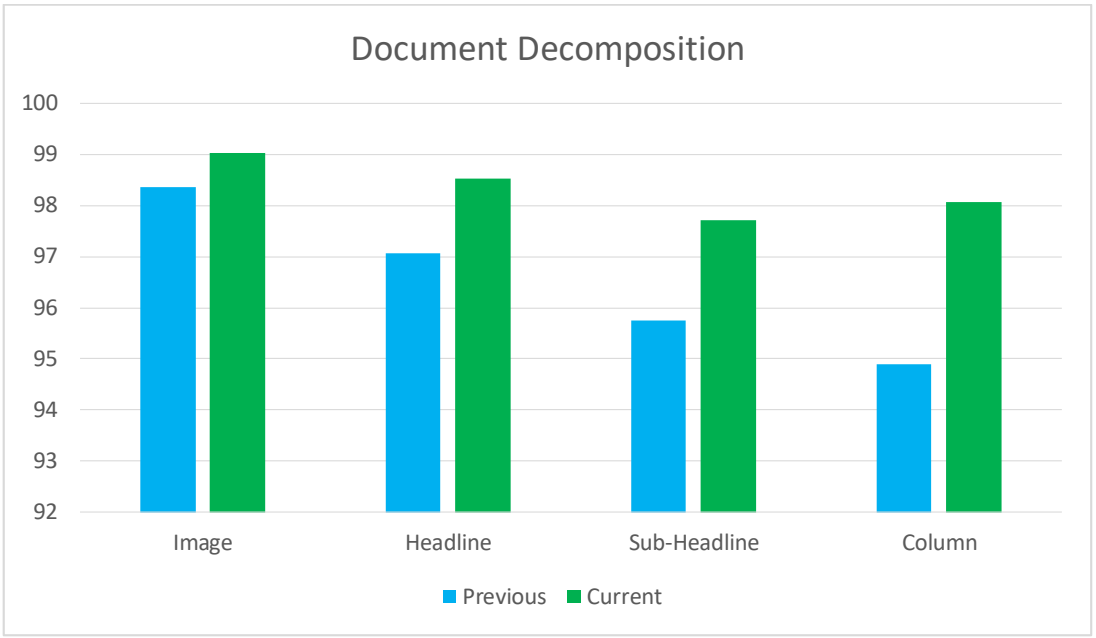


Figure 9.1: Result comparing with previous methodology.

Chapter 10

Conclusion

In this last chapter will discuss about how the whole could be more improved also what necessary steps should be taken.

10.1 Findings in our Developed Methodology

Though we wanted to solve most of the problems for segmenting a Bangla Document, there are still some lacking which should be eliminated soon. Some of these are minor issues, some should be should by another individual team. These cases are discussed below:

i. Current system doesn't have a dynamic range for decomposition. Dynamic range will solve decomposing any kind of documents perfectly. Current system was analyzed and studied by observing the newspaper most of the example was "Prothom Alo" and "Jonokantha".

ii. For the DE Skew Methodology, there will be some problem if the Contours slop is not inside the range of, $-45 >= \text{slop} >= 45$. If we exclude this range the output will not be correct and the system will failed to segment. Also our system cannot detect that, was the given document was upside down or not.

iii. In our Document Reconstruction system, we just append the result by the serial of the component from their position in the document. But this should be reconstructed as the document. Our observation was that, this part should be researched by another team. Because for reconstructing the document format of the output should be well studied as it can always differ for the layout of the document.

References

[1] L. Eikvil, Optical Character Recognition, December 1993.

[2] Wikipedia, "Optical character recognition," [Online]. Available: https://en.wikipedia.org/wiki/Optical_character_recognition. [Accessed 8 April 2016].

[3] blogdna, "Deskew Scanned Images of Documents & Text Pages with Solway Deskew," [Online]. Available: http://www.blogsdna.com/18083/deskew-scanned-images-of-documents-text-pages-with-solway-deskew.htm. [Accessed 9 April 2016].

[4] C. Project, "Detect image skew angle and deskew image," [Online]. Available: http://www.codeproject.com/Articles/104248/Detect-image-skew-angle-and-deskew-image. [Accessed 9 April 2016].

[5] B. S. Mehmet Sezgin, "Survey over image thresholding techniques and quantitative performance evaluation," p. 23, January 2004.

[6] K. Summers, "Automatic Discovery of Logical Document Structure," *Ph. D. Thesis,* 1998.

[7] C. Strouthopoulos, "Text identification for document image analysis using a neural network," *Image and Vision Compution,* vol. 16, no. 12-13, 1998.

[8] Wikipedia, "International Conference on Document Analysis and Recognition," www.wikipedia.org, 4 February 2015. [Online]. Available: https://en.wikipedia.org/wiki/International_Conference_on_Document_Analysis_and_Recognition.

[9] B. Gatos, S. Mantzaris, K. Chandrios, A. Tigris and S. J. Perantonis, "Integrated algorithms for newspaper page decomposition and article tracking," in *ICDAR*, Bangalore, India, 1999.

[10] C. Strouthpoulos and N. Papamarkos, "Text identification for document image analysis using a neural network," *Image and Vision Coputing,* vol. 16, no. 12-13, 1998.

[11] Wahl, M. F., Wong, Y. K., Casey and R. G., Block Segmentation and Text Extraction in Mixed Text/ Image Documents, 1982.

[12] B. Kong, S. Chen, Haralick, M. R., Phillips and T. I., Automatic Line Detection in Document Images Using Recursive Morphological Transforms, 1995.

[13] K. Hadjar, O. Hitz and R. Ingold, "Newspaper Page Decomposition using a Split and Merge Approach," in *International Conference on Document Analysis and Recognition*, Seattle, 2011.

[14] P. Saikrishna and A. G. Ramakrishnan, "Script independent detection of bold words in multi font-size documents.," in *In proceeding of National Conference on Computer Vision, Pattern Recognition Image Processing and Graphics (NCVPRIPG-2013)*, 2013.

[15] D. S. B. F. R. Chen and L. D. Wilcox, "Detection and location of multicharacter sequences in lines of imaged text," *Journal of Electonic Imaging,* vol. 5, p. 37 – 49, January 1996.

[16] D. B. K. R. McConnell and R. Schaphorst, FAX: Digital Facsimile Technology and Applications, Artech House, 1989.

[17] P. N. M. Javed and B. Chaudhuri, "Automatic detection of font size straight from run length compressed text documents," *International Journal of Computer Science and Information Technologies,* vol. 5 (1), p. 818–825, 2014.

[18] A. Z. L. Likforman-Sulem and . B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition,* vol. 9, pp. 123-138, 2007.

[19] J. Lu and D. Jiang, "Survey on the technology of image processing based on dct compressed domain," in *International Conference on Multimedia Technology (ICMT),*, 2011.

[20] P. N. M. Javed and B. Chaudhuri, "Extraction of projection profile, run-histogram and entropy features straight from run-length compressed documents.," in *Proceedings of Second IAPR Asian Conference on Pattern Recognition (ACPR13)*, Okinawa, Japan, 2013.

[21] U. Pal and B. B. Chaudhuri, "OCR in Bangla: an Indo-Bangladeshi language," in *Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International*, 1994.

[22] U. Pal and B. B. Chaudhuri, "A complete printed Bangla OCR system," vol. 31, no. 5, p. 531–549, 1 March 1998.

[23] Wikipedia, "Tesseract (software)," [Online]. Available: https://en.wikipedia.org/wiki/Tesseract_(software). [Accessed 11 April 2016].

[24] H. Yan, "Skew Correction of Document Images Using Interline Cross-Correlation," *Graphical Models and Image Procecssing,* vol. 55, no. 6, pp. 538-543, 1993.

[25] H.-F. Jiang, . C. . c. Han and K.-C. Fan, "A Fast approach to the detection and correction of skew documents," vol. 18, pp. 675-686, 1887.

[26] C. Singh, N. Bhatia and A. Kaur, "Hough Tansform Based Fast Skew Detection and Accurate Skew Correction Methods," *elsevler,* vol. 41, pp. 3528-3546, 2008.

[27] R. Szeliski, Computer Vision: Algorithms and Applications, 2010.

[28] Rich and Knight, Artificial intelligence, 2009.

[29] systransoft, "Machine Translation," [Online]. Available: http://www.systransoft.com/systran/corporate-profile/translation-technology/what-is-machine-translation/. [Accessed 8 April 2016].

[30] elsevier, "Pattern Recognition," [Online]. Available: http://www.journals.elsevier.com/pattern-recognition/. [Accessed 8 April 2016].