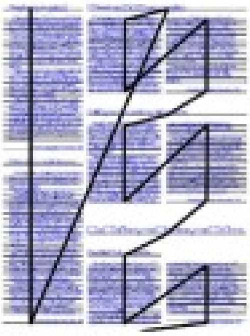
# OCRopus Addons

Internship Report

*Submitted to:*

Image Understanding and Pattern Recognition Lab
German Research Center *for* Artificial Intelligence
Kaiserslautern, Germany



*Submitted by:*

Ambrish Dantrey, B. Tech. III year , E&CE
Indian Institute of Technology, Roorkee
Roorkee, India

Supervisors: Faisal Shafait, Illya Mezhirov
Reviewer: prof. Dr. Thomas Breuel

Start Date for Internship: 15th May, 2007
End Date for Internship: 27th July, 2007

Report Date: 27th July, 2007

# Preface

This report documents the work done during the summer internship at Image Understanding and Pattern Recognition(IUPR) Lab, Deutsche Forschungszentrum für Künstliche Intelligenz(DFKI), Germany under the supervision of Prof. Dr. Thomas Breuel. The report first shall give an overview of the tasks completed during the period of internship with technical details. Then the results obtained shall be discussed and analyzed.
Report shall also elaborate on the the future works which can be persuaded as an advancement of the current work.

I have tried my best to keep report simple yet technically correct. I hope I succeed in my attempt.

Ambrish Dantrey

# Acknowledgments

Simply put, I could not have done this work without the lots of help I received cheerfully from whole IUPR. The work culture in IUPR really motivates. Everybody is such a friendly and cheerful companion here that work stress is never comes in way.

I would specially like to thank Dr. Thomas Breuel and Dr. Daniel keysers for proving the nice ideas to work upon. Not only did they advised about my project but listening to their discussions in IPeT meeting have evoked a good interest in Image analysis. I am also highly indebted to my supervisors Faisal Shafait and Ilya Mezhirov, who seemed to have solutions to all my problems.

Author

# Abstract

The report presents the three tasks completed during summer internship at IUPR which are listed below:

1. Detection of headlines in document images with black run-lengths and OCRopus performance evaluation in detecting headlines
2. Re-engineering the zone-classification module
3. Evaluation of different segmentation algorithms performance

All these tasks have been completed successfully and results were according to expectations. The detection of headlines achieved a low error rate of 2.85% as against 6.52 of previously used methods. During evaluation of segmentation algorithms XY-cut was found to gain a lot by noise cleanup, which is an interesting result as it strengthen the claim of XY-cut segmentation algorithm as a suitable method for OCRopus. The re-engineering and porting of zone-classification module to OCRopus makes it possible for OCRopus to have a text/image segmentation if it is required in future.

Author

## OCRopus : Introduction

Though the field of optical character recognition(OCR) is considered to be widely explored, the development of an efficient system for use in real world situations still remains a challenge for developers. OCRopus is a state-of-the-art document analysis and OCR system, featuring pluggable layout analysis, pluggable character recognition, statistical natural language modeling, multi-lingual capabilities and is being developed at IUPR. This being a very big project, I was assigned the tasks of developing tools for layout-analysis and evaluation.

## The Goals:

Following goals were set as I proceeded in my work:

1. Conversion of ground-truth-data in MARG database from XML format to hOCR micro-format[1].
2. Development of a rule-based headline detection method using the median black run-length of the lines.

3. Development of segmentation-classification module and evaluation of performance of different segmentation algorithms as against noise.

## 1. XML to hOCR:

hOCR is a format for representing OCR output, including layout information, character confidences, bounding boxes, and style information. It embeds this information invisibly in standard HTML. By building on standard HTML, it automatically inherits well-defined support for most scripts, languages, and common layout options. Furthermore, unlike previous OCR formats, the recognized text and OCR-related information co-exist in the same file and survives editing and manipulation. hOCR markup is independent of the presentation.

Due to all above qualities of hOCR format, it is highly desirable to have ground truth in this format. I was assigned the task of converting the MARG database ground truth into hOCR format. For this purpose I have written following script.
Script Name : xml-to-hocr
Language Used: Python
Command-line-argument form: xml-to-hocr FILE.XML

FILE.XML : The file in XML format to be converted into hOCR micro format.

Note: The script does not take care of latex characters yet. It would be an improvement to incorporate this feature.

## 2. Headline detection Based on black run-length and its integration into OCRopus:

Detection of headlines in document images is one issue that is mostly overlooked but yet is highly desirable to properly format the output of OCR. OCRopus had till now used a rule based method which used space between lines as the criteria for detection of headlines. Though this method worked for many images, it also failed many times. It was an obvious observation that black run-lengths of headlines are more than the black run-length of the normal line, and we tried to build upon this concept. We used median black run length of a line as the deciding criteria. The median was used instead of mean because mean run length could have easily been affected by the noise merging with text and would have produce errors.

The whole approach is simple as discussed below:

1. Calculate the median black run-length for the each line on page.
2. Compare this run length for each line with the lines below and above it.
3. If black run-length for a line has been found K1(a parameter) times the median run-length of line below it, and K2(another parameter) times the median run-length of the line above it,set it as a headline.

The value of parameters K1 and K2 was to be found experimentally. After many times evaluating the performance of the program, the value of K1 and K2 has been set to 1.5 and 1.1 respectively.

We used histogram based method to find the median run-length. A histogram of the number of occurrences versus run-length was calculated, once we have such a histogram we normalize it with the largest value of occurrence. Then we calculated the cumulative distribution function for this normalized histogram. The point when cumulative distribution function reches a value of 0.5, corresponds to the median runlength.

The program for detection of headlines was written in C++ and used standard OCRopus classes. The program has been successfully integrated into OCRopus and

## Evaluation:

We also designed a tool which evaluates the performance of the OCRopus in detecting headlines. As according to OCRopus standards, this tool has been developed to work with files in hOCR micro-format. This tool comprises of two programs:

1. The first program takes the OCRopus output and the corresponding ground truth file in hOCR format and outputs the total no of false positives and

false negatives which occurred in detection. It also outputs the total no of true headlines which are present in the ground-truth. The command line form of this programs is:

```
headline-eval hOCR-true hOCR-actual
```

2. The second program is for parsing the file produced by running above program on a large no of files(or on a database) and counts the total no of false positives and false negatives occurred in whole database and tells the error rate of OCRopus on whole database. The command line form of this programs is

```
count_errors FILE.TXT
```

Both of the above programs were written in PYTHON.

**Criteria for evaluation**: For evaluating the performance of OCRopus in detection of headlines we define the the error rate as:

$$e = (fp+fn)/T$$

e = percentage error
fp = total no of false positives
fn = total no of false negatives

We evaluated the performance on standard University of Washington-III (UW-III) database [2]. The results for headline-detection program showed clearly that median black run-length criteria is better than the space between lines criteria, yet errors were still present. While visually analyzing the output, an observation was made that run-length based criteria and space based criteria both produced different false negatives and positives. Hence it was clear that one of the method can be used to remove the errors produced by other. So we tried to combine the both approaches in such a way that space based criteria is used as a filter to detect false positives produced by the run-length based criteria. The rule which was used to combine them was as follows:

1. Use run-length based criteria to find the headlines.
2. Calculate the median black run-length for whole page

3. Compare the median black run-length of all lines found to be headline in step 1 with the median black run-length of the page. Since median black run-length of the page represents just the simple line not a headline, if any headline found in step 1 has a run-length less than or equal to the run-length for whole page, it is a suspicious case. Recheck for this line with space based criteria.

## Results:

The results were as expected. Only run-length based criteria performed better than only space based criteria and a combination of both the criteria as described above outperformed the both. The error rates on standard UW3 database for different approaches are as follows:

**Space based headline detection:**

total no of text lines: 138018
total no of false positives: 7356.0
total no of false negatives: 1713.0
% error = 6.52%

**Black Run-length based headline detection:**

total no of text lines: 138018
total no of false positives: 4341.0
total no of false negatives: 1386.0
% error = 4.14%

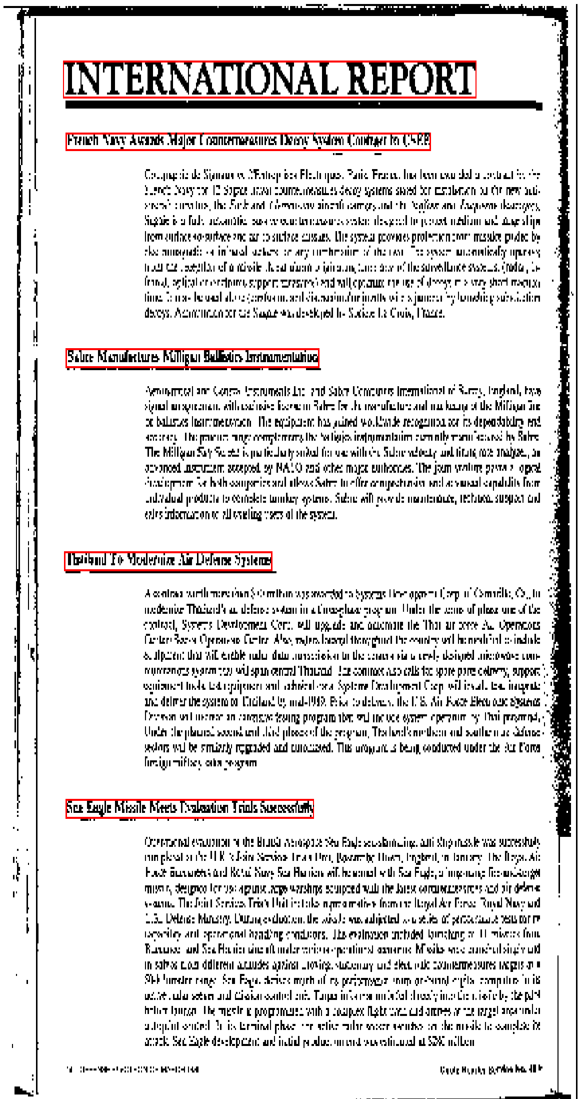
**Both approaches combined (using space based approach as a filter to remove false positives)**

total no of text lines: 138018
total no of false positives: 2452.0

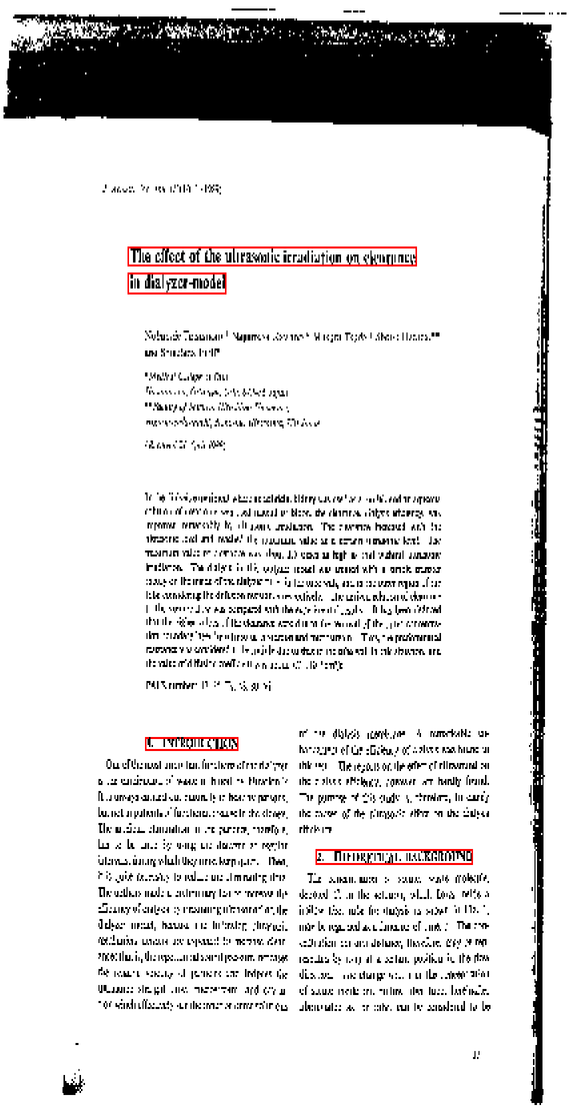
total no of false negatives: 1476.0

% error = 2.85%

Next we show some of the examples:


INTERNATIONAL REPORT
French Navy Awards Major Countermeasures Decoy System Contract to CSEE
Sabre Manufactures Milligan Ballistics Instrumentation
Thailand To Modernize Air Defense Systems
Sea Eagle Missile Meets Evaluation Trials Successfully

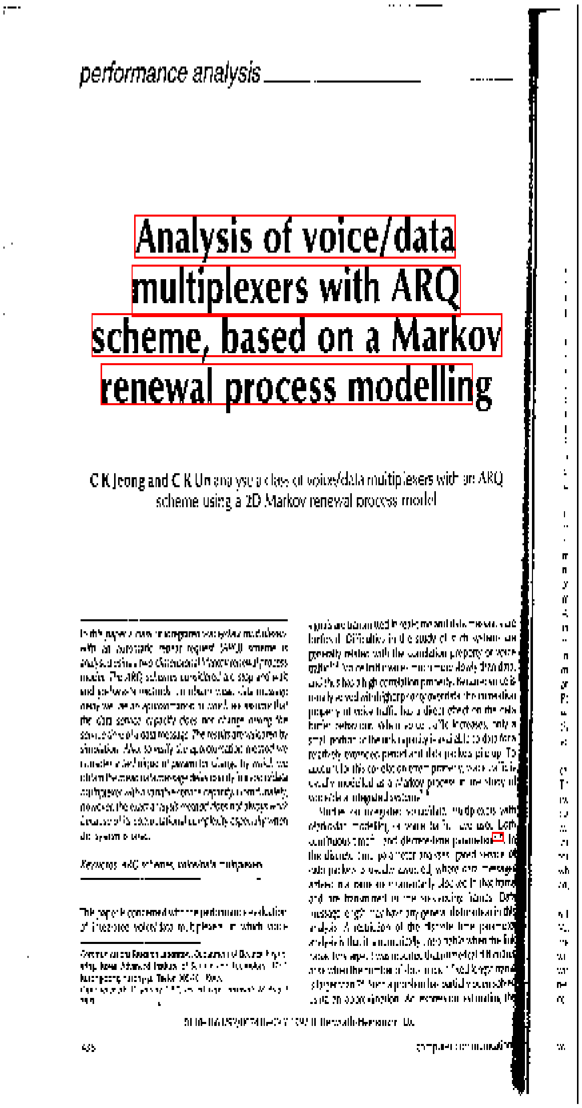
Related work
Stanford Dash
IEEE Scalable Coherent Interface
March 1992



The effect of the ultrasonic irradiation on clearance in dialyzer-model

1. INTRODUCTION

2. THEORETICAL BACKGROUND



performance analysis

Analysis of voice/data multiplexers with ARQ scheme, based on a Markov renewal process modelling

C K Jeong and C K Un analyse a class of voice/data multiplexers with an ARQ scheme using a 2D Markov renewal process model

In this paper a class of integrated voice/data multiplexers with an automatic repeat request (ARQ) scheme is analysed using a two-dimensional Markov renewal process model. The ARQ schemes considered are stop-and-wait and go-back-N methods. [illegible]

Keywords: ARQ schemes, voice/data multiplexers

This paper is concerned with the performance evaluation of integrated voice/data multiplexers in which voice [illegible]

computer communications

## 3. Text/Image Segmentation and Classification

Document image layout analysis is a crucial step in many applications related to document images, like text extraction using optical character recognition (OCR), reflowing documents, and layout-based document retrieval. Layout analysis is the process of identifying layout structures by analyzing page images. Layout structures can be physical (text, graphics, pictures, . . . ) or logical (titles, paragraphs, captions, headings, . . . ). The identification of physical layout structures is called physical or geometric layout analysis, while assigning different logical roles to the detected regions is termed as logical layout analysis [3]. The task of a geometric layout analysis system is to segment the document image into homogeneous zones, each consisting of only one physical layout structure, and to identify their spatial relationship (e.g. reading order). Therefore, the performance of layout analysis methods depends heavily on the page segmentation algorithm used. A detailed explanation of defferent segmentation algorithms and their performance comparison can be found in [4,5].

Also, another important subtask of document image analysis in the classification of physically segmented blocks into one of the predefined classes. In most of the cases the classification steps follows the segmentation and it is highly desirable to evaluate the system performance on whole segmentation/classification task. With the help of such an evaluation, it is easy to decide if the incorporation of these step in OCRopus would result in improved performance. also it would be easy to decide which segmentation algorithm to use.

For classification step we used method as described in [6] this being the best classification method. We used only two classes text and non-text which were relevent to OCRopus, instead of eight classes as described in this paper.

We already had an implementation of various segmentation algorithms and classification step. The task included re-engineering the classification step's code and porting the whole segmentation-classification module into OCRopus, making it use standard OCRopus classes and functions. The task has been completed successfully and now we have a version of whole segmentation-classification module in OCR repository and it can be integrated with OCRopus if the results and experiments comes positive. The command line form of the program is :

*ocr-classify-and-display -i IMAGE -b BOUNDING-BOX-FILE -o OUTPUT-IMAGE*

IMAGE : The image to be classified
BOUNDING-BOX-FILE : The bounding box file produced by segmentation algorithms
OUTPUT-IMAGE : The name of output image to be written

## Evaluation

As discussed earlier the evaluation of both segmentation and classification steps combined together is highly desirable. The purpose of developing a evaluation module was to decide which segmentation-algorithm would best suite the need of OCRopus. We developed a evaluation program which evaluates the performance of two steps as against the ground-truth. Our criteria for the evaluation is the hamming distance between the text/non-text zone image produced from ground truth and that from the Zone-classification module. The error rate is defined as follows:

e = HD*100/T
e = error rate
HD = Hamming distance between ground-truth text-non-text image and actual text-non-text image
T = Total no of pixels present in image

%efficiency = 100-e

This program was developed in C++. The command line argument form of the program is :

```
ocr-evaluate -gt GROUND-TRUTH-IMAGE  -ai ACTUAL-IMAGE
```

GROUND-TRUTH-IMAGE : Text/non-text image produced from ground truth
ACTUAL-IMAGE :  Text/non-text image produced from actual program

**Issue of noise cleanup:** Document Image Noise affects the performance of segmentation algorithms greatly. It was our view that the performance of all the algorithms should improve after noise cleanup. A better explanation can be found in [5]. We used noise cleanup system as explained in [7] Also we expected improvement in performance of simple segmentation algorithms like XYcut to be more than that of complex algorithms like voronoi,  reason being XYcut gets more affected by noise  than voronoi does and as we evaluated the performance of these algorithms with and without noise, we proved correct.

## Results:

Three segmentation algorithms (Voronoi, Docstrum and XYcut )performance was evaluated by our program. The results were as we had expected and hence were quite encouraging. Below are the error rates for all these algorithms with and without noise cleanup.

| Algorithm | Percentage efficiency without noise cleanup | Percentage efficiency with noise cleanup |
|---|---|---|
| Voronoi | 87.03 | 87.69 |
| Docstrum | 86.88 | 86.92 |
| XYcut | 80.16 | 85.70 |

As evident the performance of all the algorithms increase with noise cleanup, but the improvement was much more for XYcut compared to other algorithms. After noise cleanup XYcut has an efficiency much close to that of Voronoi and being a simple algorithms XYcut can be an optimum choice for the OCRopus.

## Conclusion:

The whole experience of working at IUPR was great. This organization has a superb work culture, great minds and very high quality of work. I learned a lot of about image processing and analysis. The work I could complete here was very satisfactory. I have tried to develop as many add-ons as possible for OCRopus and even got very encouraging results with some of them. I hope my work on OCRopus helps it meet its goals.

## References

1. T. M. Breuel: The hOCR Microformat for OCR Workflow and Results : ICDAR,2007 , accepted for publication
2. I. Guyon, R.M. Haralick, J.J. Hull and I.T. Phillips: Data sets for OCR and document image understanding research. In: Handbook of character recognition and document image analysis, World Scientific, (1997) 779–799
3. R. Cattoni, T. Coianiz, , Messelodi, S. Modena, C.M.: Geometric layout analysis techniques for document image understanding: a review. Technical report, IRST, Trento, Italy (1998) *
4. F. Shafait, D. Keysers, and T. M. Breuel : Performance Comparison of Six Algorithms for Page Segmentation: 7th IAPR Workshop on Document Analssis Systems(DAS),pages 368-379
5. F. Shafait, D. Keysers, T. M. Breuel : Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images : ICPR 2006, International Conference on Pattern Recognition, pages 872-875 *
6. T. M. Breuel, D. Keysers, F. Shafait : Document Image Zone Classification- A Simple High-Perfomance Approach : VISAPP 2007, pages 44-51
7. T.Gupta: OCRopus addons: tech reports, IUPR, 2007