

EXPLOITING CONTEXT FOR DOCUMENT IMAGE INTERPRETATION, INFORMATION EXTRACTION AND RETRIEVAL

ANUKRITI BANSAL



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

OCTOBER 2017

© Indian Institute of Technology Delhi (IIT Delhi) - 2017

EXPLOITING CONTEXT FOR DOCUMENT IMAGE INTERPRETATION, INFORMATION EXTRACTION AND RETRIEVAL

by

ANUKRITI BANSAL

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2017

To my parents and brother Anurag

*You were my strength when I was weak
You were my voice when I couldn't speak
You were my eyes when I couldn't see
You saw the best there was in me
Lifted me up when I couldn't reach
You gave me faith because you believed
I'm everything I am
Because you loved me*

*You gave me wings and made me fly
You touched my hand I could touch the sky
I lost my faith, you gave it back to me
You said no star was out of reach
You stood by me and I stood tall
I had your love and I had it all
I'm grateful for each day you gave me
May be I don't know that much
But I know this much is true
I was blessed because I was loved by you*

(Diane Warren)

Certificate

This is to certify that the thesis titled “**Exploiting Context for Document Image Interpretation, Information Extraction and Retrieval**” being submitted by **Ms. Anukriti Bansal** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a record of bona-fide research work carried out by her under my guidance and supervision. In my humble opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Dr. S. Dutta Roy

Associate Professor

Department of Electrical Engineering

Indian Institute of Technology Delhi

New Delhi - 110 016

Acknowledgements

This thesis owes its existence to the continuous help, conscientious support, and timely encouragement of several people. Firstly, I would like to express my deepest regards and most sincere gratitude to my PhD supervisor **Dr. Sumantra Dutta Roy** for his support and guidance. He always encouraged independent thinking and gave me the freedom to work on my ideas without any objection. His expertise in making presentations, technical writing and coding has been a source of motivation for me to learn and develop these skills. Also, he has been very tolerant and empathetic in giving me the freedom to manage both my research work and personal life with ease. I am grateful to him for his forbearance with my ignorance and for being patient with my struggles.

I owe a debt of gratitude to the members of SRC - Prof. S. D. Joshi, Prof. Prem Kalra and Dr. Brejesh Lall for sharing their illuminating views on various issues related to research. I sincerely want to thank Prof. Prem Kalra for his valuable guidance and constructive comments during presentations that helped me notice the weaknesses in my dissertation and make the necessary improvements. I would like to show my gratitude to Dr. Brejesh Lall for his extended support, constant encouragement and valuable coun-

seling, especially during some difficult phases. I must mention Prof. Santanu Chaudhury and Prof. J. B. Srivastava for their advice and support in the early days of the work. I will always be grateful to all the faculty members in Electrical Engineering and Computer Science departments at IIT Delhi, who were instrumental in adding value to my learning in the areas important to my research.

I would like to acknowledge Ministry of Human Resource Development (MHRD), Govt. of India, for financial support. I would also like to thank Department of Electrical Engineering, IIT Delhi for providing all the requisite resources and academic environment suitable for carrying out this research work peacefully. I will always be indebted to IIT Delhi for taking important measures for health and security and providing an opportunity to witness and participate in some of the magnificent technological fests, talks and cultural events.

I owe a special debt of respect and gratitude to the teachers from my former college - Prof. Deepak Bhatnagar, Dr. Sumita Kachhwaha, Dr. R. D. Agarwal and Dr. Sumit Shrivastava, for always believing in me and motivating me to explore new things in life. Their affection, understanding and personal guidance has been of great value for me. Thank you for always providing me your valuable advice and moral support.

A warm thanks to all the staff and lab-mates of Multimedia Lab and Digital Signal Processing (DSP) lab for their help in technical and official matters, as well as, for providing a happy, healthy, and friendly atmosphere in the lab.

I will never forget the chats and beautiful moments I shared with my friends Kiran and Sanchi. They are fundamental in providing support and friendship that I needed dur-

ing stressful and difficult moments. I express my thanks to my other close friends - Pooja Gopal, Arunima Shukla, Nitin Lohar, Susmit Saraswat and Mahima Raina who always tried to keep me cheerful and made my stay in Delhi memorable. I would like to thank my friend Chhanda Sen for helping me in understanding some mathematical concepts, which I was finding very difficult. Those concepts helped me a lot in my research work.

To my brother, Anurag, I owe a large debt of gratitude for his unflinching support, motivation and affection. Thanks for always helping me stay strong and focused. I am blessed and proud to have shared this journey with you. No matter where you are around the world, you are always with me.

I struggle with words to say thanks to my parents, Mrs. Madhu Bansal and Mr. Sitaram Bansal. Thanks Maa and Papa for supporting my dreams and for nourishing them with your love and prayers, while putting yours on hold. You were, are and always will be my greatest inspirations.

Above all I want to thank God for providing me the courage to pursue my dreams and surrounding me with an incredible group of family, mentors, co-workers and friends. To Him goes the credit for providing inspiration and energy to work, and for always showing the right path in life.

Anukriti Bansal

Abstract

The work presented in this thesis explores the use of context information for document image interpretation, information extraction and retrieval of the desired information. We introduce a human perception-based method for representing document images in the form of a graph of homogeneous regions of text and non-text elements. We use this representation for other document understanding operations such as document labeling, extraction of logical units and retrieval of intended information. The algorithm is evaluated on a large collection of documents from scanned newspaper images and standard datasets. We also compare the method with some popular page segmentation algorithms and show encouraging results.

Tables present in documents are important layout entities which are used to compactly communicate significant information in rows and columns. We present a novel learning-based framework to identify tables from scanned document images. The approach is designed as a structural labeling problem, which learns the layout of the document and labels its various entities as table header, table trailer, table cell and non-table region. We develop features which encode the foreground text block characteristics and

the contextual information. These features are provided to a fixed-point model which learns the inter-relationship between the blocks. The fixed-point model attains a contraction mapping and provides a unique label to each block. We compare the results with that of Condition Random Fields (CRFs). Unlike CRFs, the fixed point model captures the context information in terms of neighborhood layout more efficiently. Experiments on the images picked from UW-III (University of Washington) dataset, UNLV dataset and our dataset consisting of document images with multi-column page layouts, show the applicability of our algorithm in layout analysis and table detection.

Next, we present a hierarchical model to extract newspaper articles, which works in two stages. In the first stage, a semantic label (heading, sub-heading, text-blocks, image and caption) is assigned to each newspaper block (segmented using the human-perception-based method, mentioned earlier). The labels are then used as input to the next stage to group related blocks into news articles. A comparison with the state-of-art method shows the applicability of our algorithm for newspaper labeling and article extraction.

In the final work, we explore the utility of layouts and sub-layouts in image retrieval and propose an efficient graph-based matching algorithm, integrated with hash-based indexing, to prune a possibly large search space. During the extraction of layout entities, we handle cases of segmentation preprocessing errors (for text/non-text blocks) with a symmetry maximization-based strategy, and account for multiple domain-specific plausible segmentation hypotheses. A user can specify a combination of sub-layouts of interest using sketch-based queries. The system supports partial matching for unspecified layout

entities. Newspaper images are rich in layout and exhibit a good number of local layout matches with other documents like magazines, newsletters and journal pages. We show promising results of our system on a database of unstructured entities, containing a large number of newspaper images.

सारांश

इस थीसिस में प्रस्तुत कार्य प्रलेख चित्र विश्लेषण, सूचना निष्कर्षण और वांछित जानकारी की पुनर्प्राप्ति के लिए संदर्भ जानकारी के उपयोग की पड़ताल करता है। हम पाठ्य और गैर-पाठ्य तत्वों के समरूप क्षेत्रों को ग्राफ के रूप में प्रलेख चित्रों को दर्शाने के लिए एक मानव धारणा-आधारित विधि प्रस्तुत करते हैं। प्रलेख लेबलिंग, तार्किक इकाइयों का निष्कर्षण और इच्छित सूचना की पुनर्प्राप्ति जैसे अन्य दस्तावेज़ समझने वाले कार्यों के लिए हम इस प्रतिनिधित्व का उपयोग करते हैं। स्कैन अखबार की छवियों और मानक डेटासेट्स के दस्तावेज़ों के एक बड़े संग्रह पर कलन-विधि का मूल्यांकन किया गया है। हम कुछ लोकप्रिय पृष्ठ विभाजन कलन-विधियों के साथ विधि की तुलना भी करते हैं और उत्साहजनक परिणाम दिखाते हैं।

दस्तावेज़ों में उपस्थित सारणी महत्वपूर्ण अभिन्यास संस्थाएं हैं जो पंक्तियों और स्तंभों में महत्वपूर्ण जानकारी के साथ संक्षिप्त रूप से संवाद करने के लिए उपयोग की जाती हैं। स्कैन किए गए दस्तावेज़ छवियों से तालिकाओं की पहचान करने के लिए हम एक नया यन्त्र अधिगम-आधारित ढांचे को प्रस्तुत करते हैं। यह दृष्टिकोण संरचनात्मक लेबलिंग समस्या के रूप में बनाया गया है, जो दस्तावेज़ के लेआउट को सीखता है और इसके विभिन्न संस्थाओं को तालिका शीर्ष लेख, तालिका ट्रेलर, तालिका सेल और गैर-तालिका क्षेत्र के रूप में लेबल करता है। हम फीचर्स विकसित करते हैं जो अग्रभूमि पाठ ब्लॉक विशेषताओं और संदर्भित जानकारी का प्रयोग करते हैं। इन फीचर्स को एक निश्चित बिंदु मॉडल (फिक्स्ड पॉइंट मॉडल) के लिए प्रदान किया जाता है जो ब्लॉकों के बीच अंतर संबंध सीखता है। निश्चित बिंदु मॉडल एक संकुचन मानचित्रण प्राप्त करता है और प्रत्येक ब्लॉक के लिए एक अद्वितीय लेबल प्रदान करता है। हम परिणामों की तुलना कंडीशन रैंडम फील्ड (सीआरएफ) के साथ करते हैं। सीआरएफ के विपरीत, निश्चित बिंदु मॉडल प्रतिवेश लेआउट के अनुसार संदर्भ जानकारी को अधिक कुशलता से सीखता है। UW-III (वाशिंगटन विश्वविद्यालय) डेटासेट, यूएनएलवी डाटासेट और मल्टी-

कॉलम पृष्ठ लेआउट वाले दस्तावेज़ छवियों के साथ हमारे डेटासेट से चित्रों का प्रयोग, लेआउट विश्लेषण और तालिका पहचान में हमारी कलन विधि की प्रयोज्यता दिखाते हैं।

इसके बाद, हम समाचार-पत्रों के आलेखों को निकालने के लिए एक पदानुक्रमित मॉडल पेश करते हैं, जो दो चरणों में काम करता है। पहले चरण में, एक सिमेंटिक लेबल (शीर्षक, उप-शीर्षक, टेक्स्ट-ब्लॉक्स, इमेज और कैप्शन) को प्रत्येक अखबार ब्लॉक (मानवप्रेषण-आधारित विधि का उपयोग करके खंडित किया गया है, जो पहले उल्लेख किया गया है) को सौंपा गया है। तब लेबलों को अगले लेख में इनपुट के रूप में इस्तेमाल किया जाता है ताकि संबंधित ब्लॉक अखबारों के आलेखों के अनुसार समूहबद्ध हो। अत्याधुनिक पद्धति के साथ एक तुलना पत्रों की लेबलिंग और लेख निष्कर्षण के लिए हमारी कलन विधि की प्रयोज्यता दर्शाती है।

अंतिम काम में, हम छवि पुनर्प्राप्ति में लेआउट्स और उप-लेआउट की उपयोगिता का पता लगा सकते हैं और संभवतः बड़े खोज स्थान को छांटने के लिए हैश-आधारित अनुक्रमण के साथ एकीकृत एक कुशल ग्राफ़-आधारित मिलान कलन विधि का प्रस्ताव करते हैं। लेआउट संस्थाओं के निष्कर्षण के दौरान, हम सममिती अधिकतम-आधारित रणनीति के साथ खंडन पूर्वप्रक्रमित त्रुटियों (पाठ / गैर-पाठ ब्लॉकों के लिए) के मामलों को संभालते हैं, और कई अनुक्षेत्र विशिष्ट प्रशंसनीय विभाजन अवधारणाओं को प्रस्तुत करते हैं। एक उपयोगकर्ता स्केच आधारित प्रश्नों का उपयोग करके रुचि के उप-लेआउट के संयोजन को निर्दिष्ट कर सकता है। सिस्टम अनिर्दिष्ट लेआउट संस्थाओं के लिए आंशिक मिलान का समर्थन करता है। समाचार पत्र की छवियाँ लेआउट में समृद्ध होती हैं और अन्य दस्तावेजों जैसे पत्रिकाओं, न्यूज़लेटर्स और जर्नल पेजों के साथ स्थानीय लेआउट मेल की एक अच्छी संख्या का प्रदर्शन करती हैं। हम असंरचित संस्थाओं के डेटाबेस पर हमारी पद्धति के अच्छे परिणाम दिखाते हैं, जिसमें बड़ी संख्या में समाचार-पत्र की छवियाँ होती हैं।

Table of Contents

Certificate	i
Acknowledgements	iii
Abstract	vii
Hindi Abstract	x
List of Figures	xix
List of Tables	xxvii
1 Introduction	1
1.1 Objectives and Scope of Work	3
1.2 Major Contributions of the Thesis	6
1.3 Layout of the thesis	9
2 Related Work	11
2.1 Document Layout Analysis and Representation	12

2.2	Table Extraction from Document Images	16
2.2.1	Tables with ruling lines	17
2.2.2	Tables without ruling lines	18
2.2.3	Tables in PDF/electronic documents	20
2.3	Article Extraction from Newspaper Images	22
2.3.1	Scanned newspaper images	23
2.3.2	Portable document format and electronic documents	25
2.3.3	Web pages	26
2.4	Document and Information Retrieval	27
2.5	Motivation for the Present Work	31
3	Graph-Based Document Representation	35
3.1	Introduction	35
3.2	Segmenting images into text and non-text regions	37
3.2.1	The P/N Ratio	38
3.2.2	Parameter Optimization	39
3.2.3	Adaptive Segmentation	40
3.2.4	Results and Discussion	42
3.3	Graph-Based Representation	44
3.3.1	Preprocessing	46
3.3.2	Symmetry Maximization	49
3.3.3	Results and Discussion	50

3.4	Conclusions	53
4	Identification and Extraction of Tables from Document Images	55
4.1	Introduction	55
4.1.1	Contributions of this work	58
4.1.2	Overview of the System	60
4.2	Model Description	61
4.3	Extraction of Blocks	64
4.4	Feature Set	64
4.4.1	Appearance Features	65
4.4.2	Contextual Features	66
4.4.3	Identifying White Space Separators	67
4.4.4	Identifying Horizontal and Vertical Ruling Lines	67
4.4.5	Neighborhood Estimation	68
4.5	Block Labeling	69
4.6	Experiments and Results	72
4.6.1	Labeling Statistics and Results	73
4.6.2	Comparison with CRF	76
4.7	Conclusions	77
5	Newspaper Labeling and Article Extraction	79
5.1	Introduction	79
5.1.1	Contributions to this work	83

5.1.2	Overview	85
5.2	Mathematical Formulation	86
5.3	Preprocessing and Document Representation	89
5.4	Feature Set	91
5.4.1	Appearance Features	91
5.4.2	Contextual Features	92
5.4.3	Neighborhood Estimation	93
5.5	Block Labeling	94
5.6	Block Grouping and Article Extraction	95
5.7	Results and Discussion	96
5.7.1	Newspaper Labeling and Article Extraction Statistics	97
5.7.2	Comparison with CRF	100
5.8	Conclusions	101
6	Information Extraction and Retrieval Using Layout Components	103
6.1	Introduction: Layouts and Sub-layouts	103
6.1.1	Overview of the System	108
6.1.2	Contributions	109
6.2	Preprocessing and Document Representation	110
6.2.1	Generating Multiple Segmentation Hypotheses	111
6.3	Query Formulation	114
6.4	The Proposed Search and Retrieval Procedure	116

6.4.1	Hashing-based Pruning of the Search Space	118
6.5	Results and Discussion	119
6.5.1	Handling different types of Queries	121
6.5.2	Handling Combinations of Multiple Sub-layouts	123
6.5.3	Ranking of Results and Retrieval Statistics	123
6.5.4	Comparison With a Content-Based Document Image Retrieval Approach	125
6.6	Conclusions	128
7	Conclusions	131
7.1	Contributions	132
7.2	Scope for Future Work	134
	Bibliography	137
	Publications	154
	Biography	156

List of Figures

1.1	Layout analysis preprocessing enables proper OCR output. (a) shows the original document, (b) shows the output of a commercial OCR system on the same, (c) shows the output a layout analysis extracting article of interest, and (d) shows the OCR output of the article in (c).	3
1.2	Work flow of the thesis	5
3.1	Overview of the system for segmenting document image into text and graphics	39
3.2	In each pair, the original image is to the left, and the segmented one, to the right	43
3.3	Comparison between RLSA and ARLSA: (a) The original image, (b) output of RLSA with structuring element of smaller size giving over-segmentation errors, (c) output of RLSA with structuring element of larger size giving under-segmentation errors, and (d) output of ARLSA, structuring element according to size of connected components, with lesser over-segmentation and under-segmentation errors.	47

3.4	(a) The original image, (b) the output of text-non-text segmentation, and (c) text blocks using ARLSA.	48
3.5	Results of symmetry maximization: In each pair, the original image is to the left, and the segmented one, to the right.	51
3.6	Comparing layout extraction with Docstrum and AOSM: (a) represents the ground truth image, (b) result obtained from Docstrum, which fails to separate text blocks, as the distance between them is very small, (c) result obtained from AOSM, which fails to separate fonts with different sizes as they are aligned and present close to each other, (d) result obtained from our method which handles such cases effectively.	52
4.1	Different types of tables in document images (a) Table with horizontal lines, where color is an important feature to detect tables correctly, (b) Table with both horizontal and vertical lines (c) Tables without any ruling lines on a document with two column page layout	59
4.2	A block diagram showing an overview of the system	60
4.3	Figure showing preprocessing step for extraction of foreground block, (a) The original Image (b) blocks obtained after preprocessing step explained in Section 4.3	65
4.4	(a) An example of a document image, (b) the horizontal distance map, and (c) the vertical distance map of the background/whitespace region, with text regions shown in pink	66

4.5	Neighborhood Estimation: Current block marked in blue, neighbors at m = 1 marked in red, neighbors at $m = 2$ marked in red and green	68
4.6	The set of blocks input to the block labeling process: text blocks (gray), thick whitespace without ruling (blue), thin whitespace without ruling (red), thick whitespace with ruling (pink)	69
4.7	Results of labeling each document element as table-heading (blue), table- trailer (red), table-cell (green), non-table regions (yellow).	74
4.8	Results with incorrect labeling of table trailer blocks: (a),(d) Original Im- ages (b),(e) Ground truth images with correct labeling, (c) Results of our approach which misclassified table-trailer blocks as table cell, (f) Results of false detection of non-table regions as table region.	75
4.9	Comparison of Fixed point model and CRF: (a) Original Image (b) La- beling results of Fixed point model, (c) Labeling results of CRF	76
5.1	(a) A sample document image, (b) output of a commercial optical char- acter recognition (OCR) system, (c) shows an article extracted from (a) after layout analysis, and (d) shows the OCR output of (c)	80
5.2	Example illustrating the use of context information in correctly determin- ing the label of text block shown in (a), (b) is a part of a sample newspaper image. (Details in the text, p. 82	82

5.3	Sample newspaper images in different languages to illustrate that article headings, and thereby complete articles, can be located in newspaper images irrespective of the understanding of the language: (a) Newspaper image in English language, (b) Newspaper image in Malayalam language	84
5.4	A block diagram showing the overview of the system	85
5.5	Homogeneous regions obtained after preprocessing: (a) Original image, (b) Homogeneous regions obtained using the technique explained in Chapter 3, (c) Homogeneous regions obtained after merging the large blocks, which will be used as input for newspaper labeling and article extraction.	89
5.6	Neighborhood Selection: the current block is marked in blue, neighbors at $m = 1$ are marked in red, and neighbors at $m = 2$ are marked in red and green.	93
5.7	Results of article extraction: (a) the original newspaper image; (b) Results of Labeling each block as heading (red), sub-heading (blue), text-block (green), image (orange), caption (violet); (c) different articles identified. .	97
5.8	Comparison with CRF: (a) the original newspaper image; (b) Results of Labeling each block as heading (red), sub-heading (blue), text-block (green), image (orange), caption (violet) using proposed method; (c) labeling results obtained using CRF	101

6.1	Application of sub-layout-based retrieval: (a) represents the query for image region (shown in pink color) sandwiched between text regions (shown in blue color) on both the sides and present at the center of the page. (b), (c), (d) are the examples of retrieved document images, in which overall layout is different but contains the specified sub-layout.	105
6.2	Our system is invariant to scale and translation: an example. The first image is the query image and the second one is the retrieved image from the database, whose blocks are of different sizes, aspect ratios and are present at absolute positions different from those in the query	107
6.3	Homogeneous regions obtained after preprocessing: (a) Original image, (b) Homogeneous regions obtained using the technique explained in Chapter 3.	110
6.4	Multiple segmentation hypotheses: (a) The ARLSA output (Sec. ??), (b) Hypothesis with small (e.g., noise) or insignificant blocks (author blocks, for instance) removed, (c) Hypothesis with grouped adjacent non-text regions, and (d) Hypothesis with caption blocks removed.	112

6.5	Various types of query layouts: Blue represents text, pink represents non-text non-background blocks, and gray indicates that specific block type (text/non-text) is irrelevant. (a),(b) All blocks have their type specified, without any missing blocks, (c) The specific block type (text/non-text) is not relevant for any block, and there are no missing blocks. (d) Some blocks need to be retrieved without bothering about their specific type (text/non-text), and there are no missing blocks. (e) Some blocks missing, with the block type specified for all blocks. (f) Missing blocks, and block type (text/non-text) is irrelevant for all blocks. (g) Missing blocks, the type is specified for a few blocks.	114
6.6	(a) The reference block (highlighted in yellow) in the query image. In this example, without the Hashing-based pruning, all blocks (in (b)) get selected as candidate reference blocks (as shown in (c)). The Hashing-based pruning strategy (Sec. 6.4.1) helps to prune the large search space, as in (d).	118
6.7	Retrieved Results: (a) Result for query layout shown in Fig. 6.5(b), (b) Result for query layout shown in Fig. 6.5(c). For the query layout shown in Fig. 6.5(d), the gray block substituted by non-text block (pink color) in (c) and by text block (blue color) in (d). (e) shows the partial matching result for query layout shown in Fig. 6.5(g). (f) shows retrieved results for query layout shown in Fig. 6.5(f): the middle missing block substituted by a non-text one, and the ones to the right, as three text blocks.	120

6.8	The system is language and script-independent since it considers the relative arrangement of blocks. For a Hindi newspaper in Devanagari script, the images above show the block structure, and results of successful retrieval (Query Type 6, Figure 6.5(g)).	122
6.9	Handling irregular layouts: the two examples above correspond to lines of text interspersed with images of irregular dimensions. The system considers the closest rectangular bounding boxes and is able to perform correct retrieval: these correspond to Figures 6.7(a) and 6.7(b).	123
6.10	Retrieval results for a combination of multiple sub-layouts in different spatial locations (Sec. 6.3). Query (<i>A, bottom</i>) AND (<i>B</i>) AND (<i>NOT C</i>) (Sec. 6.5.2) results in left-most result alone, and not the other two. The middle one does not satisfy the spatial location requirements for sub-layout <i>A</i> , whereas the rightmost one also contains layout <i>C</i>	124
6.11	Ranked results for the query layout in (a): The system first generates a dummy block. The system ranks results in order of relative average discrepancy in block aspect ratios, and relative positions	125

List of Tables

3.1	Table showing average performance accuracy of our algorithm vs Docstrum and AOSM algorithms	53
4.1	Table showing experimental results of block labeling with four classes: table-header, table-trailer, table-cell, non-table	71
4.2	Comparison of block labeling with CRF	73
5.1	Experimental results of newspaper block labeling	98
5.2	Experimental results of Article Extraction	98
5.3	Comparison of block labeling with CRF	100
6.1	Statistics of search and retrieval procedure	126
6.2	Comparison of the computation efficiency of the content-based and the proposed sub-layout-based retrieval methods	128

