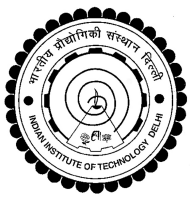
EXPLOITING CONTEXT FOR DOCUMENT IMAGE INTERPRETATION, INFORMATION EXTRACTION AND RETRIEVAL

ANUKRITI BANSAL



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2017

EXPLOITING CONTEXT FOR DOCUMENT IMAGE INTERPRETATION, INFORMATION EXTRACTION AND RETRIEVAL

by

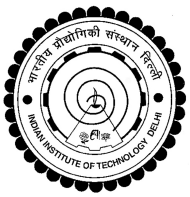
ANUKRITI BANSAL

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
Your 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 दस्तावेजों 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।

दस्तावेजों 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 है। 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 अख़बार 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 छवियां होती हैं।

Table of Contents

List of Figures

List of Tables