

**ON IMPROVING TECHNIQUES FOR
ACCESSING CONTENT IN DOCUMENT
IMAGE COLLECTIONS**

RITU GARG



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 2016

© Indian Institute of Technology Delhi (IITD), New Delhi, 2017

**ON IMPROVING TECHNIQUES FOR
ACCESSING CONTENT IN DOCUMENT
IMAGE COLLECTIONS**

by

RITU GARG

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

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

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 2016

To my family

Certificate

This is to certify that the thesis titled **On Improving Techniques for Accessing Content in Document Image Collections** being submitted by **Ms. Ritu Garg** to the **Department of Electrical Engineering**, Indian Institute of Technology Delhi, for the award of **Doctor of Philosophy** is a record of bona-fide research work carried out by her under my guidance and supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Professor Santanu Chaudhury
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110016, India.

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Santanu Chaudhury. The thesis would not have its present state without his guidance and encouragement. I have had a long association with Prof. Santanu Chaudhury, when he first appointed me as project associate in 2005. He has always motivated me to strive for excellence and put in my hundred percent to the job at hand. My sincere thanks to Santanu sir for providing financial support all throughout my Phd and other research related activities.

I thank the members of my thesis committee: Prof. S. D. Joshi, Prof. P. K. Kalra, and Dr. Brejesh Lall, for their insightful suggestions which encouraged me to widen my research from various perspectives. I am grateful to my fellow lab mates especially Dr. Ehtesham Hassan, Dr. Ayesha Choudhary, Dr. Anupama Mallik, Anupama Ray for the stimulating discussions, and for the sleepless nights we were working together before deadlines. I have been very fortunate in having Nisha, Vitesh and Utkarsh as my friends, who stood by me through the good and bad. At the end I acknowledge the most important contribution of my parents, who formed part of my vision and taught me good values that really matter in life. Their infallible love and support has always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I would like to express my heartiest gratitude towards my husband Sachin who stood by me patiently through all my efforts and encouraged me to peruse my PhD. Finally, I am thankful to my daughter Myrah who's love motivates me to work harder and achieve the best in life to make her proud.

Ritu Garg

Abstract

Major digitization efforts around the world has resulted in archiving rare and precious books in Indian languages. Effective access to such repository is limited due to heterogeneous nature of the document image collections. In this thesis, we explore different challenges and problems associated with effective access to the content in document collections and present machine learning based solutions for the same. We develop trainable algorithm for estimating the ideal parameter settings and apply it for document image enhancement and separation of text from graphics. Depending on the quality of the document images the performance of the digitization process is greatly influenced. In such environment, it is critical to have good image quality assessment method to help control the digitization performance. We present a document image enhancement scheme based upon assessment of the image quality. The experimental evaluation is presented on document collections belonging to Indian and English script.

Next, word image based document image retrieval scheme is presented. We introduce a compression scheme that exploits the basic geometric primitives to represent the word image skeleton and apply it for retrieval and content adaptation application. The word image retrieval framework presented uses proposed representation with Latent Semantic Analysis (LSA) and Probabilistic LSA for retrieving document images. A SVG representation derived from the word structure primitives for rendering and accessing document images through common browsers on desktop or mobile devices is presented. Experimental results are shown on Devanagari, Bangla, and Telugu script documents.

Further, to improve the access to content in document images we introduce an active learning based approach for improving the optical character recognition (OCR) accuracy. Experi-

mental evaluation of the proposed framework is shown on document collections of Devanagari, Bangla and Telugu script. Finally, method of document image retrieval using multi-modal information fusion is presented. A multi-modal indexing scheme for retrieving document images is presented by learning based combination of text and info-graphics. The evaluation is shown on English document images.

सार

दुनिया भर में प्रमुख डिजिटलीकरण प्रयासों ने दुर्लभ और बहुमूल्य पुस्तकों को संग्रहित किया है। भारतीय भाषाओं इस तरह के रिपॉजिटरी तक प्रभावी पहुंच सीमित है क्योंकि इसके विषम प्रकृति की वजह से है दस्तावेज़ छवि संग्रह इस थीसिस में, हम विभिन्न चुनौतियों और समस्याओं का पता लगाते हैं। दस्तावेज़ संकलन और वर्तमान मशीन में सामग्री के प्रभावी पहुंच से जुड़ उसी के लिए आधारित समाधान सीखना हम आदर्श के आकलन के लिए प्रशिक्षित एल्गोरिदम विकसित करते हैं। पैरामीटर सेटिंग्स और दस्तावेज़ छवि वृद्धि और पाठ से अलग करने के लिए इसे लागू करें ग्राफिक्स। दस्तावेज़ छवियों की गुणवत्ता के आधार पर डिजिटलीकरण का प्रदर्शन प्रक्रिया बहुत प्रभावित होती है ऐसे माहौल में, अच्छी छवि गुणवत्ता रखने के लिए महत्वपूर्ण है। डिजिटलीकरण प्रदर्शन को नियंत्रित करने में सहायता करने के लिए मूल्यांकन विधि हम एक दस्तावेज़ छवि पेश करते हैं। छवि गुणवत्ता के मूल्यांकन के आधार पर वृद्धि योजना प्रायोगिक मूल्यांकन भारतीय और अंग्रेजी लिपि से संबंधित दस्तावेज़ संग्रह पर प्रस्तुत किया गया है।

अगला, शब्द छवि आधारित दस्तावेज़ छवि पुनर्प्राप्ति योजना प्रस्तुत की गई है। हम एक परिचय संपीड़न स्कीम जो शब्द की छवि का प्रतिनिधित्व करने के लिए मूलभूत ज्यामितीय पुरालेखों का उपयोग करती है। कंकाल और पुनर्प्राप्ति और सामग्री अनुकूलन आवेदन के लिए इसे लागू करें। शब्द छवि पुनर्प्राप्ति ढांचा प्रस्तुत प्रस्तावित प्रस्तुति का उपयोग गुप्त सिमेंटिक विश्लेषण (एलएसए) के साथ करता है। और दस्तावेज़ छवियों को पुनः प्राप्त करने के लिए संभाव्य एलएसए। एक एसवीजी प्रतिनिधित्व से व्युत्पन्न सामान्य के माध्यम से दस्तावेज़ चित्रों को प्रतिपादन और एक्सेस करने के लिए शब्द संरचना पुरातनताएं डेस्कटॉप या मोबाइल उपकरणों पर ब्राउज़रों को प्रस्तुत किया गया है। प्रायोगिक परिणाम देवनागरी पर दिखाए जाते हैं, बांग्ला, और तेलगू स्क्रिप्ट दस्तावेज़।

इसके अलावा, दस्तावेज़ चित्रों में सामग्री की पहुंच में सुधार करने के लिए हम एक सक्रिय सीखने का परिचय देते हैं। ऑप्टिकल कैरेक्टर मान्यता (ओसीआर) सटीकता में सुधार के लिए आधारित दृष्टिकोण प्रस्तावित ढांचे का प्रायोगिक मूल्यांकन देवनागरी के दस्तावेज़ों के संग्रह में दिखाया गया है। बांग्ला और तेलगू स्क्रिप्ट अंत में, मल्टी-मोडल सूचना का उपयोग करते हुए दस्तावेज़ छवि पुनर्प्राप्ति की विधि संलयन प्रस्तुत किया जाता है दस्तावेज़

छवियों को पुनः प्राप्त करने के लिए एक बहु-मॉडल अनुक्रमणिका योजना पाठ और सूचना-ग्राफिक्स के आधार संयोजन को सीखकर प्रस्तुत किया गया है। मूल्यांकन दिखाया गया है अंग्रेजी दस्तावेज़ छवियों पर।

Contents

Certificate	iii
Acknowledgements	v
Abstract	vii
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 Scope and Objectives	2
1.2 Major Contribution of the Thesis	5
1.3 Layout of the Thesis	7
2 Accessing Content in Document Image Collections : A Review	9
2.1 Document Image Analysis	10
2.1.1 Document Image Enhancement	10
2.1.2 Learning based Methods for Document Image Enhancement	12
2.1.3 Document Image Segmentation	13
2.2 Intelligent Access to Document Images	15
2.3 Recognition of Document Images	18
2.4 Retrieval from Document Images	21

2.5	Motivation for the Present Work	25
3	Document Image Analysis for Information Extraction	29
3.1	Introduction	29
3.2	Parameter Optimization using EM Algorithm	30
3.2.1	Document Image Representation	31
3.2.2	Optimization Framework	32
3.3	Document Image Pre-processing : Binarization	33
3.3.1	Experimental Results and Discussion	36
3.4	Adaptive Text/Graphic Segmentation	40
3.4.1	Experimental Results and Discussions	45
3.5	Conclusion	50
4	Document Image Enhancement Using Image Quality Assessment	53
4.1	Introduction	53
4.2	Document Image Quality Assessment Methodology	54
4.3	Parameter Estimation using EM Algorithm	57
4.4	Experimental Results and Discussion	58
4.5	Document Quality Assessment : Language Model Based Approach	62
4.6	Conclusions	66
5	Accessing Document Images in Indian Scripts	67
5.1	Introduction	67
5.2	Word Image Representation : Geometric Feature Graph	69
5.2.1	Algorithm for GFG Extraction	70
5.2.2	Encoding GFG	72
5.2.3	Word Image Reconstruction from GFG	73
5.2.4	GFG Compression	74
5.3	Word Image Indexing using GFG Representation	75

5.3.1	Semantic Indexing using GFG	76
5.3.2	Results of using LSA and PLSA	79
5.4	Interactive Access to Document Images	82
5.4.1	Constructing SVG from GFG representation	83
5.4.2	Experiments: Interactive Access to Document and Compression Using SVG	87
5.5	Conclusion	89
6	Active Learning for Incremental improvement of OCR	91
6.1	Introduction	91
6.2	Propose Framework : Active OCR	93
6.2.1	Active Sample Selection	95
6.2.2	Improving Classifier : An Incremental Approach	97
6.3	Experimental Results	99
6.4	Conclusion	102
7	Multi-modal Information Retrieval	105
7.1	Introduction	105
7.2	Overview : Proposed Document Indexing Framework	107
7.3	Multi-objective Multi-Modal Distance Based Hashing for Document Image In- dexing	109
7.3.1	Kernel Distance Based Hashing : Review	110
7.3.2	Optimization Problem Formulation : MKL Based Indexing	112
7.3.3	GA based Multi-Objective Optimization for MKL Based Indexing	114
7.4	Indexing Framework for Degraded Document Image Collection	117
7.4.1	Experimental Results	118
7.4.2	Dataset Description	119
7.4.3	Features and Parameters	120
7.4.4	Experimental Results	124

7.5	Info-graphics Retrieval: A Multi-Kernel Distance Based Hashing Scheme	129
7.5.1	Multi-modal Infographics Indexing Framework	131
7.5.2	Data Description	133
7.5.3	Features and Parameters	133
7.5.4	Experimental Results	134
7.6	Conclusions	136
8	Conclusions	137
8.1	Summary of the Contributions	138
8.2	Scope of Future Work	140
	Bibliography	141
	Publications	169
	Biography	170

List of Figures

2.1	Sample Info-graphics	27
3.1	Sauvola Binarization of sample gray-scale document image with non-uniform illumination	37
3.2	Binarization of gray-scale document image with ink-bleeds	38
3.3	Binarization result of blurred gray-scale document image	38
3.4	Sample Pages with different type of degradation	40
3.5	(a) Original Document Image (b) After Connected Component Analysis	42
3.6	Sample images and their horizontal projection profile and autocorrelation plot	42
3.7	P/N ratio for sample text and graphics block	44
3.8	Experimental Results: (a) Sample document images containing text/graphics, (2) Segmentation results by Abbyy FineReader, (c) Segmentation output by our proposed approach	45
3.9	Segmentation Results: (a) Sample document images containing text/graphics, (2) Segmentation results by Abbyy FineReader, (c) Segmentation output by our proposed approach	46
3.10	Segmentation comparion with commercial softwares	48
3.11	Result of proposed adaptive binarization and segmentation on newspaper article image	50
3.12	(Continued) Experimental results of adaptive binarization and segmentation on newspaper article image	51

4.1	Parameter Estimation Framework using DIQA Methodology	54
4.2	Samples pages from English, Hindi and Bangla Datasets	59
4.3	Histogram of OCR Accuracy.	60
4.4	Binarization Result with DocIQA model, (a) Original English and Bangla sample images, (b) Binarization results with optimal Sauvola Binarization and (c) Binarized output at fixed k	62
4.5	Binarization comparison with Leptonica and other techniques	62
4.6	Modified Framework Using Statistical Language Model for Parameter Estimation	65
5.1	Basic Shape Primitives used for Word Image Representation	70
5.2	Tangent Angle Plot	71
5.3	Word Image Reconstruction using GFG string.	73
5.4	Word Image Reconstruction for Devanagari, Bangla, Telugu using GFG string	74
5.5	Retrieval statistics using LSA for Indian Document Image Collections.	78
5.6	Overview of methodology	78
5.7	Retrieval statistics using PLSA for Indian Language Document Collections.	81
5.8	Semantic coherence score for PLSA for Hindi, Bangla and Telugu document collections.	81
5.9	(a) Example binary word image (b) Word image after thinning (c) SVG reconstructed word image from GFG string	86
5.10	Sample SVG reconstruction on browser and hand-held devices	88
6.1	Overview of the Proposed Active Learning Approach	93
6.2	Block Diagram for Web OCR	94
6.3	Flow diagram for Symbol extraction	94
6.4	Illustration of scheme for identifying Confusing Classes and Class Ranking	96
6.5	Incremental training a sub-graph of DDAG SVM for sample belonging to class 2	98

6.6	Results of Active Selection with Noise Rejection Vs Random Selection Without Noise Rejection For Bangla Script	100
6.7	Results of Active Vs Random Selection with Noise Rejection for Bangla Script	101
6.8	Active Selection with Noise Rejection Vs Random Selection Without Noise Rejection for Telugu Script	101
6.9	Example document images from Google Book Search with corresponding OCR'ed text	102
7.1	Architecture of multi-modal document indexing framework	108
7.2	Sample Document Images	120
7.3	Retrieval Statistics for Devanagari and Bangla Dataset	126
7.4	Continued Retrieval Statistics for Telugu and English Dataset	127
7.5	Top 5 Retrieved Results of the proposed Word Based Document retrieval framework	128
7.6	Sample Multi-modal Document Images where Text and Info-graphics coexists	130
7.7	Overall architecture for multi-modal document indexing	131

List of Tables

3.1	Optimization Framework : Sauvola Binarization	36
3.2	Evaluation results of Sauvola binarization using character recognition metrics.	39
3.3	Evaluation results of Sauvola binarization for different degradations using character recognition metrics.	40
3.4	Segmentation evaluation results	47
3.5	Evaluation results of adaptive segmentation using character recognition metrics.	48
3.6	EM based Parameter Estimation : Newspaper Segmentation	49
3.7	Optimization Framework : Newspaper Segmentation	49
4.1	Evaluation Results : Median LCC and SROCC	61
4.2	Binarization Evaluation : PSNR and F-Measure	61
4.3	Binarization using LM based approach Results : PSNR and F-Measure	66
5.1	Compression achieved using GFG based representation compared to JBIG.	75
5.2	MAP score with LSA and without LSA.	80
5.3	MAP score with PLSA	80
6.1	Algorithm: Greedy Search	99
6.2	Dataset Description	100
6.3	My caption	101
7.1	Comparison of Image features for Document Indexing	122

7.2 MAP and Avg. Comparisons for Devanagari Dataset with different pareto optimal solutions. 124

7.3 Retrieval results for Devanagari, Bangla, Telugu and English Dataset 125

7.4 Comparison of the proposed framework with state-of-the-art 125

7.5 Retrieval Results for Multi-Modal Document Image Retrieval 135