ON IMPROVING TECHNIQUES FOR ACCESSING CONTENT IN DOCUMENT IMAGE COLLECTIONS

RITU GARG

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 2016

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 greatful 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 word 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 दस्तावेजों के संग्रह में दिखाया गया है| बांग्ला और तेलुगू 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-मोडल अनुक्रमणिका योजना पाठ और सूचना-ग्राफिक्स के आधार संयोजन को सीखकर प्रस्तुत किया गया है। मूल्यांकन दिखाया गया है अंग्रेजी दस्तावेज़ छवियों पर।

Contents

List of Figures

List of Tables