Proceedings of the ICDAR 2011 Doctoral Consortium

September 18, 2011
Beijing, China

Sponsored by

Proceedings of the
ICDAR 2011 Doctoral Consortium

September 18, 2011
Beijing, China

Sponsored by

Raytheon
BBN Technologies

Doctoral Consortium Organization

TC-10 / TC-11 Organizing Committee
TC-11 Chair: Daniel Lopresti, USA
TC-10 Chair: Jean-Marc Ogier, France
Jean-Christophe Burie, France
Masakazu Iwamura, Japan
Gernot A. Fink, Germany
Dimosthenis Karatzas, Spain
Koichi Kise, Japan
Bart Lamiroy, France
Rafael Lins, Brazil
Josep Lladós, Spain

Mentors
Henry Baird, USA
Elisa Barney Smith, USA
Abdel Belaid, France
Jean-Christophe Burie, France
Gernot A. Fink, Germany
Masakazu Iwamura, Japan
Dimosthenis Karatzas, Spain
Koichi Kise, Japan
Bart Lamiroy, France
Rafael Lins, Brazil
Cheng-Lin Liu, China
Marcus Liwicki, Germany
Josep LLadós, France
Daniel Lopresti, USA
Jean-Marc Ogier, France
Faisal Shafait, Germany
Palaiahnakote Shivakumara, Singapore
Liu Wenyin, Hong Kong
Richard Zanibbi, USA

Foreword

The leaderships of IAPR TC-10 ("Graphics Recognition") and TC-11 ("Reading Systems") are pleased to organize the inaugural Doctoral Consortium for our community in conjunction with ICDAR 2011 at the Beijing Friendship Hotel. The goal of the Doctoral Consortium is to create an opportunity for Ph.D. students to test their research ideas, present their current progress and future plans, and receive constructive criticism and insights related to their future work and career perspectives. Several months before the event, a mentor who is a senior researcher active in the field was assigned to each student to provide individual feedback. At the Doctoral Consortium, students are provided with the opportunity to present an overview of their research plans during a special poster session. A Best Poster Presentation Award is to be given at the end of the day.

Through an open call for participation, a total of 21 Ph.D. students from eight different countries are engaged in the ICDAR 2011 Doctoral Consortium. They are supported by 19 mentors representing nine different countries. As demonstrated in the pages that follow, these students represent the breadth of the international research community in the field of document image analysis. The potential for their work is evident in reading the students' research summaries and perusing their already-impressive resumes.

The schedule for the Doctoral Consortium, which takes place on September 18, is as follows:

1:40 pm – 2:10 pm: Welcome and short talk: "Advice for a Successful Ph.D. Experience"
2:10 pm – 3:10 pm: Brief oral introductions to student research plans
3:10 pm – 3:25 pm: Coffee break
3:25 pm – 5:25 pm: Student poster session with discussion and feedback
5:25 pm – 5:40 pm: Concluding remarks and Best Poster Presentation Award presentation

This program book is made available to those attending the event and will also be archived on the website for ICDAR 2011, as well as the websites for IAPR TC-10 and TC-11. It should be understood that the summaries presented here are not to be considered official publications of original research results, and should not be cited as such. No copyright is asserted by the Doctoral Consortium – those rights remain with the original authors.

A word to students concerning IAPR and its TC-10 and TC-11. IAPR is the International Association for Pattern Recognition, the premier association for those involved in all aspects of pattern recognition research. As developing researchers, you are potential future leaders in the field. We encourage you to learn more about IAPR and its many activities to help support your career through its website at www.iapr.org. Much of the volunteer effort of IAPR is led through its technical committees. TC-10 is devoted to work on graphics recognition (http://www.iapr-tc10.org/), while TC-11 focuses on research relating to reading systems, including optical character recognition and handwriting recognition (http://www.iapr-tc11.org/).

The success of this Doctoral Consortium is due in large part to the efforts of our volunteer mentors who have interacted with the students for several months, and to the Ph.D. advisors of the students who have encouraged them to participate. Special thanks are due to the ICDAR 2011 organizers in Beijing for facilitating the local logistics, especially Cheng-Lin Liu who has devoted significant attention to matters connected to the Doctoral Consortium. We express our appreciation to Jeanne Steinberg at Lehigh University for her help in assembling the hardcopy program book. Finally, we wish to take this opportunity to thank Raytheon BBN Technologies for their generous financial support which has allowed the students and mentors to attend the Doctoral Consortium with no registration fee.

Daniel Lopresti
IAPR TC-11 Chair
Lehigh University, USA
August 21, 2011

Table of Contents

Document Image Classification

Olivier Augereau

LaBRI - Laboratoire Bordelais de Recherche en Informatique

Université de Bordeaux, 351 Cours de la Libération

Talence, France

Advisor : Jean-Philippe Domenger

1 Research statement

1.1 Problematic

The subject of the thesis takes place in research field of image analysis and more especially in document analysis. The goal is to explore new techniques for document recognition and classification by analyzing document image features. We want to explore new approaches that are based on appearance (layout structure or image features) and not on OCR. One particularity of this thesis is that it is link to a digitizing company. This imply that it give the benefit of using resources of the company like the millions of documents that are digitized each month. Furthermore, researches will be applied to an industrial context so it will be important to understand and to ensure appliance to industrial constraints. The global problematic of the thesis is to find a way to accelerate or automatize identification and classification of documents.

1.2 Plan

In order to find solutions for the main problematic, research will be done in this different axes :

- Document image feature and classification algorithms.
- Classification of a large and fully unknown database.
- Identification of documents with user interaction.

1.3 Progress to date

The first step of the thesis was to research techniques for image analysis i.e. document image feature extraction. Three main types of features can be set apart : visual features, structural features and textual features [1]. According to Kumar et al. [2] general visual features are color, shape and texture. However document image are quite specific so specific features must be extracted in order to analyze document zones like text, drawing, halftones, speckles, table, logo, etc. Keysers et al. [3] extract a list of nine features in order to classify image zones. Layout extraction and comparison with the cyclic polar page layout representation [4] is very interesting but do not provide very good result on my databases. Results show that sometimes similar documents have different layout and different documents have similar layout. Maybe better result will be obtain if we could teach which blocks are fixed, which blocks move and which are significant or not. Of course, visual, structural or textual approaches are not exclusive and can be combined to obtain better results.

Then I focus to document image classification. Chen and Blostein [1] offer a detailed survey about document image classification. By studying this survey, it could be pointed out that most of classification techniques are supervised. The problem with supervised techniques is that they need to know in advance the number of classes and the type of documents in order to annotate half of the database to use it as data learning. This implies that labeling time can not be accelerate by more than two. Moreover, even if techniques give excellent results with very few mistakes, company must control all the images one by one to correct the wrong labeled images. Finally, saving time will be very low.

However, classifying automatically documents is a very complex task without knowing rules that bring documents together. Instead of trying to automate classification task, our idea is to help the user by sorting document by similarity. Clustering algorithms can be used in order to explore unknown database, to group images and to find centers of clusters. Our paper accepted in ICDAR (ID 333 : Document Images Indexing with Relevance Feedback : an Application to Industrial Context) expend this idea. The use of relevance feedback enable to upgrade similarity measure by feature selection.

The next step will be to study interactions of users in order to improve identification of documents. Feedback is a good way, but another way is to provide tools to user in order to allow him to specify more precisely his query like text, table, picture, logo, layout, etc.

References

[1] N. Chen and D. Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal on Document Analysis and Recognition*, vol. 10, no. 1, pp. 1–16, 2007.

[2] M. Kumar, K. Suneera, and P. Kumar, "Content Based Image Retrieval-Extraction by Objects of User Interest," *International Journal on Computer Science and Engineering*, vol. 3, no. 3, pp. 1068–1074, 2011.

[3] D. Keysers, F. Shafait, and T. Breuel, "Document image zone classification - a simple high-performance approach," in *2nd Int. Conf. on Computer Vision Theory and Applications*, 2007, pp. 44–51.

[4] A. Gordo and E. Valveny, "A rotation invariant page layout descriptor for document classification and retrieval," in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition-Volume 00*. IEEE Computer Society, 2009, pp. 481–485.

2 Curriculum Vitæ

Olivier AUGEREAU
Bureau 106
351, cours de la Libération
33405 Talence

Phone : +33 5 40 00 24 93

E-mail : augereau.o@gmail.com
Website : http://olivier-augereau.com
26 year-old
Single
French

PhD. Student in Image Analysis

Education

2009-2012	PhD. student at LaBRI, the computer science laboratory of Bordeaux. Working in the Image and Sound team, on image analysis field. My thesis subject is about document image analysis. Expected graduation date : December 2012.
2009	Engineer Diploma from ENSEIRB a French Graduate School (Grande École) specialized in Electronics and Computer science based in Bordeaux.
2009	Master Diploma of Image, Sound and Multimedia processing from Bordeaux University.
2006-2009	3 years as engineering student at ENSEIRB.
2003-2006	3 years in Classe Préparatoire aux Grandes Écoles : intensive years of math and physics in preparation for the selective entrance examination to French engineering schools. Lycée Turgot, Limoges.
2003	Baccalaureate S. equivalent to 'A' levels in math, physics and engineering sciences. Lycée Pré de Cordy, Sarlat.

Publications

GRETSI 11	Classification d'Images de Documents avec Retour de Pertinence : Application aux Documents de Type Ressources Humaines.
ICDAR 11	Document Images Indexing with Relevance Feedback : an Application to Industrial Context.

Teaching experience

2010-2011	University Institutes for Technology of computer science of Bordeaux : – image processing (imageJ, openGL) - 24h. – web (HTML, CSS, Joomla) - 10h. – multimedia project - 8h. Bordeaux University : C2I (openoffice, microsoft office) - 16h.
2011-2012	University Institutes for Technology of computer science of Bordeaux : – event-driven programming (C#) - 24h. – web - 10h. – multimedia project - 8h. – object-oriented programming (UML, C++, Java, C#) - 22h

Work experience

2009-2012	3 years at LaBRI and Gestform (document digitizing company), research and development.
Feb 2009 - Jul 2009	6 months at INRIA (French National Institute for research in computer science), Bordeaux. Subject : design a multitouch screen and research for interaction and computer vision.
Jun 2008 - Sep 2008	4 months in laboratory of computer science and electrical engineering. Kumamoto, Japan. Subject : random number generation based on chaos theory and applications to Markov process and CDMA.

Olivier AUGEREAU
Bureau 106
351, cours de la Libération
33405 Talence

Phone : +33 5 40 00 24 93

E-mail : augereau.o@gmail.com
Website : http://olivier-augereau.com
26 year-old
Single
French

PhD. Student in Image Analysis

Languages

English	Good working knowledge. TOEFL score : 560/650
French	Mother tongue.
Japanese	Beginner (3 years). Working towards JLPT N5.
Spanish	Basic notions.

Interests

Music	Playing piano and electric guitar.
Japanese Culture	Japanese Language class at Bordeaux University. Voluntary work at "Mandora" association.
Sports	Basket, Tennis, Volley-ball.

Document Image Enhancement (Research Statement)

Su Bolan
School of Computing, National Universtiy of Singapore
subolan@comp.nus.edu.sg

August 17, 2011

1 Introduction and Motivation

More and more documents are digitalized everyday via camera, scanner and other equipment. And many digital images are taken with text information in the scene. It would be very useful to convert the characters from a image format to a textual format using optical character recognition (OCR). That information is very important for document mining, document image retrieval and so on.
However, there are many different kinds of distortion within many digital images, which affect the performance of OCR significantly and make the textual information unaccessible. So enhancing the textual information accessibility in the digital image before applying OCR on document images is very important for ensuring document-processing tasks such as document segmentation, document layout analysis, and document retrieval. The document image enhancement techniques improve the document image quality not only to enhance human perception, but also facilitate subsequent automated image processing. There are many different kinds of document enhancement techniques which handle different distortions in document images, such as document image dewarping and document image super-resolution. And the localization and extraction of the text in natural scene images are also very important for extracting the textual information in the images. During my Ph.D. study, I focus on two aspects of the document enhancement techniques: document image binarization and document image deblurring. But the other kinds of document enhancement techniques may be explored in the future too.

2 Document Image Binarization

Document image binarization is usually performed in the preprocessing stage of different document image processing related applications such as optical character recognition (OCR) and document image retrieval. It converts a gray-scale document image into a binary document image and accordingly facilitates the ensuing tasks such as document skew estimation and document layout analysis.
Though document image binarization has been studied for many years, the optimal thresholding of degraded document images is still an unsolved problem. This can be explained by the fact that the modeling of the document foreground/background is very difficult due to various types of document degradation such as uneven illumination, image contrast variation, bleed-through, and smear. The high intensity variation within both the document background and foreground caused by degradations makes it difficult to design an uniform classification method that correctly separates text and background for all kinds of degraded document images.
The recent Document Image Binarization Contest (DIBCO) held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 and Handwritten

Document Image Binarization Contest(H-DIBCO) organized in conjunction with International Conference on Frontiers in Handwriting Recognition (ICFHR) 2010 particularly address this issue by creating challenging benchmarking datasets and evaluating the recent advances in document image binarization. The DIBCO contest received entries of 43 algorithms from 35 international research groups, and H-DIBCO contest received 17 algorithms, which partially reflect the current efforts on this task as well as the common understanding that further efforts are required for better document image binarization solutions.

3 Blurred Document Image Restoration

Blurring is one of the most common artifacts in digital photography. There are two main kinds of blurring: one is motion blur, which is caused by the relative motion between the camera and object during image capture; the other is defocus blur, which is due to the incorrect focal length setting when taking photos. Blurring induces the degradation of image quality, especially for document images where the text information is easily lost due to blur.
Many techniques have been proposed to address this problem that can be broadly classified into two categories, namely, blind deconvolution and non-blind deconvolution. In non-blind deconvolution, the PSF is assumed to be known, only the clear original image I needs to be estimated. For the blind deconvolution, only the blurred image is known, the PSF needs to be estimated as well as the original image. To find a unique and meaningful solution, some constraints and prior knowledge have to be incorporated in the deconvolution process. For the document image domain, the purpose of deblurring is more specific, which is not only to improve the image visual quality, but also the OCR precision of the text within the image. And we can try to extract some domain knowledge such as the bimodal distribution of document image to help in estimating the PSF and restoring the clear image.

4 Current Progress and Future Plan

Currently we have developed two document image binarization techniques for severely degraded document images. The proposed techniques have been tested on the DIBCO 2009 and H-DIBCO 2010 datasets. Experimental results have shown the superior performance of our methods. And we have won the best performance in the H-DIBCO 2010 among 17 algorithms around the world. The following is a list of the work I have done during my Ph.D. study.

- **Blurred Image Region Detection and Classification:**(ACM MM 2011)
  We proposed a simple and effective automatic image blurred region detection and classification technique that first detects blurred image regions by examining singular value information for each image pixel. The blur types (i.e. motion blur or defocus blur) are then determined based on certain alpha channel constraints that requires neither image deblurring nor blur kernel estimation. The proposed technique can be used in many different multimedia analysis applications such as image segmentation, depth estimation and information retrieval. And we can use it as a preprocessing step for document image deblurring.

- **Document Image Binarization Using Background Estimation:**(IJDAR 2010 December)
  We develop a binarization technique that makes use of the document background surface and the text stroke edge information. The text documents usually have a document background of uniform color and texture, but the document text within it has a different intensity level compared with the surrounding background. It first estimates a document background surface through an iterative polynomial smoothing procedure. The variation of the image contrast resulted from document degradations such as shading and smear is then compensated for by using the estimated document

background surface. The text stroke edges are further detected based on the local image variation within the compensated document image. After that, the document text is extracted based on the local threshold that is estimated from the detected text stroke edge pixels. At the end, a series of post-processing operations are performed to further correct the error that is introduced during the thresholding by using the detected stroke edge pixels.

- **Document Image Binarization Using Local Maximum and Minimum:**(DAS 2010)
  Another binarization technique we developed makes use of the local image contrast evaluated by local maxima and minima. Such image contrast is more capable of detecting the image pixels lying around the text stroke edge from historical documents that often suffer from different types of document degradation. Given a historical document image, the proposed technique first determines a contrast image based on the local maximum and minimum. The high contrast image pixels around the text stroke boundary are then detected through the global thresholding of the determined contrast image. Lastly, the historical document image is binarized based on the local thresholds that are estimated from the detected high contrast image pixels.

- **Improving Document Image Binarization Techniques:**
  Many practical binarization techniques have been developed and applied successfully on commercial document analysis systems. However, the current state-of-the-art methods may favor certain kinds of document images, but fail to produce good binarization results for other kinds of badly degraded document images. So we believe that it is better to adaptively increase the performance of existing document image binarization methods by employing domain knowledge and image statistics, compared with inventing new thresholding methods for document image binarization. Based on the results of existing document binarization methods, we can view the binarization as a learning problem, and make use of the initial classification results produced by existing methods. Two frameworks have been proposed:

  - **A Self-Learning Document Binarization Framework.**(ICPR 2010)
    We use confidently classified pixels that are far away from the threshold surface created by existing thresholding method to train a classifier to re-label those remaining pixels.
  - **Combination of Document Binarization Methods.**(ICDAR 2011)
    We use those pixels that labeled the same by different methods (which is usually correctly classified) to iteratively re-classify the remaining pixels.

In my remaining Ph.D. study, I will use about 2 months to complete my work on document image binarization techniques where I can try to propose a better framework for combining existing binarization techniques and improving the performance of reported methods. And I will use 6 to 8 months to develop some techniques for document image blur detection and restoration. Lastly, I want to use 2 to 4 months to explore other document enhancement techniques, apply methods to other domains, such as musical score documents, documents with complex background and figures, localization and extraction of text in natural scene images.

Curriculum Vitae

Bolan Su Ph.D. Candidate(Expected to graduate in August 2012)
School of Computing *Phone:* (65) 6516 2784
National University of Singapore *E-mail:* subolan@comp.nus.edu.sg
13 Computing Drive,Singapore 117417 *Homepage:* http://www.comp.nus.edu.sg/∼subolan

Research Interesting

- Information Retrieval, Document Image Restoration and Enhancement, Document Image Binarization, Image Processing, Text Recognition, Computer Vision.

Education

- **National University of Singapore** Singapore
  Ph.D. Candiate, Computer Science, School of Computing *August 2008 - August 2012(expected)*
  - Dissertation Topic: Document Image Enhancement
  - Supervisor: Prof. Tan Chew Lim; Dr. Lu Shijian
  - Cumulative Average Point (CAP): 4.5/5.0
- **Fudan University** Shanghai,China
  Bachelor, Computer Science *September 2004 - July 2008*
  - Dissertation Topic: Research on Image Based Rendering Algorithms for 3D Scene Construction
  - Supervisor: Dr. Jin Cheng
  - Grade Point Average (GPA): 3.5/4.0, 13th of 140 students

Award

- First Prize at Handwritten Document Binarization contest(H-DIBCO 2010) organized in conjunction with ICFHR 2010, 2010 among 17 submitted algorithms for the contest

Academic Experience

- **National University of Singapore** Singapore
  Graduate Student in School of Computing *August 2008 - present*
  - Includes current Ph.D. research, Ph.D. level coursework and research projects.
- **Institute for Infocomm Research** Singapore
  Postgraduate in Computer Vision & Image Understanding Department *May 2009 - present*
  - Supervision by Dr. Lu Shijian on my Ph.D. research topics.
- **Indian Statistical Institute** Kolkata,India
  Visiting Ph.D. in Computer Vision & Pattern Recoginition Unit *March 18 - March 27, 2011*
  - Work under Prof. Umapada Pal on historical document image analysis during this short term visit.
  - Discussion with Prof. Umapada Pal for the future study.
- **Fudan University** Shanghai, China
  Undergraduate Researcher in Dept. of Computer Science and Technology *July 2006 - July 2008*

Publications

- **Bolan Su**, Shijian Lu, Chew Lim Tan: Blurred Image Region Detection and Classification. In Proceedings of the eighteen ACM International Conference on Multimedia, 2011 (to appear).
- **Bolan Su**, Shijian Lu, Chew Lim Tan: Combination of Document Image Binarization Techniques. International Conference on Document Analysis and Recognition(ICDAR), 2011 (to appear). [**Oral**]
- P. Shivakumara, S. Bhowmick, **Bolan Su**, Chew Lim Tan, U. Pal: A New Gradient based Character Segmentation Method for Video Text Recognition. International Conference on Document Analysis and Recognition(ICDAR), 2011 (to appear).
- D. Rajendran, P. Shivakumara, **Bolan Su**, Shijian Lu, Chew Lim Tan: A New Fourier-Moments based Video Word and Character Extraction Method for Recognition. International Conference on Document Analysis and Recognition(ICDAR), 2011 (to appear).
- Trung Quy Phan, P. Shivakumara, **Bolan Su**, Chew Lim Tan: A Gradient Vector Flow-Based Method for Video Character Segmentation. International Conference on Document Analysis and Recognition(ICDAR), 2011 (to appear).
- Shijian Lu, **Bolan Su**, Chew Lim Tan. Document Image Binarization Using Background Estimation and Stroke Edges. International Journal on Document Analysis and Recognition. 2010, vol.13:303-314.
- **Bolan Su**, Shijian Lu, Chew Lim Tan. A Self-training Learning Document Binarization Framework. International Conference on Pattern Recognition, Istanbul, Turkey, 23-26 August 2010, 3187-3190.
- **Bolan Su**, Shijian Lu, Chew Lim Tan. Binarization of Historical Document Images Using the Local Maximum and Minimum. International Workshop on Document Analysis Systems, Boston, MA, USA, 9-11 June 2010, 159-166.[**Full paper, Oral**]
- Hui Yu, **Bolan Su**, Hong Lu, Xiangyang Xue. News Video Retrieval by Learning Multimodal Semantic Information. International Conference on Visual Information Systems, 28-29 June 2007, SHANGHAI, CHINA.

Professional Services

- **Pattern Recognition and Machine Intelligence Association** Singapore
  Student Helper *April 2010 - present*
  - Help to manage the premia website and event preparation.
- Reviewer of 22th International Conference on Tools with Artificial Intelligence(ICTAI) 2010.
- Reviewer of 11th International Conference on Document Analysis and Recognition(ICDAR) 2011.
- Reviewer of First Asian Conference on Pattern Recognition(ACPR) 2011.

Automatic understanding of sketch maps

by

Klaus Broelemann
Department of Mathematics and Computer Science
University of Muenster, Germany

Advisor:

Xiaoyi Jiang
Department of Mathematics and Computer Science
University of Muenster, Germany

1 Introduction

During the last years, Geographic Information Systems (GIS) became a widely-used technology in daily life. While the ability and complexity of GI systems are continuously increasing, there is still an absence of easy-to-use interaction methods. According to Schlaisich and Egenhofer [1] hand-drawn sketch maps can be an intuitive way to interact with GIS.

One way of using sketch maps is for volunteer geographic information (VGI) systems, which allow users to annotate, add and modify content of maps. Using sketch maps as input would enable users to provide their knowledge in a natural way to automatic systems and to share it with other users.

By doing so, a system has to understand, process and align sketch maps. The initial step for that is the understanding of sketch maps. That means to locate and recognize objects of the sketch map. Subsequent processing works with these objects and their spatial and topological relations.

1.1 Goal of my work

For my Ph.D. research I propose to develop algorithms for automatic offline semantic recognition and integration of objects in images of hand-drawn sketch maps. The goal of the recognition is to transform a low-level pixel representation into a high-level semantically enabled object representation. In this context, objects are elements of sketch maps that have a meaning, like the representations of buildings, trees, lakes and streets. Single lines normally do not have a meaning and, hence, are not considered as objects in this context.

My research is connected to the Sketchmapia project[1] and can be seen as an initial step of automatic sketch map processing in this project.

[1] See `http://sketchmapia.de/` for further information.

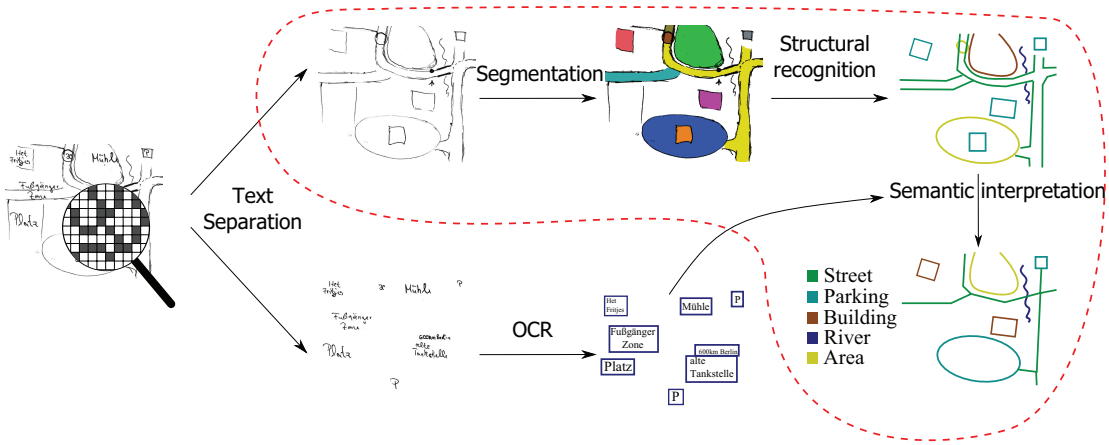


Figure 1: Proposed steps for an automatic sketch understanding system: text separation, optical character recognition (ocr), segmentation, structural recognition and semantic interpretation. The focus of my work is shown inside the dotted line.

2 Research plans

I propose to divide the sketch understanding into the following substeps: text separation, optical character recognition (ocr), segmentation, structural recognition and semantic interpretation. The interaction between these substeps can be seen in Fig. 1. In my work, I will focus on Segmentation, Structural Recognition and Semantic Interpretation of objects.

2.1 Work packages

Segmentation The aim of this work package is to find algorithms to segment a sketch map image into regions for single objects.

Structural recognition The goal for this work package is to recognize the objects structurally. This means to classify the objects shapes, but also to compute attributes for the shape like curvature or compactness. A third task for this part is to recognize the spatial and topological relations between objects.

Semantic interpretation In contrast to the previous work package, the goal of the semantic interpretation is to find the meaning of objects and not only their shape. Since the same shape can have different meanings, there is no direct matching between shapes and semantics. Thus other information like the relations between objects and the written labels has to be used for this step. I am planning to build an ontology of sketch map objects for this step. Though ocr is not focus of my work, I will investigate how recognized labels can be used for semantic interpretation.

2.2 Current state of research

For the first work package I developed a region based segmentation algorithm. This method will be presented at the GREC workshop. Furthermore I developed a method for street graph detection [2]. This method covers all three work packages, but is restricted to the recognition of streets.

My current work is on hierarchical segmentation and interpretation for combining segmentation and structural recognition (see Ahuja and Todorovic [3]). Furthermore, the gained structural information can be used for the subsequent semantic interpretation.

References

[1] I. Schlaisich and M. Egenhofer, "Multimodal spatial querying: What people sketch and talk about," in *1st International Conference on Universal Access in Human-Computer Interaction*, 2001, pp. 732–736.

[2] K. Broelemann, X. Jiang, and A. Schwering, "Automatic street graph construction in sketch maps," in *Proceedings of the 8th Workshop on Graph-based Representations in Pattern Recognition*, Mnster, Germany, 2011, pp. 275–284.

[3] N. Ahuja and S. Todorovic, "From region based image representation to object discovery and recognition," in *Structural, Syntactic, and Statistical Pattern Recognition*, ser. Lecture Notes in Computer Science, E. Hancock, R. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, Eds., vol. 6218. Springer Berlin / Heidelberg, 2010, pp. 1–19.

Klaus Broelemann

Personal Details

Name	Klaus Peter Broelemann
Date of birth	21.11.1980
Place of Birth	Rheine, Germany
Nationality	German
Family status	single

Current Status

since February 2010 — **PhD student**, *in Computer Science*, Topic: Automatic understanding of sketch maps, Part of: International Research Training Group (IRTG) on Semantic Integration of Geospatial Information.

January 2013 — **Expected Graduation**.

Education

2001–2010 — **Diploma**, *Institute for computer science*, Westfälische Wilhelms Universität Münster.
Thesis: Registration of city maps on mobile devices

Experience

WS 2004/05 — **Student assistant**, *tutor for "Basic Principles of theoretic computer science"*.

WS 2004/05 — **Student assistant**, *tutor for "Artificial Intelligence"*.

2005-2006 — **Student assistant**, *European Institute of Molecular Science*.

2000-2008 — **Corrector**, *Mathematics Olympiad in the district of Steinfurt and in NRW*.

Publications

Klaus Broelemann, Xiaoyi Jiang, and Angela Schwering. Automatic street graph construction in sketch maps. In *Proceedings of the 8th Workshop on Graph-based Representations in Pattern Recognition*, Münster, Germany, 2011.

Xiaoyi Jiang, Klaus Broelemann, Steffen Wachenfeld, and Antonio Krüger. Graph-based markerless registration of city maps using geometric hashing. *Computer Vision and Image Understanding*, 115(7):1032 – 1043, 2011.

Steffen Wachenfeld, Klaus Broelemann, Xiaoyi Jiang, and Antonio Krüger. Graph-based registration of partial images of city maps using geometric hashing. In *Proceedings of the 7th Workshop on Graph-based Representations in Pattern Recognition*, pages 92–101, Venice, Italy, 2009.

Institute for Computer Science
Einsteinstr. 62 – 48149 Münster
✉ *broele@uni-muenster.de*
• *cvpr.uni-muenster.de/organisation/broelemann.html*

Generic Layout Analysis of Diverse Collection of Documents

Syed Saqib Bukhari
Image Understanding and Pattern Recognition (IUPR) Research
Technical University of Kaiserslautern, Germany
bukhari@informatik.uni-kl.de, bukhari@iupr.com
Advisor: Prof. Dr. Thomas M. Breuel

1 Introduction

Diverse Collection of Document Images: document image processing is the subfield of digital image processing. It mainly deals with the transformation of digitized documents into electronic/symbolic form for storage, transmission, reuse, and modification. Traditionally, scanners are used for document digitization, which produce planer surface document images as shown in Figure 1(a) and 1(b). Digital cameras are also being used for document digitization nowadays, which usually produce perspective and geometric distortions in document images as shown in Figure 1(c). English script (Figure 1(a) and Figure 1(b)) is one of the simplest script in the world with respect to document image processing [Nag00] as compared to complex scripts [KKJ07] like Arabic, Persian, African, Urdu (Figure 1(c)), and Indic scrips (Kannada, Telugu, etc.). Other than printed-text documents (Figure 1(a), Figure 1(b), and Figure 1(c)), there exits a huge amount of handwritten documents especially in the form of historical documents in libraries all over the world. A sample historical document image is shown in Figure1(d). A corpus of document images as shown in Figure 1 is referred to as *diverse collection of document images*.

(a) Scanned, Binary, Printed-Text Latin Script, Manhattan Layout

(b) Scanned, Color, Printed-Text Latin Script, Non-Manhattan Layout

(c) Camera-Captured, Grayscale, Printed-Text Urdu Script, Warped Document

(d) Scanned, Grayscale, Handwritten/Historical Arabic Script, Irregular Layout

Figure 1. Diverse collection of document images.

Geometric Layout Analysis: the first step of document image processing in the pipeline of knowledge acquisition from a document image is *geometric layout analysis*–decomposition of document image its homogeneous regions and determination of their reading order. For diverse collection of document images, which are composed of rectangular as well as non-rectangular/irregular blocks, *text-line* is the dominant geometrical layout structure. Therefore, text-line extraction is considered as one of the most important geometric layout analysis steps. Over the

last four decades, several text-line extraction methods have been proposed. Comprehensive overview of the state-of-the-art page segmentation (or text-line extraction) methods has been provided in [CCMM98, Nag00]. Most of these algorithms perform well for scanned, binarized, cleaned document images with simple script [SKB08], but fail on complex document images such as warped camera-captured document images [BSB10], complex scripts [KKJ07], and handwritten/historical document images [LZDJ08, LSZT07] because of their specific challenging problems. Text-line extraction for complex document images is an open challenging field. Every year, several researcher proposed different text-line extraction methods for solving specific problems in each of these categories of complex document images. Detailed overview of the state-of-the-art text-line extraction methods for warped camera-captured document images has been provided in [BSB10] and for handwritten and historical document images in [LSZT07], and some of the specialized text-line extraction approaches for complex script document images can be found in [KKJ07, SHKB06].

Generic Text-Line Extraction Method: there is no universal or generic text-line finding method that can be robustly applied to a diverse collection of document images. Many researchers [NJ07, LSZT07, KB06] highlighted a problem that, the universal/generic text-line extraction method is an elusive goal so far and still beyond the reach of the state-of-the-art in the field. Even though, a generic text-line finding method can solve variety of document image processing problems/tasks such as:

- it can be applied equally on diverse collection of document images and overcome the requirement of specific text-line extraction methods for different categorize of document images,
- it can improve the performance of existing OCR softwares for complex printed-text document layouts
- it can help in enhancing the quality of camera-captured document images by removing geometric and perspective distortions through monocular dewarping methods,
- it can solve layout analysis problem for complex scripts (like Urdu, Telugu, Kannada, etc.) document images,
- it can upgrade/promote the document image processing pipeline of handwritten/historical document images to page level, which traditionally focus on the recognition at line, word, or character levels because of complex (irregular) page layouts.

2 Contributions of this Thesis

The main contributions that are presented in this PhD thesis are:

- **A Generic Text-Line Extraction Method:**
  - presented a novel text-line extraction method using two standard image processing techniques: filter bank smoothing and ridge detection (Bukhari CAIP'09 [15], Bukhari ICDAR'11 [1]) [1]
  - it can be equally applied on a large variety of document images, which are composed of scanned or camera-captured images, different types of writing styles (printed-text or handwritten), different types of intensity values (binary or grayscale), or different scripts (Latin, Chinese, Arabic, etc.)
- **Text/Non-Text Segmentation**
  - introduced a novel discriminative learning based text and non-text segmentation method (Bukhari DAS'10 [8])
  - improved the Leptonica's multiresolution morphology based page segmentation method (Bukhari DRR-SPIE'11 [7])
- **Text-Line based Preprocessing of Camera-Captured Document Images**
  - **Binarization:** proposed a novel foreground (text-line) guided local adaptive thresholding method for degraded camera-captured document images (Bukhari CBDAR'09 [14], Bukhari JUCS'09 [10])

[1] All references starting with "Bukhari" can be found in the "Publications" section of the CV.

- **Document Cleanup (Page Frame Detection):** developed a text-line based page frame detection method for noise cleanup in camera-captured document images (Bukhari CBDAR'11 [4])
- **Monocular Dewarping:** presented a text-line based monocular dewarping method for rectifying camera-captured document images (Bukhari CBDAR'09 [13])
- **Performance Evaluation Methodology for Dewarping Methods:** introduced an image based performance evaluation method for page dewarping algorithms using SIFT (Bukhari CBDAR'11 [3])

- **High Performance Layout Analysis of Arabic and Urdu Document Images**
  - developed a layout analysis system for segmenting text and non-text elements and extracting text-lines in reading order from scanned Arabic script document images written in different languages (Arabic, Urdu, Persian, etc.) and different styles (Naskh, Nastaliq, etc.) (Bukhari ICDAR'11 [2], Bukhari Springer'11 [6])

- **New Active Contour (Snake) Models for Document Image Segmentation**
  This thesis also contributes in presenting new active contour (snake) based image segmentation models, which can be adapted for solving various image segmentation problems where a traditional active contour model does not work. The adaptation of these new active contour models for document image segmentation are also presented.
  - **Baby-Snake Model:** introduced a novel "baby-snakes" model (multiple open-curve snakes that initializes automatically on discrete set of points and deform in targeted direction to achieve segmentation) and demonstrated its application for curled text-line segmentation from camera-captured document images (Bukhari DAS'08 [18], Bukhari ICDAR'09 [12])
  - **Coupled Snakelets Model:** introduced a novel "Snakelets" model (an extension of baby-snake model that introduces an automatic initialization of multiple wighted-coupled pairs of snakes on discrete points and their deformation in evolving fashion) and adapted it for curled text-lines segmentation (Bukhari ICDAR'09 [11])

References

[BSB10] S. S. Bukhari, F. Shafait, and T. M. Breuel. Performance evaluation of curled textline segmentation algorithms on CBDAR 2007 dewarping contest dataset. In *Int. Conf. on Image Processing, 2010 17th*, pages 2161 –2164, Cairo, Egypt, sept. 2010.

[CCMM98] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena. Layout analysis techniques for document image understanding: a review. In *available from http://citeseer.nj.nec.com/, IRST, Trento, Italy, Tech. Rep. 9703-09*, 1998.

[KB06] D. J. Kennard and W. A. Barrett. Separating lines of text in free-form handwritten historical documents. In *2nd Int. Conf. on Document Image Analysis for Libraries.*, pages 12 – 23, Los Alamitos, CA, USA, april 2006.

[KKJ07] K. S. Kumar, S. Kumar, and C. Jawahar. On segmentation of documents in complex scripts. In *9th Int. Conf. on Document Analysis and Recognition*, pages 1243–1247, Washington, DC, USA, 2007.

[LSZT07] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *Int. Journal on Document Analysis and Recognition*, 9:123–138, 2007.

[LZDJ08] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger. Script-independent text line segmentation in freestyle handwritten documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1313 –1329, aug. 2008.

[Nag00] G. Nagy. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38 –62, jan 2000.

[NJ07] A. M. Namboodiri and A. Jain. Document structure and layout analysis. pages 29–48, London, UK, 2007. Springer-Verlag.

[SHKB06] F. Shafait, A. U. Hasan, D. Keysers, and T. M. Breuel. Layout analysis of Urdu document images. In *IEEE Int. Multitopic Conference, INMIC '06*, pages 293–298, Islamabad, Pakistan, 2006.

[SKB08] F. Shafait, D. Keysers, and T. M. Breuel. Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):941 –954, june 2008.

Curriculum Vitae: Syed Saqib Bukhari

Image Understanding and Pattern Recognition (IUPR) Research
Technical University of Kaiserslautern, Germany
E-mail: bukhari@iupr.com, *Web:* sites.google.com/a/iupr.com/bukhari

Research Interests

- Computer Vision, Image Analysis, Pattern Recognition, and their Applications
- Efficient and Reliable Algorithms for the Layout Analysis of Document Images

Education

PhD in Computer Science: Technical University of Kaiserslautern, Germany, April 2008 - March 2012 (expected graduation date). Thesis Topic: *Generic Layout Analysis of Diverse Collection of Documents.* Advisor: *Prof. Dr. Thomas M. Breuel*

Masters of Engineering in Computer Systems: NED University of Engineering and Technology, Karachi, Pakistan, Jan 2004 - July 2006. Result: *4.0 GPA (excellent)*

Bachelor of Engineering in Computer Systems: NED University of Engineering and Technology, Karachi, Pakistan, Jan 1999 - Feb 2003. Result: *93% (5th Position out of 76 Students)*

Publications

16 publications in peer reviewed conferences, one in journal, and one book chapter.

Projects

DECAPOD [2010 - to-date]: 3D Capture, Dewarping, and Archival Conversion of Books and other Historical Objects (*http://sites.google.com/site/decapodproject/*)

Workshops

Selected for the "International Computer Vision Summer School, ICVSS'09" and ICVSS'11", Sicily, Italy.

Honors and Awards

- DAAD 'German Academic Exchange Service' PhD Scholarship, 2007- 2011
- Outstanding Performance in Masters of Engineering in Computer Systems
- Position in B.E. Computer Systems Engineering : 5/76

Academic and Professional Experience

Image Understanding and Pattern Recognition (IUPR) Research (Prof. Dr. Thomas M. Breuel), Technical University of Kaiserslautern, Kaiserslautern, Germany. (Researcher Scholar: October 2009 - to-date).

German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany. (Researcher Scholar: April 2007 - September 2009).

Fraunhofer-Gesellschaft, Kaiserslautern, Germany. (Research Assistant: August 2008 - September 2009).

NED University of Engineering and Technology, Karachi, Pakistan. (Teaching: Feb 2003 - Jan 2007)

Reference

Prof. Dr. Thomas M. Breuel
Image Understanding and Pattern Recognition
Technical University of Kaiserslautern, Germany, Email: tmb@iupr.com, www.iupr.com

Prof. Dr. Andreas Dengel
German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern
Email: Andreas.Dengel@dfki.de, www.dfki.de

Publications

1. S. S. Bukhari, F. Shafait, T. M. Breuel, "Text-Line Extraction using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters", 11th Int. Conf. on Document Analysis and Recognition, ICDAR'11, China, 2011.
2. S. S. Bukhari, F. Shafait, T. M. Breuel, "High Performance Layout Analysis of Arabic and Urdu Document Images", 11th Int. Conf. on Document Analysis and Recognition, ICDAR'11, China, 2011.
3. S. S. Bukhari, F. Shafait, T. M. Breuel, "An Image Based Performance Evaluation Method for Page Dewarping Algorithms using SIFT Features", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
4. S. S. Bukhari, F. Shafait, T. M. Breuel, "Border Noise Removal of Camera-Captured Document Images using Page Frame Detection", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
5. S. S. Bukhari, F. Shafait, T. M. Breuel, "The IUPR Dataset of Camera-Captured Document Images", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
6. S. S. Bukhari, F. Shafait, T. M. Breuel, "Layout Analysis of Arabic Script Documents", in Guide to OCR for Arabic Scripts, Springer-Verlag, 2011.
7. S. S. Bukhari, F. Shafait, T. M. Breuel, "Improved Document Image Segmentation Algorithm using Multiresolution Morphology", Document Recognition and Retrieval XVIII, SPIE 2011, USA, 2011.
8. S. S. Bukhari, F. Shafait, T. M. Breuel, "Document Image Segmentation using Discriminative Learning over Connected Components", 9th IAPR Workshop on Document Analysis Systems, DAS'10, USA, 2010.
9. S. S. Bukhari, F. Shafait, T. M. Breuel, "Performance Evaluation of Curled Textlines Segmentation Algorithms on CBDAR 2007 Dewarping Contest Dataset", Int. Conf. on Image Processing, ICIP'10, Hong kong, 2010.
10. S. S. Bukhari, F. Shafait, T. M. Breuel, "Adaptive Binarization of Unconstrained Hand-Held Camera-Captured Document Images", Journal of Universal Computer Science, 2009.
11. S. S. Bukhari, F. Shafait, T. M. Breuel, "Coupled Snakelet Model for Curled Textline Segmentation of Camera-Captured Document Images", 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09, Spain, 2009.
12. S. S. Bukhari, F. Shafait, T. M. Breuel, "Script-Independent Handwritten Textlines Segmentation using Active Contours", 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09, Spain, 2009.
13. S. S. Bukhari, F. Shafait, T. M. Breuel, "Dewarping of Camera-Captured Document Images", 3rd Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'09, Spain, 2009.
14. S. S. Bukhari, F. Shafait, T. M. Breuel, "Foreground-Background Regions Guided Binarization of Camera-Captured Document Images", 3rd Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'09, Spain, 2009.
15. S. S. Bukhari, F. Shafait, T. M. Breuel, "Ridges based Curled Textline Region Detection from Grayscale Camera-Captured Document Images", 13th Int. Conf. on Computer Analysis of Images and Patterns, CAIP'09, Germany, 2009.
16. S. S. Bukhari, F. Shafait, T. M. Breuel, "Curled Textline Information Extraction from Grayscale Camera-Captured Document Images", Int. Conf. on Image Processing, ICIP'09, Egypt, 2009.
17. S. F. Rashid, S. S. Bukhari, F. Shafait, T. M. Breuel, "A Discriminative Learning Approach for Orientation Detection of Urdu Document Images", 13th IEEE Int. Multitopic Conf., INMIC'09, Pakistan, 2009
18. S. S. Bukhari, F. Shafait, T. M. Breuel, "Segmentation of Curled Text Lines using Active Contours", 8th IAPR Workshop on Document Analysis Systems, DAS'08, Japan, 2008.

Effects of Artificial Sample Generation Models for On-line Handwritten Japanese Character Recognition

Bin Chen

Department of Computer and Information Sciences

Tokyo University of Agriculture and Technology

Tokyo, Japan

Current research

My research topic is online handwritten Japanese character recognition. To improve the generalization performance of classifier, I constructed six models of artificial sample generation to produce artificial samples for training classifiers and have evaluated their performance on the TUAT databases Nayayosi and Kuchibue. Though meaningful results have been obtained, there are some unsolved problems to consider for the future research.

This work is based on a theory, the more learning patterns employed for training pattern recognition systems, the higher recognition rate is obtained. Especially for languages of a large character set, like Japanese and Chinese, etc. Therefore, for researching on on-line handwritten Japanese character recognition, I had constructed six linear distortion models and combine them with a nonlinear distortion model to generate a large amount of artificial patterns. I also evaluated distortion models by the performance of classifier trained by artificial pattern.

I try to change the combination approach of distortion model, so obtain two kinds of combined method by change the order of distortion models in process. I also evaluated distortion models by the performance of classifier trained by artificial pattern.

I sort the original patterns by recognition score, then employ the part of original patterns to generate the pattern by the distortion models. I selected the highest and the lowest pattern with the same percentage, obtain a pair of comparison result. In my experiment, I choose the different percentages including 5%,10%,15%,20%,…100%. From my experiment, the accuracy of using the highest pattern part is better when the percentage is less than about 20%. Moreover, lowest pattern part get the better result when the percentage is more than about 20%.

Problem

As the description above, we evaluated unique models and current provide models for on Japanese character recognition classifier. And present some interesting results already.

But the distortions which are proposed are quite simple and remain quite classic. Even if the

proposed approach permits to improve the recognition process, the proposed distortions are not proved to be representative of on-line distortions. The improvement of the recognition process may be due to the simple fact of the increasing number of training samples.

Research plan

First, in my past experiment the class number in training data is different with test data. I had started to combine these two data sets. I take part of this combined data set for training and the rest for test.

Second, in order to know more effect of these models, I plan to finish an extend experiment use all the artificial patterns distorted from LDM and NLDM, It might take 15 days to get the result. It takes more time in training process, when time is saving in recognizing process, further a higher accuracy is obtained.

Third, I receive an article published in IGS2011. It applied a model named Sigma-Lognormal Model. It calculates writing speed of one stroke, analyzed it`s profile after Fourier Transformation. During this process, the author obtains several static features. The article is base on biology theory. Nearly, optical transmit theory, by this theory I can control distortion of stroke in frequency domain, instead of time domain. Therefore, it distort more easily and clearly.

This paper is based on symbols with just one stroke, so we need to find out a way to apply this model on Japanese characters. As we know, there are more than one stroke for most of Kanji, even for hiragana and katakana in Japanese。

This work may help me to obtain a generative model base on the structural characteristics of characters. Taking into account the hierarchical structures of Chinese/Japanese characters composed of radicals and strokes, we can characterize the variations and relations of radicals and strokes using probabilistic models such as Gaussian distributions. From this kind of probabilistic models, artificial samples of natural shapes with variable distortion degree can be generated.

(2) CV: include your expected graduation date (1 page).

Curriculum Vitae

Education

- ✧ 2010.4~ Ph.D, Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology (graduate at April,2013)
- ✧ 2008.4~2010.4 Master Course, Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
- ✧ 2002.9~2006.6 Undergraduate, Dept. of Information Science and Electronic Engineering , Faculty of Information Technology, Zhejiang University City College
- ✧ 1999.9~2002.6 Middle School, Hangzhou No.14 Middle School

Research

- ✧ Research field: On-line Handwritten Japanese Character Recognition
- ✧ [1] Bin Chen, Bilan Zhu and Masaki Nakagawa，Effects of a Large Amount of Artificial Patterns for On-line Handwritten Japanese Character Recognition（CJKPR201, Nov, 4th, 2010）
- ✧ [2] Bin Chen, Bilan Zhu and Masaki Nakagawa，Effects of Generating a Large Amount of Artificial Patterns for On-line Handwritten Japanese Character Recognition（ICDAR2011, Sep, 18, 2011）

Work Experience

- ✧ None

Exploiting Metadata in Off-line Handwritten Documents: Modeling and Applications

Jin Chen
Lehigh University
Bethlehem, PA, United States
{jic207@cse.lehigh.edu}
Advisor: Daniel Lopresti

I. Introduction

Document Image Analysis (DIA) is the subfield of digital image processing that aims at converting document images to a symbolic form for modification, storage, retrieval, reuse, and transmission [1]. One primary task in this area is to translate an input document image into a character transcription in ASCII or Unicode. However, the bulk of information that a document corpus conveys goes beyond such a transcription. For example in handwritten documents, writer idiosyncrasies are an example of document metadata that can be exploited for many applications, including handwriting recognition and document mining. Other examples include the date of creation, the language of the script, the topics involved, etc. In the discussion, I define document metadata as a class of information that is beyond the symbolic transcription of a handwritten page. I hypothesize that by exploiting such metadata information, we are able to restore the original relationships between documents, build new relationships from them, and facilitate problem solving tasks. Such document metadata include the physical and logical structure of the tables, writer idiosyncrasies, and perhaps the specifications of paper sheets people write on.

II. 2-D Arrangement Metadata

The two-dimensional arrangement of text cells conveys more critical information than the symbolic form itself. For example, for each table cell, its relationship with the *row head* and

column head build a logical relationship for these three table components which is valuable for knowledge extraction. In this thesis, I address table understanding in noisy handwritten documents by considering the involved two sub-problems separately: table detection and table recognition. So far, by evaluating an inside-space based correlation method on handwritten document, I have acquired an area-ratio based *precision* of 80% and a *recall* of 85% [2].

III. Handwriting Metadata

Handwritten annotations are valuable information since they differ from the remaining text on the temporal manner and more importantly they usually indicate attitudes, opinions, comments, and questions of the annotator. In contrast to traditional authorship analysis where sufficient data is usually assumed for extensive classifier training, in such a scenario we would always face severe data constraints. Then the research problem becomes: how can we identify authors given such limited training data? By perturbing real handwriting guided by a series of user studies, I have observed that adding perturbed handwriting for writer identification is as effective as having three pages of real handwritten documents (each page has about 25 text lines) [3]. Considering each page has about 25 text lines, we consider it a positive sign of our idea. Future work might address the possibility of simulating handwriting variations directly in the classification space.

IV. Page Metadata

Within a large collection of unstructured documents, it is common that several of them are closely related by physical materials, in addition to the contents. Such metadata are critical in restoring original relationships within the large collection. Currently, we have been collecting handwritten notebooks from Lehigh students. All these notebooks are filled with students' course notes prior to our data collection announcement. During the data collection, we strive for collecting spontaneous handwriting and least curation afterwards. This is important since we attempt to work on real-life handwritten documents which have various kinds of artifacts and noise comparing to those published datasets.

So far, by modeling ruling lines using a multi-line regression model, I have achieved quite good performance on capturing ruling lines in handwritten documents. In addition, our new algorithm has outperformed a previously published paper on similar task using the public dataset *Germana* [4]. As an ongoing work, I am evaluating our algorithm on the Lehigh notebook dataset.

V. About Myself

I started my pursuit in document analysis related work at Lehigh University since Fall 2007. Over the years, I have been working on problems from biometric security [5], [6], [7], [8], to handwriting recognition [9], to metadata exploitation [10], [11]. Thanks to my advisor Professor Daniel Lopresti being tremendously supportive, I attended almost every important conference for the past years and built great connections with other researchers in this area. Now, I have become a fifth-year Ph.D candidacy at Lehigh, and am aiming for a dissertation defense by April 2012.

References

[1] G. Nagy, "Twenty years of document image analysis in pami," *IEEE Transactions on Pattern Analysis and Machine Inteligence*, vol. 22, no. 1, pp. 36–62, 2000.

[2] J. Chen and D. Lopresti, "Table detection in noisy off-line handwritten documents," in *Proceedings of the 2011 11th International Conference on Document Analysis and Recognition*, 2011, to appear.

[3] J. Chen, W. Cheng, and D. Lopresti, "Using perturbed handwriting to support writer identification in the presence of severe data constraints," in *Proc. Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging)*, 2011.

[4] J. Chen and D. Lopresti, "A model based ruling line detection algorithm for noisy handwritten documents," in *Proceedings of the 2011 11th International Conference on Document Analysis and Recognition*, 2011, to appear.

[5] L. Ballard, J. Chen, D. Lopresti, and F. Monrose, "Biometric key generation using pseudo-signatures," in *Proceedings of The 11th International Conference on Frontiers in Handwriting Recognition*, Montreal, Canada, August 2008.

[6] J. Chen, D. Lopresti, L. Ballard, and F. Monrose, "Pseudo-signature as a biometric," in *Proceedings of the IEEE 2nd International Conference on Biometrics Theory, Applications and Systems*, 2008, arlington, VA, USA.

[7] J. Chen, D. Lopresti, and F. Monrose, "Towards resisting forgery attacks on pseudo-signatures," in *Proceedings of the 10th International Conference on Document Analysis and Recognition*, July 2009.

[8] J. Chen and D. Lopresti, "On the usability and security of pseudo-signatures," in *Proc. Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging)*, 2010.

[9] J. Chen, H. Cao, R. Prasad, A. Bhadwaj, and P. Natarajan, "Gabor features for offline arabic handwriting recognition," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, Boston, United States, June 2010.

[10] J. Chen, D. Lopresti, and E. Kavallieratou, "The impact of ruling lines on writer identification," in *Proc. of the 12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[11] E. Kavallieratou, D. Lopresti, and J. Chen, "Ruling line detection and removal," in *Proc. Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging)*, 2011.

Jin Chen **www.cse.lehigh.edu/∼jic207**
19 Memorial Drive West **(+1)484-560-8866**
Bethlehem, PA 18015 **jic207@cse.lehigh.edu**

Education

- **Lehigh University** **2007.09 - 2012.05** (expected)
  Ph.D Candidacy in Computer Science Bethlehem, PA, United States
  Research Assistant in *Pattern Recognition Lab*, working on off-line/on-line handwriting analysis.
  Thesis Title: Exploiting Metadata in Off-line Handwritten Documents: Modeling and Applications.
  Doctorate Committee: **Daniel Lopresti** (Thesis Advisor, Professor and Chair at Lehigh), **Xiaolei Huang** (Assistant Professor at Lehigh), **Brian Davison** (Associate Professor at Lehigh), **George Nagy** (Professor Emeritus at Rensselaer Polytechnic Institute), **Huaigu Cao** (Research Scientist at Raytheon BBN Technologies).

- **Lehigh University** **2007.09 - 2009.05**
  Master's Degree in Computer Science Bethlehem, PA, United States

- **Nanjing University of Science & Technology** **2002.09 - 2006.07**
  Bachelor's Degree in Computer Science Nanjing, Jiangsu, China
  Thesis work on biological image stitching has been published.

Research

- **Research Assistant** **2007.09 - Present**
  Pattern Recognition Lab **Lehigh University**
  - One primary task for document analysis is to translate a document into a symbolic transcription for knowledge retrieval. However, large amount of structural information exists beyond such a transcription, e.g., the handwriting idiosyncrasies, the 2-D arrangement of information, and the paper medium people write in. I proposed several algorithms to detect and recognize such information for extracting knowledge from an unstructured document corpus.
  - Traditionally, writer identification assumes sufficient amount of data for classifier training. In real-world scenarios, however, we might face severe data constraints. Thus I proposed a methodology for writer identification under severe data constraints by synthesizing handwriting where human subjects were involved for calibrating "realistic-looking."
  - Textual passwords are hard to remember and easy to guess, while physical biometrics such as fingerprints, faces are limited to create multiple passwords. Thus I proposed a graphical password scheme called "pseudo-signatures" that was designed to be easy to remember/use and also robust to forgery attacks from skilled human beings and intelligent machines.

- **Intern** **2011, 2010, 2009**
  Document Analysis Group **Raytheon BBN Technologies**
  - Applied region dependent feature transform for Arabic off-line handwriting recognition. Part of the work will be submitted to DAS 12'. (**2011**)
  - Developed a hierarchical SVM classifier for the Chinese OCR problem and built the classification system for writer identification under severe data constraints. Part of the work was published in DRR 11' and I delivered an oral presentation for it. (**2010**)
  - Extended former work on using Gabor features for Arabic off-line handwriting recognition. Developed other structural feature sets that combine with Gabor to achieve better performance. This work was published in DAS 10' and I delivered an oral presentation for it. (**2009**)

Professional Activities

- **Presentation**
  - Oral presentations at international conferences: ICDAR 09', DAS 10', DRR 09', 10'.

- Two-time speaker for Lehigh CSE Graduate Research Seminar Series (GRSS 09', 11').

- **Service**
  - Reviewer of ACM Computing Reviews, since 2011.
  - Conference paper review: ICDAR 11', DAS 08', DAS 10'.
  - Journal paper review: IEEE T-PAMI, IEEE T-MM, IEEE T-SMC, IJDAR.

- **Training**
  - Lehigh CSE Rossin Doctoral Fellows Program: to promote abilities for academic careers.

Publications

1. "Table Detection in Noisy Off-line Handwritten Documents," **J. Chen** and D. Lopresti, *International Conference on Document Analysis and Recognition*, September, 2011. [**Poster**]
2. "A Model-based Ruling Line Detection for Noisy Handwritten Documents," **J. Chen** and D. Lopresti, *International Conference on Document Analysis and Recognition*, September, 2011. [**Poster**]
3. "A Real-World Noisy Unstructured Handwritten Notebook Corpus for Document Image Analysis Research," **J. Chen** , D. Lopresti, and Bart Lamiroy, *the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, September, 2011. [**Oral**]
4. "Using Perturbed Handwriting to Support Writer Identification in the Presence of Severe Data Constraints", **J. Chen**, W. Cheng, D. Lopresti, *Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging*, January 2011. [**Oral**]
5. "Ruling Line Detection and Removal", E. Kavallieratou, D. Lopresti, and **J. Chen**, *Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging*, January 2011. [**Poster**]
6. "The Impact of Ruling-lines on Arabic Off-line Writer Identification," **J. Chen**, D. Lopresti, and E. Kavallieratou, *International Conference of Frontier Handwriting Recognition*, November 2010. [**Poster**]
7. "Gabor Features for Arabic Handwriting Recognition," **J. Chen**, H. Cao, R. Prasad, A. Bhadwaj, and P. Natarajan, *International Workshop on Document Analysis System*, June 2010. [**Oral**]
8. "On the Usability and Security of Pseudo-signatures," **J. Chen** and D. Lopresti, *Document Recognition and Retrieval XVII (IS&T/SPIE International Symposium on Electronic Imaging)*, January 2010. [**Oral**]
9. "Toward Resisting Forgery Attacks using Pseudo-signatures," **J. Chen**, D. Lopresti, and F. Monrose, *International Conference on Document Analysis and Recognition*, July 2009. [**Oral**]
10. "Pseudo-signatures as a Biometric," **J. Chen**, D. Lopresti, L. Ballard, and F. Monrose, *International Conference on Biometrics: Theory, Applications, and Systems*, September 2008. [**Poster**]
11. "Biometric Key Generation using Pseudo-signatures," **J. Chen**, L. Ballard, D. Lopresti, and F. Monrose, *International Conference of Frontier Handwriting Recognition*, August 2008. [**Poster**]

Work Experience

- **Summer Intern, Raytheon BBN Technologies, MA, United States** **2009, 2010, 2011**
- **Software Engineer, SNDA Networks, Shanghai, China** **2006.07 - 2007.07**

Skills

- **Programming**: C/C++, Tcl/Tk, Perl, Matlab, OpenMP/MPI. Familiar with Win/Linux/OSX programming platforms.
- **Language**: fluent English, native Mandarin.

Syntactic Model for Semantic Document Analysis

Lluís-Pere de las Heras
Computer Vision Center – Universitat Autònoma de Barcelona
lpheras@cvc.uab.es
Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola)
Barcelona, Spain

Ph.D Advisor: **Gemma Sánchez** (Computer Vision Center, Spain)
ICDAR D.C. Mentor: **Jean-Marc Ogier** (La Rochelle, France)

1. Problem Statement

The aim of this thesis is to create a general syntactic approach that, by taking into account the hierarchical and structural information between elements, will be capable to interpret and recognize different kinds of engineering documents. This task entails a recognition step over the elements contained in documents. Then, a representational model that allows describing hierarchical and structural relations between elements must be defined. After that, a grammar formalism for describing possible document class representations has to be inferred. Finally, a parser has to be implemented and applied as a classifier engine over the representational model accordingly with the grammar rules. Nevertheless, since the objective of creating a syntactic model capable to interpret and recognize any kind of engineering document is so ambitious, we have centered our work on a realistic environment in order to evaluate its viability. In order to do so, our model will be specifically used for architectural floorplan (see figure 1a) interpretation and recognition, but without forgetting that this model must be able to be extrapolated to any kind of engineering plan drawings, or even to any kind of document. Recently, several works on floorplan interpretation have been presented with different possible final applications [2, 3]. One of the possible final applications is to automatically interpret and render a building in three dimensions to allow users to navigate in a realistic environment. However, interpretation of floorplan documents is a non-solved problem, essentially because there is no standard notation defined, as it is shown in figure 1b. Elements such as walls, windows, furniture, indications, etc. are drawn distinctively depending on the architect and the country. Therefore, existing approaches are usually focused in one specific notation convention, and are not usable for the rest ones. These existing problems encourage us to construct a syntactic model capable to interpret every floorplan independently of its notation.

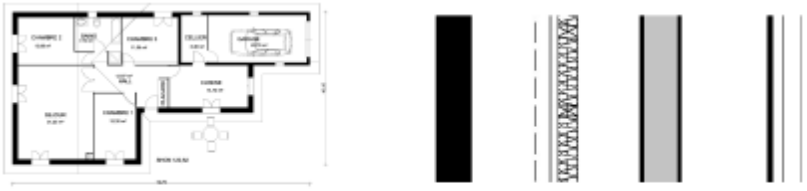


Figure 1: (a) Architectural floorplan example. (b) Wall intra-class variability.

2. Progress up to date

2.1 A basic syntactic system for floorplan interpretation

A basic syntactical model based on a stochastic grammar of images has been created in order to interpret floorplans. The grammar is built over an And-Or graph that allow to represent all plans hierarchically by means of a tree structure, and structurally and semantically by attribute nodes between elements at the same level of the tree. This grammar is augmented with three probabilistic models that were learned from a training-set to account the appearing frequency of the plan components and their spatial and structural characteristics. In the recognition step, a Bottom-up/Top-down parsing strategy is implemented to construct the And-graph that better represents the plan from its primitives to the root. Later, the branches of the graph that are not consistent with the grammar are pruned. Finally, an input instance is classified as a valid plan whether its parsing graph is consistent with the grammar productions and, in that case, the hierarchical, semantic and structural representation of the plan is returned by the process. In addition to that, since primitives at different levels of abstraction in the graphs would be extracted from plans in order to build the graph representation of a plan, two different extraction approaches were implemented: a patch window detection, and a region based room detection. The complete explanation of our syntactic system can be found in [1].

2.2 Notation invariant wall detector

One of the biggest drawbacks of the syntactic model explained above is that, since elements (primitives) of plans have to be extracted in order to construct its graph representation, our model was only able to recognize properly those plans with the notation that our extraction methods were oriented to. Therefore, an effective primitive extraction method capable to deal with the maximum number of notations has been studied and created. Actually, we have created two different patch-based detectors for walls: a Bag-of-Patches Wall Detector and a Descriptor-based Wall Detector.

2.2.1 Bag-of-Patches Wall Detector

The basic pipeline of the process is explained briefly below. Please, refer to [4] for the extensive explanation. In both cases, at learning and testing steps, there is a preprocessing that consists of binarization, text graphic separation and a common step for patch extraction based on defining a grid over the whole image. Regarding the grid, three different topologies have been considered: squared rigid-grid squared overlapped-grid and squared deformable-grid. After extracting features for every patch, in the learning phase, a dictionary of representative patches is created by clustering PCA feature vectors. Then, a probability of belonging to every class of objects is assigned to each word. In the testing phase, each patch is assigned to the nearest word in the dictionary, inheriting the class probabilities of the word. Moreover, a specific evaluation methodology has been proposed in order to achieve a realistic evaluation of our method at a pixel level. Finally, the method has been tested on two datasets containing plans with different notations and resolutions each. The good results obtained confirm the suitability of our method to detect walls in plans with different notation and resolutions.

2.2.2 Descriptor-based SVM Wall Detector

This approach is based on extracting different descriptors from each patch of the topology defined over the image, and later trains a Support Vector Machine. The process, which has been presented in [5], extracts all the patches from the learning-set images and computes a desired descriptor from each patch. Then, a SVM classifier is trained by selecting a specified number of randomly selected patches of each class. For a given test plan, every patch of the image is classified accordingly with its descriptor. This approach is tested for squared rigid-grid an overlapped topologies, using as patch descriptor the pixel intensity, PCA and Blurred Shape Model descriptor (BSM) separately. The method has been tested over the two same datasets used in [4] with the same evaluation methodology in order to compare their respective performances. The results obtained for all the descriptors in this method are closely similar than those ones obtained by Bag-of-Patches Wall Detector.

3. References

[1] Lluís-Pere de las Heras and Gemma Sánchez. And-or graph grammar for architectural floor plan representation, learning and recognition. A semantic, structural and hierarchical model. In Pattern Recognition and Image Analysis, volume 6669 of LNCS, pages 17--24. Springer Berlin / Heidelberg, 2011.

[2] Tong Lu, Chiew-Lan Tai, Feng Su, and Shijie Cai. A new recognition model for electronic architectural drawings. Computer-Aided Design, 37(10):1053--1069, 2005.

[3] Sébastien Macé, Hervé Locteau, Ernest Valveny, and Salvatore Tabbone. A system to detect rooms in architectural oor plan images. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pages 167--174, 2010.

[4] LP. de las Heras, J. Mas, G. Sánchez and E. Valveny: "Wall Patch-Based Segmentation in Architectural Floorplans". In *International Conference in Document Analysis and Recognition (ICDAR2011).* In press, Bejing (China), 2011.

[5] LP. de las Heras, J. Mas, G. Sánchez and E. Valveny: "Descritor-based Wall Detection on Floorplans". In *Graphic Recognition (GREC2011).* In press, Seoul (South Korea), 2011.

Lluís-Pere de las Heras Caballero

Name: Lluís-Pere de las Heras Caballero
Identity number: 46980076-T
Date of birth: 23.07.1985, in Barcelona
Address: Cavallers 12 K Baixos 2ª
08034 Barcelona, Spain
Nationality: Spanish
Telephone: Land-line: +34 932 033 406
Mobile: +34 617 967 319
e-mail: lpheras@cvc.uab.es
luigggi@gmail.com

EDUCATION

- 2010 - 2014 expected.
Ph.D Student in Computer Vision and Artificial Intelligence:
Syntactic Model for Semantic Document Analysis.
Computer Vision Center - UAB, Barcelona, Spain.

- 2010 - 2011
Master in Computer Vision and Artificial Intelligence.
Universitat Autònoma de Barcelona, Barcelona, Spain.

- 2009
Bachelor in Computer Science
Universitat Autònoma de Barcelona, Barcelona, Spain.

PUBLICATIONS

- LP. de las Heras, J. Mas, G. Sánchez and E. Valveny: "Wall Patch-Based Segmentation in Architectural Floorplans". In *International Conference in Document Analysis and Recognition (ICDAR2011).* Pre-print, Bejing (China), 2011.

- LP. de las Heras, J. Mas, G. Sánchez and E. Valveny: "Descritor-based Wall Detection on Floorplans". In *Graphic Recognition (GREC2011).* Pre-print, Seoul (South Korea), 2011.

- LP. de las Heras and Gemma Sánchez: "And-Or Graph Grammar for Architectural Floorplan Representation, Learning and Recognition. A Semantic, Structural and Hiararchical Model". *In Iberian Conference in Pattern Recognition and Image Analysis (IbPRIA2011).* pre-print, Las Palmas de Gran Canaria (Spain), 2011.

- LP. de las Heras and Gemma Sánchez: "Syntactic Model for Semantic Document Analysis". *In Fifth CVC Workshop on the Progress of Research and Development (CVCRD2010).* Computer Vision Center, Barcelona (Spain) , 2010.

RESEARCH PROJECTS

- **ScanPlan: Architectural Floor-plan Interpretation** (2010 - X)
  Director: Josep Lladós
  Institution: Computer Vision Center
  Group: Document Analysis Group
  URL: http://www.cvc.uab.es/projectes2.asp?id=246&scanplanarchitectural-floorplan-interpretation

- **IOCS: Investigation Oriented Car Simulator (**2009-2010) (Catalan language)
  Director: Felipe Lumbreras
  Institution: Universitat Autònoma de Barcelona / Computer Vision Center
  Group: Automatic Driver Assistance Systems
  URL:
  http://ddd.uab.cat/pub/trerecpro/2009/hdl_2072_43812/PFC_LluisPdelasHerasCaballero.pdf

EXTRA R&D WORK

- **Editorial Assistant of an Electronic Journal in Computer Vision** (2011 - X)
  El. Journal: Electronic Letters on Computer Vision and Image Analysis
  Editors: Josep Lladós and Simone Marinai.
  URL: http://elcvia/index.php/elcvia

LANGUAGES

- Catalan and Spanish: Native languages
- English: Fluent oral and written. (B2.2 European Level)

COMPUTER VISION KNOWLEDGE

– Computer graphics.
– Human and computer perception strategies.
– Basic image processing techniques.
– Image segmentation strategies.
– Knowledge representation.
– Reasoning and Uncertainty.
– Basic knowledge in robots action.
– Heuristic Search and Machine Learning

Table Recognition and Evaluation in PDF Documents

Jing Fang
Advisor: Zhi Tang
Institute of Computer Science & Technology,
Peking University, China
{fangjing, tangzhi}@icst.pku.edu.cn

1. Research Background

In recent years, mobile reading has rapidly gained popularity and handheld devices (e.g. smartphones, Kindle, iPad, etc.) are more and more used as platform for rendering and displaying of electronic documents (e.g. e-Books). Generally, Portable Document Format (PDF) is widely used because of its ability to preserve the appearance of the original documents. However, on the relatively small screens of those devices, the documents usually need to be re-flowed and recomposed to avoid readers moving the screen back and forth. This proposes logical structure extraction requirements for the fixed-layout and untagged documents.

Table, as an efficient and compact means to present data, is an important structural document component. Automatic recognition of tables is of great meaning for handheld device reading. In the lower level, the table regions should be separated from other elements of pages, so that they could be taken as integrated objects to be rendered on handheld device screens. While in the higher level, the detailed structure of tables, namely columns, rows and cells should be recognized. In this way, a large table can be re-edited and displayed in continuous screen pages, or people can choose to hide some columns or rows according to their reading preference. Apart from these, tables are also significant data source to be indexed and searched. Consequently, I choose table recognition as the main research work of my thesis.

Since a good number of research efforts have been made on table recognition, there is another important issue “How to evaluate which algorithm is better in different application scenarios”. It is well known that, a large quantity and representative ground-truthed dataset and performance metrics are two important factors of performance evaluation problem. For the image documents which have gone through several decades’ research, there are already some datasets widely used, such as UW-III, UNLV etc. However, researches on PDF format have been carried out only in recently years, and most of the proposed algorithms are evaluated on their in-house and small datasets. Therefore, building a publicly accessible and large-sized dataset, and proposing table recognition performance metrics are also included in my thesis.

2. Proposed Plan

Goal 1: Table Detection

Table detection, also named table spotting or table location, refers to separating table regions from non-table regions in a given page. It is the first and crucial step for subsequent recognition

stages. A good number of researches have been addressed on this topic. According to the document media type, different algorithms can be classified into four categories: image documents table detection, plain text (e.g. ASCII) table detection, web pages table detection and PDF documents table detection.

My thesis research aims at detecting tables from PDF documents, which is concerned just in recent years. Most of existing published works made use of merely cell layout information to locate tables. This method works well for regular tables, but become ineffective when dealing with irregular tables (e.g., sparse tables) or tables in complex-layouted pages. Through observation, we found that the more complex a document and the more irregular the layout of a table, the more graphic lines are employed as borders and rules. As a result, in my thesis, the ruling lines and table contents would both be treated as valuable sources for spotting table regions.

Goal 2: Table Structure Analysis

The purposes of table structure analysis are three fold: *i)* give a physical description of the table, i.e., identify its cells and their relative positions, as well as its rows and columns from the detected table regions; *ii)* extract logical structure of table cells, i.e., determine the heading rows and columns, and relationship between the indication cells and body data cells; *iii)* identify the affiliated table attributes, such as table caption, table footnotes and descriptions from text paragraphs. These optional components provide basic semantic information of tables.

In recent years more and more researchers were addressing the topic of table structure analysis. However, the performances are still not satisfactory. The challenges are mainly caused by varied ways tables can show up in real-world documents, e.g. nested tables, sparse tables, tables with spanning cells etc. Therefore, table structure analysis, including physical segmentation, logical relation mining and shallow semantic understanding will be an important part of my thesis research.

Goal 3: Table Recognition Evaluation

Almost all of the existing table recognition algorithms were evaluated on their in-house data set. No general table ground-truth dataset for PDF documents is publicly available. This often makes it difficult to define what constitutes a 'correct' recognition result. Therefore, we plan to build and make public a table dataset in PDF format with ground-truths.

Another issue of table performance evaluation is about performance metrics. Most of existing published papers analyze experiment results using precision & recall metrics, which are widely used in the classification and information retrieve field. Actually, these are not enough for table recognition evaluation. For the table detection subtask, error type is more than false positive and false negative and tables may be partially recognized, amplified, etc. While for structure analysis, evaluation metrics should work for not only physical segmentation, but also logical relationship and semantic interpretation. None of current work covers all these aspects. Therefore, proposing proper evaluation metrics will also be covered in my thesis research.

3. Progress to Date

- **Table Detection**

I have proposed a table detection method via visual separators and geometric content layout information, targeting at PDF documents. The visual separators refer to not only the graphic ruling lines but also the white spaces to handle tables with or without ruling lines. Furthermore, page columns are detected in order to assist table region delimitation in complex layout pages. Evaluations of our algorithm on an e-Book dataset and a scientific document dataset show competitive performance. This part of work is to be published by ICDAR 2011.

- **Table Detection Evaluation**

Currently, I have proposed and implemented a set of metrics to evaluate table detection algorithms. The construction of a dataset with both Chinese and English pages in PDF format is nearly complete, together with XML-based ground-truths. We are going to evaluate and compare several existing table detection algorithms soon.

- **Relation between Tables and Non-table Components**

On the basis of table detection, the other logical page components are also recognized from the non-table regions, including text paragraphs, titles, captions, lists, footnote, etc. A reading order detection algorithm is implemented to recover reading order among all the page components so that the re-flowed content conforms to human’s reading custom.

CURRICULUM VITAE

PERSONAL DATA

SURNAME:	Fang
FIRST NAME:	Jing
ADDRESS:	Institute of Computer Science & Technology, Peking University, 100871, P.R.China
PHONE NUMBER:	86(10)-15120094118
E-MAIL ADDRESS:	fangjing315@gmail.com; fangjing@icst.pku.edu.cn
SUPERVISOR:	Prof. Zhi Tang

EDUCATION

Sep. 2008 – Present Ph.D. Candidate of Computer Science (expected to graduate in 2013),
Institute of Computer Science & Technology, Peking University, China

Sep. 2004 – Jul. 2008 Bachelor of Software Engineering,
School of Computer Software, Tianjin University, China

RESEARCH INTERESTS

Table Recognition and Performance Evaluation;
Document Analysis and Understanding;
Pattern Recognition; Machine Learning;

PUBLICATIONS

1. **Jing Fang**, Liangcai Gao, Kun Bai, Ruiheng Qiu, Xin Tao and Zhi Tang. A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures. To appear in ICDAR 2011**.**
2. **Jing Fang**, Zhi Tang and Liangcai Gao. Reflowing-driven Paragraph Recognition for Electronic Books in PDF. Proc. of the SPIE, Volume 7874, pp. 78740U-78740U-9 (2011).
3. Liangcai Gao, Zhi Tang, **Jing Fang**, Xiaofan Lin, "Multi-page document analysis based on format consistency and clustering", IJCAI, Vol.38 (4), 2010.
4. Liangcai Gao, Zhi Tang, Xin Tao, **Jing Fang**, "An Approach to Auto-detection, Segmentation and Tagging of Bibliographic Metadata", Acta Scientiarun Naturalium Universitatis Pekinensis,Vol.46(6), 2010.
5. Liangcai Gao, Zhi Tang, Xiaofan Lin, Yinyan Yu, **Jing Fang**, “A Table of Content Recognition method of Book Documents Based on Clustering Techniques”, Acta Scientiarun Naturalium Universitatis Pekinensis, Vol.46 (4), 2010.

PATENTS

1. **Jing Fang**, Zhi Tang, Liangcai Gao, Xin Tao, ”An approach and apparatus to detect reading order in digital documents”, 2010. Accepted. 201010559135.3
2. Liangcai Gao, Zhi Tang, **Jing Fang**, Ruiheng Qiu, "An approach and apparatus to recognize page information in digital documents “, 2010. Accepted. 201010193898.0.

Investigations on the use of linear-chain CRF based method to segment old newspapers

David Hebert
Laboratoire LITIS EA 4108 Université de Rouen, France
Email: David.Hebert@univ-rouen.fr
Thesis Advisor: Pr. Thierry Paquet

Accepted paper at ICDAR'11: "Continuous CRF with multi-scale quantization feature functions, Application to structure extraction in old newspaper"

I. PhD Thesis context

My PhD thesis subject is about the use of Conditional Random Fields (CRF) models for structure extraction in complex data such as old documents digitized from archives. The aim is to develop some extraction techniques based on the CRF formalism to extract and label some entities from document images.

Various choices can be made to achieve this kind of task. The use of image data suggests the use of a bi-dimensional approach to model a 2D mesh of pixels or sites but 2D approaches are computationally complex and results are theoretically non-optimal, due to restrictions to ensure that the calculations are feasible.

II. The linear-chain CRF formalism

The formalism introduced in 2001 by Lafferty et al. for Part Of Speech tagging tasks is a simple and easy to understand concept: keep in memory some configurations of observation – word– realizations in a defined context and save the label associated with each configurations that have been seen on training data. This can be seen as a memorization process of sub-parts of a training set with a lightning procedure giving more importance (higher potential values) to stored configurations that are the most often encountered and/or the most discriminative.

Given a new observation, the labeling process consists in finding known configurations in the context of this new data and combining potentials associated with.

III. Our approach

We decided to use this formalism, so we had to adapt the image data in order to be able to use the general idea. Indeed, numerical features are computed to characterize image content more than the simple use of pixel colors. These values are not discrete such as words are in POS tagging tasks. Some approaches use the discrete output of a classifier as the input of a CRF so a classification stage is required, leading to a need for knowledge, and a local decision. These local decisions are then contextualized by a CRF model.

We decided to maximize the CRF contribution in the system by avoiding the use of a classification stage to discrete data. The adaptation of continuous data to discrete one is performed by a quantization stage, the simplest way to convert continuous to discrete values. This data adaptation stage does not take any local decision and only reorganize input data. To avoid the choice of a fixed (and irrelevant) quantization step, we use a set of J quantization functions each using different quantization steps. Each original continuous value is quantized J times by J distinct quantization steps. These J values are then given as input to a discrete CRF model working on the formalism described in section II. Doing this, we assume that the parameter estimation of CRF during the training process will perform a selection of the best quantization functions by increasing or decreasing potentials.

We also use a sequential model of pixel lines and each image is analyzed line by line. This complexity restriction is only possible because a document is highly structured with horizontal and vertical dependencies.

IV. Discussions

These choices have resulted in a pixel lines model to extract structures from documents that use the formalism well used in automatic language analysis domain. However, the use of a horizontal model does not take into account the all vertical dependencies, useful to extract vertically oriented entities in a document such as a vertical separator. One of our research perspectives is the use of additional information. Several improvements will be investigated in this way, without going to the use of a 2D CRF model at a pixel level:

- The use of a context of vertical combinations of observation realizations
- The use of a second sequential CRF to model vertical dependencies and combine results with the horizontal one
- The integration of the CRF formalism described above in a 2D agglomerative, multi-resolution method for pixels (eg. quad tree) to make a multi-resolution analysis
- The use of a low level, local, feature classification stage (SVM or Deep Neural Network) prior to the contextual analysis made by the CRF

The results obtained with the system previously described are very encouraging but we ask ourselves the question of the influence of redundancy introduced in the input data by the multi-scale quantization stage. CRF is theoretically able to select the most appropriate values but is the redundancy complicates this selection process?

David HEBERT

Phd Student

LITIS EA 4108
Ufr Sciences & techniques
Avenue de l'Université
76800 Saint Etienne du Rouvray, FRANCE
Email : david.hebert@univ-rouen.fr

Education

- I obtained a Master's degree in Multimedia information processing with honors at the University of Rouen (France) in 2009. During these two years I studied pattern recognition and discovered the main method for data classification, image treatment and segmentation, signal processing and sequence modeling apply on handwriting documents.

Medical images processing

- After a preliminary study of 3 month, I did my first research experience in 2008 during a short training period where I worked on medical images of kidneys where the aim was to determine a kidney volume using a sequence of slices. The chosen method used the belief function theory and more generally information combination.

- During a student project, I also did a short work on a lung images classification task. The aim was to investigate the use of a boosted cascade of classifiers as defined by Viola and Jones for face detection, to classify lung images as "healthy" or "non-healthy" and identify critical parameters in this method used in a classification context. Results on this work have been submitted for publication and I am currently waiting for reviews.

Document images processing

- At the end of my master's degree, I did a second training period of 6 months. The subject was the extraction of structures in document images, digitized from regional newspaper archives, using linear-chain HMM like method. This work [1] has been presented in RFIA'10, a French conference on pattern recognition and artificial intelligence.

- I had started my PhD studies at the end of 2009 with Pr. Thierry Paquet as advisor at the LITIS laboratory. My thesis title is "Conditional Random Fields for structure extraction in complex data". By developing a CRF based approach, I have to extract structures in document images. The firsts 8 months was dedicated to the improvement of previous results obtained during my 6 months training period.
  The first international communication on this work will be made during ICDAR'11 [2].
  My expected date of graduation is October 2012

Publications

[1] D. Hebert, T. Paquet, S. Nicolas, Champs de Markov Cachés 2D à composantes séparables, application à l'extraction de structures de journaux anciens, RFIA'10, Caen, France, 2010

[2] D. Hebert, T. Paquet, S. Nicolas, Continuous CRF with multi-scale quantization feature functions, Application to structure extraction in old newspaper, ICDAR'11, Beijing, China, 2011 [to be published]

RECOGNITION AND RETRIEVAL OF HANDWRITTEN MATHEMATICAL EXPRESSIONS

LEI HU
ROCHESTER INSTITUTE OF TECHNOLOGY
LEI.HU@RIT.EDU
ADVISOR: DR. RICHARD ZANIBBI

1. The Problem

Recognition and retrieval of textual information is fairly mature, but recognition and retrieval of mathematical expressions are in comparatively early stages of research [1].

Mathematical expressions are an indispensable component of scientific and technical literatures [2]. So far the most popular way to enter mathematical expressions is either in a linear format (e.g., TEX), or by using a structured editor (e.g., equation editor available with MS-Word) [3]. Producing large and complicated expressions in these two ways requires a lot of time and mental effort. With the emergence of pen-based electronic devices, such as PDAs and tablet PCs, it will be exciting that people can simply write mathematical expressions on the electronic tablet to let the computer recognize them automatically.

Recognition of math expression is the basis of the retrieval of math expression, which means recognition of math expression could provide the query to the retrieval. There are many fast and effective text retrieval methods. Buy when we want to find a math expression in the documents, the text retrieval methods are not enough. Now people have to remember the name or corresponding keyword of the math expression to search it. It would be very helpful if people could locate the math expression based on the math expression that he or she draws online.

2. Solution

Our research now focus on the recognition of online handwritten mathematical expressions. Recognition of mathematical expressions includes two major steps: symbol recognition and structural analysis [2]. Symbol recognition is the basis of the structural analysis. It consists of two phases: symbol segmentation and isolated symbol recognition. The input data of online handwritten mathematical expression is a set of strokes, and a mathematical symbol may comprise more than one stroke. Symbol segmentation aims to transform the sequence of strokes into a set of symbols, which will be classified in the isolated symbol recognition stage.

A number of approaches have been proposed for online handwritten mathematical symbol recognition. Then can be divided into four categories: nearest neighbor based methods [4] [5] [6], rule-based methods [7] [8], combination of different classifiers [3] and statistical approaches [9] [10].

We have established a recognition system based on Hidden Markov Model (HMM) for isolated online handwritten mathematical symbols. Fig. 1 shows the flow chart of our system. We design a continuous left to right HMM for each symbol class and use four online local features, including a new feature: normalized distance to stroke edge (NDTSE). The new feature can be computed as :

$$NDTSE(s,t) = \begin{cases} 1 - \frac{|d_e - d_b|}{l_s}, & \text{for actual stroke} \\ -(1 - \frac{|d_e - d_b|}{l_s}), & \text{for interpolated stroke,} \end{cases} \tag{1}$$

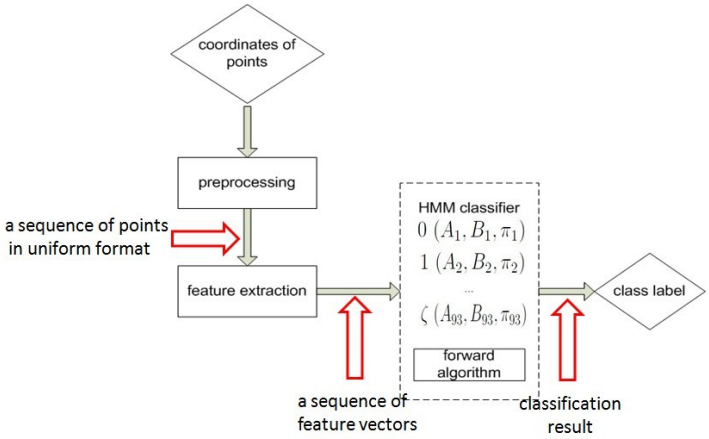


FIGURE 1. Flow chart of our system.

where l_s represents the length of the stroke s which the current point $x(t), y(t)$ belongs to; d_e represents the distance between the current point and the last point of s; d_b represents the distance between the current point and the first point of s. Actual stroke is the visible stroke, while interpolated stroke is the hidden parts of the trajectory, where the digital pen does not contact with the electronic tablet. For the point belongs to actual stroke, NDTSE is nonnegative; for the point belongs to interpolated stroke, NDTSE is nonpositive. Fig. 2 visualizes the new feature.

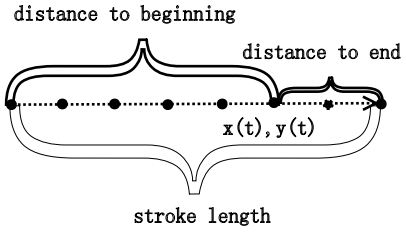


FIGURE 2. New feature: normalized distance to stroke edge, containing the pen-up/down information and the location information of the current point

A variant of segmental K-means is used to get initialization of the Gaussian Mixture Models' parameters which represent the observation probability distribution of the HMMs. The system has acquired encouraging results on the symbols from a new publicly available, ground-truthed corpus of handwritten mathematical expressions [11].

3. FUTURE PLAN

In the next year, we will focus on implementing new segmentation and parsing algorithms based on Bayesian network techniques or syntactic techniques. We also plan to extend the research on recognition of online handwritten mathematical expressions to the recognition of offline handwritten mathematical expressions. Our final goal is to create a retrieval algorithm for mathematical expressions to make locating them in documents as easy searching for text.

References

[1] R. Zanibbi and D. Blostein, "Recognition and Retrieval of Mathematical Expressions," *International Journal on Document Analysis and Recognition*, to appear.

[2] K.-F. Chan and D.-Y. Yeung, "Mathematical expression recognition: a survey," *International Journal on Document Analysis and Recognition*, vol. 3, no. 1, pp. 3–15, Aug. 2000.

[3] U. Garain and B. Chaudhuri, "Recognition of online handwritten mathematical expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 6, pp. 2366–2376, 2004.

[4] S. Smithies, K. Novins, and J. Arvo, "A handwriting-based equation editor," *Proc. Graphics Interface*, pp. 84–91, June 1999.

[5] B. Vuong, Y. He, and S. Hui, "Towards a web-based progressive handwriting recognition environment for mathematical problem solving," *Expert Systems with Applications*, vol. 37, no. 1, pp. 886–893, 2010.

[6] S. Maclean and G. Labahn, "Elastic matching in linear time and constant space," in *Proc. Ninth IAPR Int'l. Workshop on Document Analysis Systems*. ACM, 2010, pp. 551–554.

[7] J. Fitzgerald, F. Geiselbrechtinger, and T. Kechadi, "Mathpad: A fuzzy logic-based recognition system for handwritten mathematics," *In Proc. International Conf. on Document Analysis and Recognition*, vol. 2, pp. 694–698, 2007.

[8] A. Belaid and J.-P. Haton, "A syntactic approach for handwritten mathematical formula recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, pp. 105–111, Jan. 1984.

[9] N. Matsakis, "Recognition of handwritten mathematical expressions," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, May 1999.

[10] H.-J. Winkler, "HMM-based handwritten symbol recognition using on-line and off-line features," *In Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 3438–3441, 1996.

[11] S. MacLean, G. Labahn, E. Lank, M. Marzouk, and D. Tausky, "Grammar-based techniques for creating ground-truthed sketch corpora," *International Journal on Document Analysis and Recognition*, pp. 1–21, May 2010.

Lei Hu **http://people.rit.edu/lxh9868**
Rochester Institute of Technology lei.hu@rit.edu
Rochester, NY, United States (+1)479-871-9667

Education

- **Rochester Institute of Technology** Rochester, NY, United States
  Ph.D. in Computing and Information Sciences (in progress) *2010.09 – 2014.05(expected)*
  - Advisor: Richard Zanibbi
  - Research Interests: Recognition and Retrieval of Handwritten Mathematical Expressions
- **Wuhan University of Technology** Wuhan, Hubei, China
  M.S. in Computer Science and Technology *2008.09 – 2010.06*
  - Advisor: Hongxia Xia
  - Thesis: Research and Application of Fuzzy Controller Based on Granular Computing
- **Wuhan University of Technology** Wuhan, Hubei, China
  B.S. in Computer Science and Technology *2004.09 – 2008.06*
- **Wuhan University** Wuhan, Hubei, China
  B.S. in Business Administration *2006.03 – 2008.06*

Research

- **Research Assistant** 2010.09 – Present
  Document and Pattern Recognition Lab *Rochester Institute of Technology*
  - Apply Hidden Markov Model on online isolated handwritten mathematical symbols.

Professional Activities

- **Service**
  - Conference paper review: ICDAR 11'.

Recent Publications

1) **L. Hu** and R. Zanibbi. *HMM-Based Recognition of Online Handwritten Mathematical Symbols Using Segmental K-means Initialization and a Modified Pen-up/down Feature.* Int'l. Conf. Document Analysis and Recognition, Beijing, China, September 2011(to appear).

2) **L. Hu**, H. Xia and H. Wang *Fuzzy Control Applied to Drainage System of City Highway Tunnel.* The 2nd International Workshop on Education Technology and Computer Science, pp 810-813, Wuhan, China, 2010.

Touching Text Segmentation and Shape Analysis

Le Kang Advisor: David Doermann
University of Maryland
College Park, MD 20742, USA
{lekang, doermann}@umiacs.umd.edu

Segmentation of a document image into basic units is typically required by handwritten text recognition system, and when text touching affects segmentation, recognition failure may result. To solve the text touching problem, a good understanding of character shapes is considered necessary. My research is therefore focused on touching text segmentation and shape analysis.

Touching exists in different kinds of situations. If we consider the property of text, touching can be categorized into two kinds: heterogeneous and homogeneous. Heterogeneous means the components that form the touching come from different types of text, like machine print vs. handwritten, or Chinese vs. English. Comparatively, homogeneous indicates same sort of text. If we focus on the position of touching, it can be divided into two categories: in-line and between-line. In-line touching occurs between adjacent letters, characters or words in the same text line and usually embodied in joined-up writing. Between-line touching involves text components from two adjacent lines, mostly caused by protruding characters.

In current stage, my concentration is on homogeneous between-line touching. In previous work, between-line touching segmentation is usually closely combined with text line segmentation, or sometimes even ignored: touching components are simply cut by the boundaries of text lines. However, to further improve the final OCR rate, we need to treat the between-line touching more carefully. Techniques from traditional in-line character segmentation may be applied to this problem.

The segmentation of touching components in general has been addressed extensively in the literature where they are typically segmented according to character size, profile, or junction points. Two primary approaches to the segmentation of touching characters can be found, recognition-free and recognition-based. Recognition-free segmentation techniques generally include contour, skeleton and projection profile analysis, and use only structural information. These empirical methods may not be robust enough to handle a lot of variations in practice, though they can efficiently address certain touching problems. Recognition-based segmentation usually generates multiple candidate segmentation hypotheses and selects the optimal one based on recognition or other evaluation functions. The problem with this kind of segmentation methods is that they naturally require heavy computation in pursuit of accuracy and they succeed only when the correct segmentation is in one of the candidates and a proper recognition kernel is available.

The proposed plan is to address this problem in two aspects. One is to set up a framework that handles the segmentation in a machine-learning manner with minimum heuristics, which should be easily adapted to various situations (different languages, stroke width, style, etc). The other is

to exploit the geometric property of character shapes so that simple touching patterns can be segmented very efficiently.

Majority of work in the first aspect has already been done. We proposed a template based approach to the segmentation of touching components in handwritten text lines. For simplicity, it is assumed that local patches around touching components can be identified. In fact, locating the touching patch may be based on some text line detection techniques. A dictionary is created consisting of template touching patches together with their correct segmentations. We use shape context based methods to compute similarity between input patches and dictionary templates to find the best match. Actually we compared the application of original shape context and inner-distance shape context in this process and the former seems more suitable. The template's known segmentation is then transformed using the thin plate spline to segment the input patch. Experiments are carried on a dataset containing handwritten Arabic documents where text lines are moved close to their neighbors to create touching artificially.

For the second aspect, graph and clustering based methods are in development. The basic idea is to design some quantity that can describe the connectivity between sections of strokes, and then cut at low strength connections to obtain reasonable segmentation. We need to understand properties of character shapes and how human groups multiple visual elements.

In future we will improve the template based segmentation in accuracy, and complete the graph based method. It is possible to combine the two in a real system to strike a balance between efficiency and accuracy. Finally we expect to address various touching problems, heterogeneous or homogeneous, in-line or between-line, in the same framework as proposed.

Le Kang

Research Assistant (Aug.2009 - present)
Laboratory for Language and Media Processing
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
lekang@umiacs.umd.edu

EDUCATION

University of Maryland, College Park
Ph.D. student in Dept. of Electrical and Computer Engineering Aug. 2009 – Jun. 2014

Tsinghua University
B.S. in Electronic Engineering/ Information Science Jul. 2009

RESEARCH EXPERIENCE

LAMP, Univ. of Maryland, College Park Jul. 2010 - Mar. 2011
Project Title: Template based Touching Components Segmentation in Handwritten Text Lines

- Created a semi-synthetic dataset of local touching regions through text line proximity technique and constructed a dictionary of patches of touching components
- Designed a shape contexts and thin plate spline based framework to segment query touching patches according to solutions in the dictionary

LAMP, Univ. of Maryland, College Park Sep. 2009 - Jun. 2010
Project Title: Synthetic Dataset Generation and Handwritten Text Line Segmentation Evaluation

- Designed relative/absolute text line proximity methods and touching labeling protocol for synthetic handwritten text dataset generation
- Evaluated the performance of text line segmentation algorithms under different levels of proximity

PUBLICATIONS

- Le Kang and D. Doermann, “Template based Touching Components Segmentation in Handwritten Text Lines”. *Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, 2011, in press.
- J. Kumar, Le Kang and D. Doermann, “Segmentation of Handwritten Text Lines in Presence of Touching Components”. *Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, 2011, in press.
- J. Kumar, W. Abd-Almageed, Le Kang, D. Doermann. “Handwritten Arabic Text Line Segmentation using Affinity Propagation”. *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 135-142, 2010.

Research Plan
Learning-Based Word Spotting for Arabic Handwritten Documents Using Hierarchical Classi er

Student Name: Muna Khayyat
Supervised By: Dr. Ching Y. Suen and Dr. Louisa Lam

August 14, 2011

1 Introduction

Word spotting for Arabic handwritten documents has been receiving more attention recently. We proposed methods and techniques to look for Arabic words within handwritten documents. This research plan provides the essence of my work which may be summed up in the following steps:

- Pre-processing and line segmentation of Arabic handwritten documents.
- Extracting feature from Arabic handwritten word.
- Segmenting Arabic handwritten documents and words into Pieces of Arabic Words (PAW).
- Document pruning for the proposed Arabic words spotting system.
- Implementing our proposed hierarchical classifier that aims to solves the lack of boundaries problem.

2 Problem

A great number of handwritten documents have been digitized to preserve, analyze, and disseminate them. These documents are of different categories, being drawn from fields as diverse as history, commerce, finance, and medicine. As the sheer number of handwritten documents being digitized continues to increase, the need for indexing them becomes vital. Word spotting is an approach that allows the user to search for keywords in spoken or written text.

Most research on word spotting has been implemented for Latin-based and Chinese documents. However, few word spotting systems have been implemented for Arabic handwritten documents. Yet, Arabic is spoken by a significant number of the world's population. Arabic script is cursive by nature; besides, in Arabic writing words have no clear boundaries. These facts make the implementation of word spotting for Arabic handwritten documents a significant challenge.

We proposed a learning-based word spotting system that uses Support Vector Machine (SVM) to recognize sub-words rather than complete words. We will use this partial segmentation concept to resolve the word boundary problem in Arabic handwriting. In addition, a shortest distance algorithm is used to spot Arabic handwritten words.

3 Proposed Plan

The exact form a word spotting system takes is application dependent, but can be divided into two broad classes depending on the lexicon: open lexicon and closed lexicon. We are using the Center for Pattern Recognition and Machine Intelligence (CENPARMI) Arabic database. It is a closed lexicon of 69 Arabic handwritten words with about 200 handwritten documents to validate our system. These documents contain all the lexicon words.

In our work we will make use of the concept of PAW to train and detect the lexicon words. The CENPARMI isolated Arabic words database contains *handwritten* words and documents of diverse writing styles. We are reconstructing this database by partially segmenting it into PAWs. In addition, we extracted gradient, Gabor, and Fourier transformation features from Arabic handwritten words.

Matching techniques of Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) are frequently used to spot words. In our word spotting system the SVM — which has good classification accuracy and the ability to discriminate between classes — will be used in conjunction with the Dijkstra algorithm. Given all the above, we proposed a learning-based word spotting system that trains PAWs rather than words in a hierarchical routine. For the first time, the Dijkstra algorithm will be employed for the detection of words. Thus, the outcome of our system will be a method that attempts to overcome the difficulty of finding the boundaries of Arabic handwritten word within a document. Figure 1 shows the flowchart of our proposed system for Arabic handwritten word spotting.

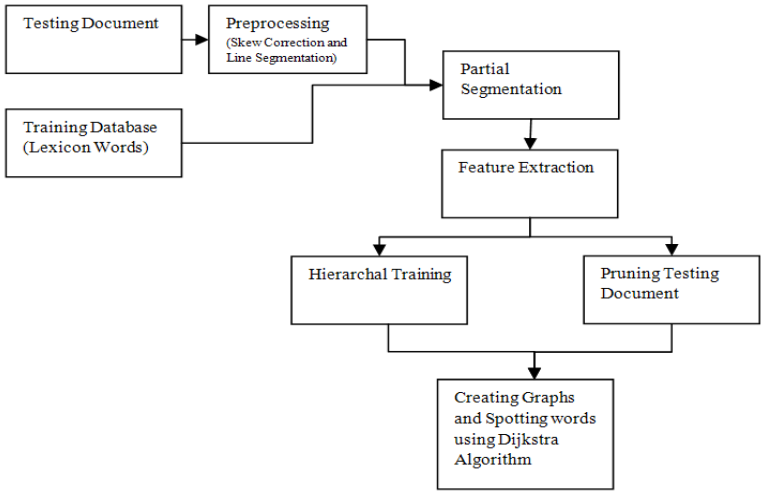


Figure 1: The Proposed Word Spotting System Flowchart

4 Progress to date

1. **Arabic handwritten word recognition:**
   We constructed a recognition system to recognize handwritten Arabic Words and Sub-words (PAW) as well. This system extract three different feature sets from the words images, namely: Gradient features using Roberts filter, Gabor features, and Discrete Fourier transform from the upper, lower and projection profiles of the word images. The three different feature sets were fed into three different classifiers and the classifiers were combined to recognize the word images. This system was submitted to ICDAR-11 Arabic handwritten words recognition competition.

2. **Partial Segmentation of the CENPARMI Arabic isolated handwritten word database:**
   We partially segmented our database into PAWs using heuristics. The 69 Arabic handwritten classes of our

database were partially segmented into 127 classes. These classes are within two different groups: The first group is the PAWs that are formed from segmenting the words into the PAWs resulted from partitioning the printed words and those bring up 92 classes. The second group resulted from tolerating different writing habits that result on touching PAWs and disconnecting some of those that should be connected, those form the rest of the classes. To end up with a database of 127 classes of PAWs. This database is going to be used to train the pruning classifier, also some PAWs will be merged to train our proposed hierarchical classifier.

3. **Arabic text line extraction:**
   Many handwritten text line extraction has been proposed. Some of them has shown promising results and they were successfully tested on ICDAR text line competition database. However, the Arabic language has some characteristics that makes it different from any other language. Not because of its cursiveness, but because of the presents of dots and diacratics that make it more difficult when it goes to text line extraction. We proposed an Algorithm for text line extraction from Arabic handwritten documents. We extracted the lines from the document using this algorithm with a promising accuracy. This algorithm is to be published.

5 Current Work

Currently, we are partially segmenting the lines that we extracted into PAWs. We constructed a classifier for the pruning process. This classifier uses the PAWs that were extracted from the CENPARMI isolated words database for training. We already trained the classifier. Accordingly, after extracting all the PAWs from the document we will prune our documents and analyze the effect of the pruning to some published word spotting systems. The CENPARMI handwritten isolated words and documents will be used to test our pruning system.

6 Future Work

For future work we are planning to do the following:

1. We plan to extract one extra set of features, that may bring higher classification accuracy. The features will be extracted from the PAWs and tested as well.
2. We will construct our hierarchical classifier and test it for Arabic handwritten words.
3. We will perform our word spotting system using the aforementioned classifier.

Muna Khayyat

2250 Guy Street, Apt. 1404
Montreal, QC, Canada, H3H 2M3

Phone: +1 (514) 261-4022
Email: m_khay@encs.concordia.ca

Objectives

To participate in the PhD consortium in the International Conference of Document Analysis and Recognition - ICDAR-11

Education

Ph.D. Computer Science, Concordia University (Montreal, QC, Canada), September 2009 - April 2013. Research Direction: Document Analysis and Recognition.

M.A. Computer Science, University of Jordan (Amman, Jordan), September 2002 - August 2004. Thesis Title: A Comparative Analysis of A Machine Ability to Recognize Handwritten Hindu and Arabic Digits Using Extracted Features.

B.S. Computer Science, Birzeit University (Ramallah, Palestine), September 1996 - January 2000. Graduation Project Title: Internet Shopping Using Java Servlets.

Employment

Concordia University – The Center of Pattern Recognition and Machine Intelligence (CENPARMI) – Montreal, QC, Canada, Research Assistant, September 2009 – to date.

Concordia University – Department of Computer Science and Software Engineering – Montreal, QC, Canada, Teaching Assistant for Data Structures and Algorithms Course, January 2010 – December 2010.

Birzeit University – Department of Computer Science – Ramallah, Palestine, Instructor, February 2005 – August 2009.

Al-Quds Open University – Department of Computer Science – Ramallah and Jericho, Palestine, Instructor, October 2004 – July 2005.

University of Jordan – King Abdulla II School for Information Technology – Amman, Jordan, Teaching Assistant, October 2003 – August 2004.

United Nations World Food Program (WFP) – Amman, Jordan, IT Assistant, May 2003 – July 2003.

Palestine OnLine – Ramallah, Palestine, Web Developer, April 2000 – November 2002.

Professional Qualification

National Laboratory of Pattern Recognition (NLPR) Institute of Automation of Chinese Academy of Sciences – Research Studentship – Beijing, China, July 2011 – September 2011.

The São Paulo Advanced School of Computing – Seminars in Image Processing and Visualization – São Paulo, Brazil, July 12 2010 – July 17 2010.

Concordia University Center for Teaching and Learning Services (CTLS) – Seminars in University Teaching – Montreal, QC, Canada, September 2009 – December 2009.

The Abdus Salam International Centre for Theoretical Physics (ICTP) – School on Digital and Multimedia Communication Using Satalite Radio Links – Trieste, Italy, February 12 – March 2, 2001.

Galaxy – Course on Network Essentials and TCP/IP Protocol – Ramallah, Palestine, Summer 1999.

Programming Languages

C, C++, Java, Matlab, Prolog, Assembly, HTML, ASP, Pascal.

Last updated: August 14, 2011

Adaptive Methods for Robust Document Image Understanding

Ph.D. Thesis Overview, 15.08.2011

Iuliu Vasile Konya,
Fraunhofer IAIS, NetMedia dept./ University of Bonn, Germany
Advisor: Dr. Stefan Eickeler, Fraunhofer IAIS

Increasingly many libraries and publishers are starting to digitize their complete archives as part of their plan to offer access to the materials in electronic form in order to reach a wider audience. Consequently, vast amounts of printed documents are awaiting processing and all over the world new digitization projects are being initiated. A few well known examples Google's mass book digitization project, the Theseus project endorsed by the German Federal Ministry of Economics and Technology and the IMPACT project supported by the European Commission. The content providers for these projects feature renowned universities, such as Harvard, Stanford, Oxford, as well as major libraries, such as the New York Public Library, the German National Library and the French National Library. From the context of such large document collections, we will give a brief overview of some of the most significant obstacles in document image analysis (DIA) and describe the corresponding research questions to be addressed in the current dissertation.

Problem 1: **large document quality variations** - robust pre-processing algorithms are necessary for achieving an image quality which is acceptable/usable by the DIA engine. DIA algorithms have in general strict requirements regarding their input and their output quality is highly dependent on how well those preconditions are satisfied by the pre-processing step. In contrast, the scanning process is still partly performed by humans, making errors inevitable. A recent study on a sample of about 1,000,000 pages done by a major German scan service provider has shown that scanning errors caused by humans are more than twice as frequent as errors coming from all other causes taken together. At the same time the current trend goes towards full-color scans of documents in order to capture the most amount of potentially relevant information. While the intent is laudable, this greatly increases the burden of the automatic DIA systems which have to deal with much larger amounts of data. From a computer science point of view, this represents a significant growth of the search space (for the information of interest), making the exploration process exponentially more difficult ("curse of dimensionality").

Problem 2: **adaptive logical layout analysis** - logical layout analysis (LLA), i.e. the identification of higher-level logical structures (such as tables, articles, captions, footnotes) has received little attention from the research community. In contrast, geometric layout analysis or page segmentation, i.e. the process of dividing a document page into text and non-text areas has already been relatively well explored. A major factor contributing to the low number of published LLA algorithms is the difficulty of dealing with continuous layout changes. Even for the same publisher, the layout of its publications unavoidably changes (sometimes drastically) over time. This is especially visible when dealing with publications spanning over many decades or even centuries. The layout rules employed in producing the document needing to be processed are in general not known by the DIA system, and most of the time they are not known even by the publishers themselves (e.g. they were lost over time). Thus it is extremely challenging to

have the same algorithm consistently delivering good results over the whole range of scanned documents.

The **objective of the current dissertation** is to advance the state-of-the-art in the areas previously mentioned so as to allow a satisfactory (semi-)automatic processing of real-life large document collections featuring complex layouts. More specifically:

- Provide **automatic immediate feedback** about the quality of the scanned images. Low quality images may arise from defective/poorly calibrated equipment as well as human error (**quality assessment**).
- Based on a computed image quality and an available document type, attempt to enhance the document image by removing **paper degradations** and **improving text areas** for better page segmentation- and OCR results.
- **Skew detection** and correction working reliably without relying on any non-generic assumptions about the document layout. Document scans exhibiting wildly varying degrees of rotation are practically unavoidable in large document collections, because of different scanning techniques and human error.
- **Color reduction** techniques and **binarization** algorithms must preserve as much relevant information as possible in their mapping of the input data into a much lower dimensional space. At the same time they must be fast enough allow the processing of the available large data quantities in reasonable time.
- Provide a **general DIA framework**, capable of handling to different types of printed documents. Special focus will be put on newspapers and books, as two of the most widespread and information-rich types of documents.
- Develop **LLA algorithms** which are more easily adaptable to different document layouts and scripts.

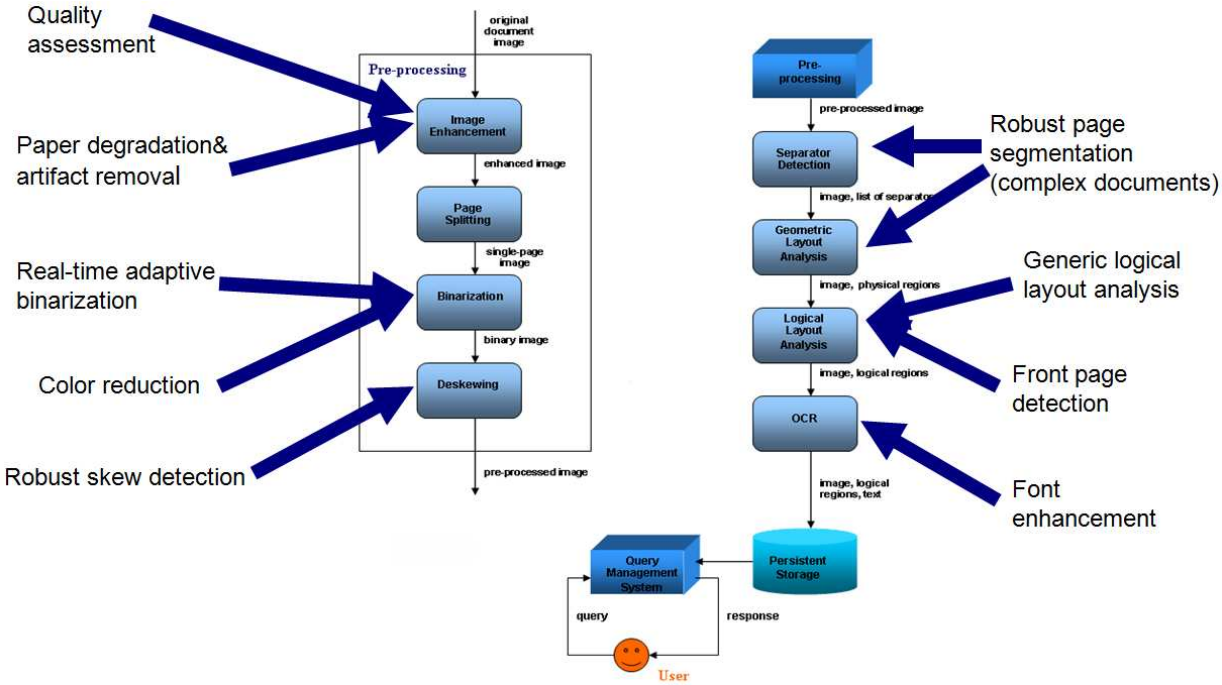


Figure 1: Generic document image analysis system with superimposed research areas

Current status of the Ph.D. thesis: As of the current date, the thesis is nearing its completion with an expected hand-in date sometime in November/December 2011. All aforementioned objectives have been reached and are partially described (or submitted for publication) in international journals (IJDAR, EURASIP), conferences (ICPR, ICDAR) and a published book chapter on trainable logical layout analysis. Work is currently focused on writing the thesis by assembling all research results into the final structure. The oral defense of the thesis is expected to take place in early 2012.

Publications (in chronological order)**:**

Konya, I.; Seibert, C.; Glahn, S.; Eickeler, S.: *A Robust Front Page Detection Algorithm for Large Periodical Collections.* In: Proc. Int'l Conf. on Pattern Recognition (ICPR), pp. 1-5, Dec 2008, Fraunhofer IAIS paper of the month (Dec 2008).

Konya, I.; Seibert, C.; Eickeler, S.; Glahn, S.: *Constant-Time Locally Optimal Adaptive Binarization*. In: Proc. Int'l Conf. on Document Analysis and Recognition (ICDAR), pp. 738-742, Jul 2009.

Konya, I.; Eickeler, S.; Seibert, C.: *Fast Seamless Skew and Orientation Detection in Document Images.* In: Proc. Int'l Conf. on Pattern Recognition (ICPR), pp. 1924-1928, Aug 2010, Fraunhofer IAIS paper of the month (Apr 2011).

Paaß, G.; Konya, I.: *Machine Learning for Document Structure Recognition.* In: Mehler, A.; Kühnbergerand, K.; Lobin, H.; Lüngen, H.; Storrer, A.; Witt, A. ed., *Modeling, Learning and Processing of Text Technological Data Structures*, Springer, Berlin/New York, 2011.

Liu, M.; Konya, I.; Nandzik, J.; Eickeler, S.; Flores-Herr, N.; Ndjiki-Nya P.: *A New Quality Assessment and Improvement System for Print Media*. In: EURASIP Journal on Advances in Signal Processing - Special Issue on Image and Video Quality Improvement Techniques for Emerging Applications, submitted May, 2011.

Konya, I.; Seibert, C.; Eickeler, S.: *Fraunhofer Newspaper Segmenter – A Modular Document Image Understanding System.* In: Int'l Journal on Document Analysis and Recognition (IJDAR), submitted July, 2011.

Konya, I.; Eickeler, S.; Seibert, C.: *Character enhancement for historical newspapers printed using hot metal typesetting.* In: Proc. Int'l Conf. on Document Analysis and Recognition (ICDAR), Sep 2011.

Kurbiel, T.; Konya, I.; Eickeler, S.: *A novel preprocessing method for hectography prints based on independent component analysis.* In: Proc. Int'l Conf. on Document Analysis and Recognition (ICDAR), Sep 2011.

Curriculum Vitae

Full Name	Iuliu Vasile KONYA
Address	Auf der Schleide 2, 53225 Bonn, Germany
E-mail	kiuliu@yahoo.com
Phone	+49(0)15205427914
Birth date	March 4, 1981
Citizenship	Romanian
Marital status	Single

Education

Jan2009 – Mar2012 University of Bonn/ Fraunhofer IAIS, Germany - **PhD candidate** in the area of document image analysis

Oct2004 – Nov2006 Bonn-Aachen International Center for Information Technology (B-IT), Bonn, Germany - **graduate student** in **MSc Program "Media Informatics"** (study language: English), GPA 1.1 (/1.0) - *summa cum laude*

Oct2003 – Jun2004 "Babeş-Bolyai" University, Cluj-Napoca, Romania - **graduate student** in **MSc Program "Intelligent Systems"** (study language: English), GPA: 10.00 (/10.00)

Oct1999 – Jun2003 "Babeş-Bolyai" University, Cluj-Napoca, Romania - **undergraduate student** in **Computer Science**, GPA: 9.52 (/10.00)

Honors

M.Sc. in **Media Informatics** – Nov 2006
1. M.Sc. paper entitled "Development of a Newspaper Image Understanding System" (written in **English**)
2. Practical project: implemented proposed system in C++ (Linux, IDE: Eclipse) and tested its performance on a real-life document dataset provided by Fraunhofer IAIS.
Overall grade: 1.0 (/1.0)

M.Sc. in **Intelligent Systems** – June 2004
1. M.Sc. paper entitled "Multi Expression Programming – Applications in Clustering and Digital Circuit Evolution" (written in **English**)
2. Practical project: developed C++ applications (IDE: Visual C++ 6.0) for testing, visualizing and validating the proposed evolutionary techniques for data clustering and digital circuit generation/optimization.
***Overall grade:* 10.00(/10.00)**

Bachelor of Science - June 2003
1. Licence paper entitled "Database protection".
2. Practical project: developed "AUROORA" (Analyzer for User Rights& Objects – ORAcle version), a visual manager for users, objects, roles and profiles for Oracle 8i databases). Language used: Object Pascal (IDE: Borland Delphi 6).
***Overall grade:* 9.85 (/10.00)**

Awards / Scholarships

- **ICDAR'09 page segmentation competition** *(Sep 2009)*, winner as leader of the Fraunhofer Newspaper Segmenter team: www.cse.salford.ac.uk/prima/papers/ICDAR2009_Competition.pdf
- **Computerworld Honors Program**, winner in section Media, Arts and Entertainment, as part of the NZZ Project "Archive 1780" team, *June 2006*.
- "Babeş-Bolyai" University **Academic Scholarship**:
  10.2003 - 02.2004 as MSc student.
  10.1999 - 10.2000, 03.2001 - 06.2003 as undergraduate student.
- **ACM International Collegiate Programming Contest** for Southeastern Europe (focus: data structures and algorithms), member of the "Babeş-Bolyai" University team - rank 13th (/47 universities), *October 2001*.

Areas of specialization

- Digital image& video processing
- Data structures and algorithms
- Evolutionary algorithms (genetic alg., multi expression programming)
- Object-oriented software design
- Concurrent and distributed programming

Work experience

Jan2009 – Jan2011	**Research Fellow (half-time)** in Fraunhofer IAIS, NetMedia dept., Germany. Continued work in document image analysis within industry and research projects (Theseus).
Jan2007 – Jan2009	**Research Fellow** in Fraunhofer IAIS, NetMedia dept., Germany. Lead developer of an in-house document image analysis library used in several large-scale (>1.000.000 pages) industry digitization projects (C++, Java, cross-platform Win/Linux/MacOS).
May2005 – Dec2006	**Student Research Assistant** in Fraunhofer IMK, Germany. Involved in the Neue Zürcher Zeitung (NZZ) Project "Archive 1780". Implemented tools for automatic newspaper page segmentation using state-of-the-art algorithms (C++, Linux); fuzzy string search/matching for automatic newspaper/book metadata extraction (Java).

Work interests

- Media analysis (focus on document, still image and video processing)
- Machine learning
- Parallel and distributed computing

Computer skills

- Programming languages/libraries: **C/C++** (very good), incl. STL, WinAPI, Boost, OpenCV; **Java** 2 SE (very good); **Pascal**/Object Pascal (good); MATLAB (good).
- Operating systems: **Windows** (very good), **Linux** (very good)
- Other technologies/environments: **XML** (good) – SAX, DOM, DTD, XML Schema (Java and C++); **Java RMI**, **servlets**, **JavaBeans**, Java Server Pages (**JSP**) (good); **relational databases** and **SQL** (good), incl. Oracle8i, MS Access; **LaTex** (good); **HTML** (average).

Segmentation and Labeling of Mixed-type Noisy Handwritten Documents

Jayant Kumar
PhD Student, Dept. of Computer Science,
University of Maryland College Park, USA
email: jayant@umiacs.umd.edu

Advisor: Dr. David Doermann
Director, Language and Media Processing Lab.
Institute of Advanced Computer Studies
University of Maryland College Park, USA
email: doermann@umiacs.umd.edu

My research interests lie in the areas of *Document Image Analysis and Retrieval*, *Computer Vision* and *Large Scale Machine Learning*. Over the span of four years, I have worked on problems like *handwritten text-line segmentation*, *document image enhancement*, *document image classification and labeling*, *scene-text detection and binarization*. As a part of DARPA's **MADCAT** project, I have been involved in the development of algorithms for processing of large collection of handwritten Arabic document images which may contain other types of content like printed text, logos, signatures, figures etc. The main challenge has been to handle the noise and free-style nature of these mixed documents. The details of the problems I am currently working, proposed methods and progress till date are as follows:

A. Handwritten Text-line Segmentation

Segmentation of text-lines can be a crucial step in many document processing tasks including document skew estimation, layout analysis and word/character recognition. For handwritten text, the problem is even more difficult due to its free style nature, character size variations and non-uniform spacings between components. To address these issues, I proposed an affinity-propagation based text-line segmentation method which combines local and global techniques [1]. Our method consists of four steps: *coarse text-line estimation*, *error detection and correction*, *touching component localization and separation* and *diacritic/accent component assignment*. We first estimate local orientation at each primary component to build a sparse similarity graph. We then, use a shortest-path algorithm to compute similarities between non-neighboring components. From this graph, we obtain coarse text-lines using two estimates obtained from *Affinity propagation* and *Breadth-first search*. In the next step, we use the distances along the nodes in the local-orientation graph to automatically detect touching and proximity errors. Expectation-maximization(EM) is then iteratively applied to split the touching lines. Finally, the error components are localized and cut to obtain an accurate estimate of text-lines [2]. The proposed method is fast and robust to non-uniform skew and character size variations, normally present in handwritten text lines. We evaluated our method using a pixel-matching criteria, and reported 98.6% accuracy on a dataset of 125 Arabic document images. We also presented a proximity analysis on datasets generated by artificially decreasing the spacings between text-lines to demonstrate the robustness of our approach. We showed improvement in accuracies using our correction method. Results on proximity images show that the method is effective for handling touching errors in handwritten document images. We also evaluated our method on ICDAR 2009 segmentation competition dataset(200 images) [8] and obtained a F-score of 97.8%. This dataset consists of handwritten document images of different scripts like English, French, German and Greek. I have implemented this method in C++ using DocLib library [7]. Currently, I am working on adapting the proposed method for scripts other than Arabic/Latin.

B. Document Image Enhancement

A recurring source of noise in handwritten documents is the underlying page rule-lines that interfere with the foreground text. These rule-lines significantly reduce the accuracy of subsequent document processing tasks, especially if the task is to be performed on a binary document image where the contrast between foreground text and underlying rule-lines is lost. I implemented a linear subspace-based rule-line removal technique in C++ for handwritten document images[3]. During a training phase, we incrementally construct linear subspaces representing horizontal and vertical lines using a set of rule-line images. During the testing phase, we compute the distance between features extracted from the test image and the previously

constructed subspaces. Pixels that belong to foreground text exhibit larger distances to the subspace and are not removed. The computation of central-moment features used in this approach was found to be very time-consuming for high-resolution document images. I proposed another approach for rule-line removal using integral-images which is much faster than our previous approach [4]. I also proposed a technique to sample points from a large training data sets for large scale Support Vector learning [Submitted].

C. Document Image Classification and Labeling

The labeling of large sets of document images for training/testing analysis systems or indexing images for search can be a very costly and time-consuming process. *Supervised learning* based approaches segment the image into different zones and attempt to classify each zone. This requires the training images to be fully annotated. Multiple instance learning (MIL) is a generalization of traditional supervised learning which relaxes the need for exact labels on training instances. Instead, the labels are required only for a set of instances known as *bags*. We applied MIL to the retrieval and localization of signatures and the retrieval of images containing machine-printed text, and showed that a gain of 15-20% in performance can be achieved over the supervised learning with weak-labeling [5]. The results are competitive to those obtained from supervised learning with fully-annotated data. Using our experiments on real-world datasets, we show that MIL is a good alternative when the training data has only document-level annotation. In the future, we would like to apply MIL for logo and stamp detection.

REFERENCES

[1] Jayant Kumar, W. Abd-Almageed, Le Kang and David Doermann. Handwritten Arabic Text Line Segmentation using Affinity Propagation. DAS 2010, Boston USA, 2010.

[2] Jayant Kumar, Le Kang, David Doermann and Wael Abd-Almageed. Segmentation of Handwritten Textlines in Presence of Touching Components, Intl. Conf. on Document Analysis and Recognition (ICDAR 11), Beijing, 2011.

[3] W. Abd-Almageed, Jayant Kumar and David Doermann. Page Rule-Line Removal using Linear Subspaces in Monochromatic Handwritten Arabic Documents. ICDAR 2009, Barcelona, 2009.

[4] Jayant Kumar and David Doermann. Fast Rule-line Removal using Integral Images and Support Vector Machines. Intl. Conf. on Document Analysis and Recognition(ICDAR 11), Beijing, 2011

[5] Jayant Kumar, Jaishanker Pillai and David Doermann. Document Image Classification and Labeling using Multiple Instance Learning. Intl. Conf. on Document Analysis and Recognition(ICDAR 11), Beijing, 2011.

[6] J. Kumar, R. Prasad, H. Cao, W. Abd-Almageed, D. Doermann and P. Natarajan, *Shape Codebook based Handwritten and Machine Printed Text Zone Extraction*, DRR, vol:7874(06), pp. 1–8, 2011

[7] Stefan Jaeger, Guangyu Zhu, David Doermann, Kevin Chen and Summit Sampat, *DOCLIB: a Software Library for Document Processing*, International Conference on Document Recognition and Retrieval XIII, pp. 1-9, San Jose, CA, 2006

[8] B. Gatos, N. Stamatopoulos and G. Louloudis, *ICDAR2009 Handwriting Segmentation Contest*, Intl. Conf. on Document Analysis and Recognition, pp. 1393-1397, Barcelona, Spain, 2009

Jayant Kumar

Contact Information	3348 A. V. Williams Department of Computer Science University of Maryland College Park, MD 20742 USA	*Voice:* (240) 601-5180 *Office:* (301) 405-1732 *E-mail:* jayant@umiacs.umd.edu *WWW:* www.umiacs.umd.edu/~jayant

Research Interests

Document Image Analysis & Retrieval, Computer Vision & Pattern Recognition, Large Scale Machine Learning, Natural Language Processing

Education

University of Maryland, College Park, MD USA

Ph.D. Candidate, Computer Science — Sept., 2008 - present

- Advisors : Dr. David Doermann and Prof. Larry Davis
- Expected graduation date: Dec., 2012

M.S., Computer Science — Sept., 2008 - Dec., 2010

- GPA: 3.9/4.0

R. V. College of Engineering, Bangalore, India

B.E., Computer Science and Engineering — Sept., 2002 - May, 2006

- Obtained *First Class with Distinction*
- Project: *Parallelization of Samba on NonStop/OSS*, NonStop Enterprise Divison, Hewlett Packard

Research Experience

Language and Media Processing Lab., MD USA
Research Assistant, Dr. David Doermann — August, 2008 - present
Topics: *Multilingual Automatic Document Classification Analysis and Translation* (DARPA's MADCAT Project), *Document Image Enhancement*, *Document Image Retrieval*

Fuji-Xerox Palo Alto Lab., Palo Alto, CA USA
Research Intern, Dr. Francine Chen, Multimedia Systems — May, 2011 - August, 2011
Topic: *Smart Document Capture Using Mobile Devices*

Raytheon BBN Technologies, Boston, MA USA
Image Processing Scientist(Intern), Dr. Huaigu Cao/Rohit Prasad — June, 2010 - Aug., 2010
Speech and Language Technologies group
Topic: *Segmentation and Classification of Mixed-type Noisy Document Images*

Indian Institute of Science, Bangalore, India
Research Assistant, Prof. A. G. Ramakrishnan — Feb., 2007 - May, 2008
Topics: *Online Handwriting Recognition of Indian Scripts*, *Automatic Field Extraction in Document Images*

Selected Publications

Jayant Kumar, Jaishanker Pillai and David Doermann, *Document Image Classification and Labeling using Multiple Instance Learning*, Intl. Conf. on Document Analysis and Recognition, 2011

Jayant Kumar, Le Kang, David Doermann and Wael Abd-Almageed, *Segmentation of Handwritten Textlines in Presence of Touching Components*,

Intl. Conf. on Document Analysis and Recognition, 2011

Jayant Kumar and David Doermann, *Fast Rule-line Removal using Integral Images and Support Vector Machines*, Intl. Conf. on Document Analysis and Recognition, 2011

Jayant Kumar, Rohit Prasad, Huiagu Cao, W. Abd-Almageed, David Doermann and Prem Natarajan, *Shape Codebook based Handwritten and Machine Printed Text Zone Extraction*, Document Recognition and Retrieval, 2011

Jayant Kumar, W. Abd-Almageed, Le Kang and David Doermann, *Handwritten Text Line Segmentation using Affinity Propagation*, Document Analysis System, 2010

Wael Abd-Almageed, **Jayant Kumar** and David Doermann, *Page Rule-Line Removal using Linear Subspaces in Monochromatic Handwritten Arabic Documents*, Intl. Conf. on Document Analysis and Recognition, 2009

T Kasar, **Jayant Kumar**, and A G Ramakrishnan, *Font and Background Color Independent Text Binarization*, Camera Based Document Analysis and Recognition (workshop of ICDAR), 2007

Arvind K R, **Jayant Kumar**, A G Ramakrishnan, *Line removal and Restoration of Handwritten strokes*, Intl. Conf. on Computational Intelligence and Multimedia Applications, 2007

Achievements

Obtained a grade of A+ in the courses: *Image Segmentation* and *Computational Linguistics II*

Qualified National Level Mathematics Olympiad (1999) with 84% marks

Qualified Indian Statistical Institute entrance exam(2002) for B. Maths with All India Rank - 18

Secured highest marks(100%) in Engineering Mathematics in first year (2002-03) of BE

Secured highest marks (98%) in University in subject - Concepts of C programming

Secured third rank in Computer Science department in first year of bachelors

Professional Services

Reviewer: ICPR 2010, ICDAR 2011, Computer Vision and Image Understanding

Member of working committee in *Summer School on Document Image Processing*, IISc Bangalore, June, 2008

Graduate Courses

Computer Vision: *Image Understanding*, *Image Segmentation*, *Object Recognition*
AI and Learning: *Machine Learning*, *Neural Computation*, *Computational Linguistics II*
Algorithms: *Design and Analysis of Computer Algorithms*, *Computational Geometry*, *Scientific Computing I*
Software Engineering: *Fundamentals of Software Testing*

References

Dr. David Doermann
Director, Language and Media Processing Lab.
University of Maryland College Park
email : doermann@umiacs.umd.edu

Dr. Francine Chen
Senior Research Scientist, Fuji-Xerox Palo Alto Lab
email : chen@fxpal.com

Mathematical Formula Recognition and Retrieval in PDF Documents

Xiaoyan Lin
Advisor: Zhi Tang
Institute of Computer Science and Technology
Peking University, Beijing, China

1. Abstract of Proposed Research

Mathematical formula is a common type of page component, especially in scientific documents. However, most of the mathematical formulas in documents are in poor electronic forms, such as images or unstructured symbols. Therefore, mathematical formulas in documents are difficult to extract, manipulate and retrieval. This is the most straightforward motivation I do research on mathematical formula recognition.

Nowadays an increasing number of scientific documents are available in PDF format, which can greatly facilitate document exchange and printing. With the tremendous popularity of PDF format, recognizing mathematical formulas in PDF documents becomes a new and important problem in document analysis field. The existing formula recognition methods have been widely discussed in image-based documents. To recognize mathematical formulas in PDF documents, one way is to transfer the PDF documents into images and then apply the current methods targeting at image documents. However, the converting and OCR procedures of this approach cost extra processing time and bring in character recognition errors. Moreover, the precise content information embedded in PDF documents is lost during the process. Character and layout information obtained from the PDF parser is much richer and more accurate than that acquired from OCR. In this sense, we can expect better results from PDF document recognition. This motivates me to do research on mathematical formula recognition targeting at PDF files.

The result of mathematical formula recognition can be utilized to make the formulas in documents accessible and searchable. For users to benefit from the mathematical resources, they need to search effectively not only by text but also by formulas. For instance, students who study mathematics want to search for information (e.g. name, author, background, etc) of the formula which they are unfamiliar with. For scientists, they may want to search for information of a new mathematical formula which he/she has never seen before. However, mathematical expressions are objects with complex structures and distinct symbols, therefore, users can’t search for relevant literatures centering on mathematical formulas in current search engines which are generally text based, e.g. Google or Yahoo!. In recent years, some researchers were addressing on the mathematical formula retrieval. However, some obstacles still exist in this area, such as, building user-friendly interface for querying formulas, approximate matching between formulas, and sub-structure matching between formulas, etc. Few of the current mathematical formula retrieval technologies can be used in practical applications. Therefore, my further work will focus on mathematical formula retrieval.

My thesis research aims at developing effective approaches to recognize mathematical formulas in PDF documents and facilitate mathematical formulas retrieval. To accomplish this goal, I will work on the following sub-tasks:

1. Extract the mathematical symbols information from PDF documents. Match the content streams (text, image, and graph) parsed from PDF documents into math symbols, and obtain the information (bounding box, baseline and font, etc) of the mathematical elements.
2. Mathematical formulas identification in PDF documents. The existing formula identification methods focus on image documents and most the existing methods are rule-based. Mathematical formulas, especially the embedded formulas, are still detected at accuracy too low to be used in practical applications. To improve the performance of formula identification, we will try to utilize the precise information of PDF and apply the Support Vector Machine (SVM) classification techniques to identify mathematical formulas in PDF documents.
3. Mathematical formula structure analysis in PDF documents. Similar to mathematical formulas identification, most of the existing formula structure analysis approaches are designed for image documents. They heavily depend on character recognition results of the OCR system, in which recognition errors are inevitable. We will try to adopt the existing methods with alteration in PDF documents and overcome the problems in traditional methods through fully utilizing the character and layout information in PDF files.
4. Build a ground-truth dataset for formula recognition in PDF documents. As formula recognition in PDF document has become a new and important research field, a need has arisen for a ground-truth dataset which can be used to evaluate the performances of different algorithms for formula recognition in PDF documents. As far as we know, there is no available PDF document dataset for mathematical formula recognition. Therefore, we plan to build such a ground-truth dataset.
5. Design and implement a mathematical formula retrieval system which will be able to query by mathematical formulas, support approximate matching and sub-structure matching between formulas. Current mathematical formula retrieval methods and systems are mostly based on text information retrieval technology which would ignore the structure information of the formulas. For the lack of the structure information, these methods are difficult to support approximate matching and sub-structure matching between formulas. To overcome these problems, we would like to investigate the indexing and relevant ranking algorithm considering the structure of mathematical formulas.

2. Progress

1. Mathematical character recognition

Content stream, including text, graph, and image objects, can be successfully parsed from the PDF documents through a tool developed by our research group. In the PDF content stream, a mathematical expression element may be composed of several different types of objects (e.g.,

text, image, graph). Therefore, the content stream extracted from the PDF document cannot be directly used as the logical mathematical expression elements. We have proposed a preprocessing step to focus on this problem. Matching from various objects in content stream to mathematical elements (e.g. root symbol, fraction, and vertical delimiter, etc) has been implemented in the preprocessing step. Precise attributes of the mathematical symbols like bounding box, baseline and font information are also available.

2. Mathematical formula identification

We have proposed a hybrid method by combining rule-based and SVM-based methods to detect isolated mathematical expressions in PDF documents. Rule-based method is applied to extract embedded mathematical formulas. In our method, various features of formulas, including geometric layout, character and context content, are used to adapt to a wide range of formula types. Experimental results show satisfactory performance of the proposed method. This work is the basis of the paper accepted by ICDAR 2011 conference.

3. Structure analysis of mathematical formula

The existing baseline structure analysis algorithm has been adapted to analyze mathematical formulas structure in PDF documents. The parse trees representing the layout structures of the math symbols are created and exported into MathML (in Presentation Markup). By utilizing the rich and precise character information extracted from PDF documents, some problems of baseline analysis in image documents can be avoided, e.g. threshold setting and math symbol recognition, etc.

3. Challenging problems

To achieve my research goals for mathematical formulas recognition and retrieval in PDF documents, the most challenging problems are concluded as follows:

1. PDF documents are generated through different tools. The objects used to render the mathematical expressions vary in the different types and versions of PDF generation programs. Hence, it is challenging to build a parser for recognizing mathematical symbols in various versions of PDF documents.
2. The embedded formula is more difficult to identify than isolated formulas, because the embedded formulas are generally short expressions, which are difficult to discriminate from ordinary text.
3. Semantic analysis for formulas benefits mathematical formula understanding and retrieval, but remains challenging now. Current structures analysis algorithms are mostly focused on the layout structure analysis of mathematical formula, whereas effective semantic analysis (or logical structure analysis) algorithms are few.
4. Significant problem in mathematical formula retrieval should be designing indexing and ranking algorithms which consider the structure of mathematical expressions and support approximate and sub-structure matching between formulas.

Curriculum Vitae

PERSONAL INFORMATION

Name:	Xiaoyan Lin
Email:	linxiaoyan@icst.pku.edu.cn
Tel:	(86)10- 1371 8837 024 (86)10- 8252 9613
Address:	Institute of Computer Science & Technology, Peking University, Beijing, China. 100871.
Citizenship:	China

RESEARCH INTERESTS

Machine learning; Pattern recognition; Document recognition and retrieval; Math recognition and retrieval

EDUCATION

2009 – Present Ph.D. candidate (Expected to graduate in 2014), Peking University, Institute of Computer Science & Technology, Beijing, China.

Research on mathematical formula recognition and retrieval in PDF documents.

2005 – 2009 Bachelor of Computer & Science, Beijing Normal University, Institute of Information Science & Technology, Beijing, China

SCHOLARSHIPS

2009 Wang Xuan Scholarship, First Prize, Institute of Computer Science & Technology, Peking University

2006 – 2008 Professional Scholarship, First Prize (for each academic year) Beijing Normal University

AWARDS & HONORS

2010 Academic Excellent Award of Peking University

2010 Outstanding Student Award, Institute of Computer Science & Technology, Peking University

2009 Outstanding Graduates Award, city of Beijing

2007 Bronze Medal, ACM International Collegiate Programming Contest, Asia Programming Contest, Changchun Site

WORK EXPERIENCE

2008 – 2009 Intern in DRM lab, Institute of Computer Science & Technology of Peking University, Beijing, China

Code and test the client software (Apabi Reader) of the EBook system.

Design and implement a Bluetooth technology based digital license obtaining method for electronic books.

2008 Venue Technical Volunteer of Beijing 2008 Olympic Games. Technical support Coordinator

Handle technical support requests in venue and assist technology and telecommunication engineers to exception handles.

Outstanding Volunteer Award of Beijing 2008 Olympic Games.

PUBLICATIONS

2011 Xiaoyan Lin, Liangcai Gao, Zhi Tang, Xiaofan Lin and Xuan Hu. "Mathematical Formula Identification in PDF Documents", To appear in the International Conference on Document Analysis and Recognition 2011.

HOBBIES

Tennis; Traveling; Movies

EFFICIENT INDEXING AND RETRIEVAL OF GRAPHS USING TECHNIQUES FOR EMBEDDING GRAPHS IN REAL-VALUED FEATURE SPACES

Muhammad Muzzamil Luqman [1,2]
mluqman@cvc.uab.es

Professor Jean-Yves Ramel
[1] Laboratoire d'Informatique
Université François Rabelais de Tours
37200 France
ramel@univ-tours.fr

Professor Josep Llados
[2] Computer Vision Center
Universitat Autonoma de Barcelona
08193 Spain
josep@cvc.uab.es

1. PROBLEM

My doctoral research is concerned with efficient indexing and retrieval of graphs, using techniques for embedding graphs in real-valued feature spaces and addresses the problem of lack of efficient computational tools for graph based representations of structural pattern recognition. The goal of the thesis is to enable these mature and powerful structural representations to employ the computational strengths offered by efficient state-of-the-art machine learning models of statistical pattern recognition. To achieve this I work on graph embedding, which is an emerging research domain and has attracted the attention of many researchers over the past half decade.

The application domain being used for investigation is graphical document elements, such as architectural drawings and electronic diagrams. The use of symbolic data structures has remained very popular in graphics recognition research community for last three decades for solving the problems of recognition and localization of the graphic entities in document images [1]. The increasing trend of mass digitization of document archives and the automation of office procedures has resulted in ever growing size of document image repositories. To cope with the user's demands of efficient querying and browsing mechanisms for the document image repositories, as a result the graphics recognition research community is now focusing on not only on designing efficient novel methods but also on porting the existing mature symbolic data structures based techniques to state-of-the-art efficient computational models of machine learning (i.e. the various clustering and classification techniques).

In short, my thesis research employs the representational power of structural pattern recognition together with the computational efficiency of statistical pattern recognition; for recognition, indexing and retrieval of graphic document images of architectural drawings and electronic diagrams.

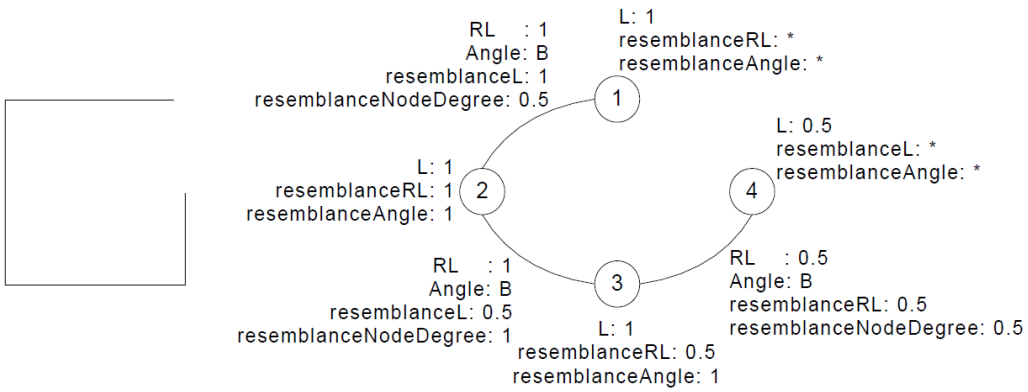
2. WORK PLAN AND TO-DATE PROGRESS

Targeting the above problem, the research is carried out in two complementary directions. These are briefly detailed in next subsections.

A. Graph embedding

A first direction of our work is to transform a graph into a feature vector. This immediately enables the graph based representations to use all the theory and models of the range of state-of-the-art efficient computational tools of machine learning.

Work realized: We have proposed an explicit graph embedding method for mapping a graph to a suitable point in vector space [2]. The proposed method exploits multilevel analysis of graph for embedding it into a feature vector. The feature vector contains graph level features (graph order and graph size), along-with structural level features (node degree and the homogeneity of nodes and edges in graph) and elementary level features (node attributes and edge attributes). We have used fuzzy overlapping trapezoidal intervals for minimizing the information loss while mapping from continuous graph space to discrete vector space. These intervals are employed for constructing fuzzy interval encoded histograms for embedding structural and elementary level features. Figure 1 illustrates an example of our explicit graph embedding method.



Feature vector: 4; 3; 2; 2; 1; 3; 0; 0; 1; 1; 0; 2; 1; 2; 0; 0; 3; 0; 2; 0; 0; 2; 1

Figure 1: A primitive shape, its attributed graph representation and the feature vector representation obtained after explicit graph embedding.

An important contribution of this work is that it enables graph based structural representations to access the state-of-the-art computational models of machine learning. The proposed methodology of first embedding graphs into feature vectors and then applying clustering/classification tools turns an impossible operation in original graph space into a realizable operation in feature vector space. Another important contribution of this work is that it can embed attributed graphs with numeric as well as symbolic attributes on both nodes and edges. To the best of our knowledge, there is no work in literature which permits to embed attributed graphs with numeric as well as symbolic attributes on both nodes and edges.

Future work: Future directions of work include the improvement of the graph embedding technique by incorporating more information on the topology/structure of graph.

B. Application of graph embedding for indexing and retrieval of graphs

A second direction of work is the application of the graph embedding technique to real problems of indexing and retrieval of relational (or symbolic) data structures. This is achieved by bag-of-subgraphs model; which is an extension of the original bag-of-words model.

Work realized: We have applied the graph embedding method for achieving subgraph spotting in a graph repository [3]. The goal is to retrieve a set of graphs, containing a given query subgraph, from a graph repository. Subgraph spotting is a very interesting research problem for various application domains and addresses the problems of indexing and retrieval of graph repositories. Our proposed method accomplishes subgraph spotting through graph embedding. We achieve automatic indexing of a graph repository during off-line (unsupervised) learning phase and achieve the subgraph spotting during on-line querying phase.

The proposed methodology does not rely on any domain specific details and offers a very general solution to the problem of subgraph spotting; indeed equally applicable to a wide range of application domains where the use of graph as a data structure is mandatory. Apart from incorporating learning abilities in structural representations (without requiring any labeled training set) and offering the ease of query by example (QBE) and the granularity of focused retrieval, the system does not impose any strict restrictions on the size of query subgraph.

Future work: Future directions of work include the building of multi-resolution index of graph repository by exploiting higher order cliques (>2) in graphs.

References

[1] D. Conte, P. Foggia, C. Sansone, and M. Vento, “Thirty years of graph matching in pattern recognition,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, no. 3, pp. 265–298, 2004.

[2] M. M. Luqman, J. Lladós, J-Y. Ramel, and T. Brouard, “A Fuzzy-Interval Based Approach for Explicit Graph Embedding,” in Lecture Notes in Computer Science, *Recognizing Patterns in Signals, Speech, Images and Videos*, 2010, vol. 6388, pp. 93-98.

[3] M. M. Luqman, J-Y. Ramel, J. Llados, and T. Brouard, “Subgraph Spotting through Explicit Graph Embedding: An Application to Content Spotting in Graphic Document Images,” in *Eleventh International Conference on Document Analysis and Recognition (ICDAR)*, 2011, p. to appear.

Muhammad Muzzamil LUQMAN

64, Avenue Jean Portalis, 37200 Tours, France.
Dob: 22nd April 1982, Peshawar Pakistan.
☎ : +33 2 47 36 14 43
✆ : +33 6 59 22 63 63
✆ : +34 6 88 41 68 57
✉ : mm.luqman@gmail.com
💻 : http://www.sites.google.com/site/mmluqman

PhD Candidate (Computer Science)

EDUCATION

2008 – to date	**Doctorate in Computer Science**	**Université François-Rabelais de Tours, France** **Universitat Autonoma de Barcelona, Spain** Title of thesis: Fuzzy Graph Embedding for Recognition, Indexing and Retrieval of Graphic Document Images. **Expected completion date: February 2012**
2007 – 2008	**Master-2 Research in Computer Science**	**Université François-Rabelais de Tours, France** Title of thesis: Recognition of complex graphic symbols in document images.
2000 – 2004	**Bachelor of Computer Science (Hons.)**	**Government College University Lahore, Pakistan** Title of thesis: Iris recognition.

ACADEMIC AWARDS & ACHIEVEMENTS

2008-2011	**Scholarship by Higher Education Commission of Pakistan for pursuing PhD in France** (http://www.hec.gov.pk/)
2007-2008	**Scholarship by Higher Education Commission of Pakistan for pursuing master in France** (http://www.hec.gov.pk/)
2006	**Academic roll of honor and medal for first position in Bachelor of Computer Science (hons.)** (http://www.gcu.edu.pk/)

PROFESSIONAL EXPERIENCE

2008 – to date	**Doctorate research**	**Laboratoire d'Informatique, Tours France** (http://www.li.univ-tours.fr) **Computer Vision Center, Barcelona, Spain** (http://www.cvc.uab.es/)
2006 – 2007	**Software development (LAMP)**	**Kolachi Advanced Technologies, Karachi Pakistan** (www.kolachi.net)
2005 – 2006	**Software programming (LAMP)**	**Hauka, Karachi Pakistan** (www.hauka.com)
2004 – 2005	**Software Programming (LAMP)**	**Dovemex Solutions, Karachi Pakistan** (www.dovemex.com)

RESEARCH INTERESTS

Pattern Recognition
Machine Learning
Document Image Analysis and Recognition
Graphics Recognition

PUBLICATIONS

- Subgraph Spotting through Explicit Graph Embedding: An Application to Content Spotting in Graphic Document Images. International Conference on Document Analysis and Recognition. 2011.
- Fuzzy-Interval Based Approach for Explicit Graph Embedding. Lecture Notes in Computer Science, Volume 6388, p. 93-98. 2010.
- A Content Spotting System For Line Drawing Graphic Document Images. International Conference on Pattern Recognition. p. 3420-3423. 2010.
- Fuzzy Intervals for Designing Structural Signature: An Application to Graphic Symbol Recognition. Lecture Notes in Computer Science, Volume 6020, p. 12-24. 2010.
- Graphic Symbol Recognition using Graph Based Signature and Bayesian Network Classifier. International Conference on Document Analysis and Recognition. p. 1325-1329. 2009.

LANGUAGE SKILLS

Urdu (fluent)
English (fluent)
French (proficient)

HOBBIES & INTERESTS

Traveling
Reading
Comparative religions
Meaning of life

Geometric-based Symbol Spotting, with Application to Symbol Retrieval in Document Image Databases

Nibal Nayef

Technical University Kaiserslautern, Germany

nnayef@iupr.com

Advisor: Prof. Dr. Thomas M. Breuel

Executive Summary

The ultimate goal of my research is the reliable and efficient symbol retrieval form large document image databases, in particular technical line drawings. Content-based image retrieval is becoming a necessary component in search engines whether regular or on mobile devices.

My approach to solve the retrieval problem has two main complementary directions. First, the recognition of the symbols, both isolated recognition and in-context spotting. Second, the off-line content analysis of line drawings, which makes us able to index the documents and the regions inside them for later fast retrieval. The two work directions constitute an important step towards the automatic understanding of line drawings.

For realizing my approach, I use geometric matching techniques for recognizing and spotting symbols. The use of shape features and geometry has shown to be very suitable for line drawings. As for content analysis, it is based on finding repeating symbols patterns and clustering them in an indexable symbol library. Finding the patterns is based on a statistically justified grouping method.

All my methods have been applied on standard datasets and have achieved significantly better results for spotting than the state-of-the-art approaches.

My future work will continue in the same directions. The short term goals include making the content analysis of line drawings more reliable, and the recognition faster. The long term goal is making symbol retrieval practical to be used in search engines.

Progress to date

1- Symbol recognition, both isolated and in-context spotting [1, 2]:

Matching two patterns is a key step for doing recognition. In isolated symbol recognition, people developed various kinds of descriptors: statistical, structural or combined, and then match the descriptors of the query against the descriptors of the library models. Some approaches use learning in different classifiers for recognition.

For symbol spotting, the common approach is first to identify regions of interest in a line drawing, and then using different descriptors to describe those regions and matching them with a query symbol. Sometimes a verification step is used after the initial matching to reduce the false positives.

In my work I have applied a geometric matching algorithm [3] for matching patterns. For symbol representation, pixels and/or vectorial primitives are used, and then the recognition is done via geometric matching under similarity transformations. This algorithm works with large

amounts of clutter, hence it can spot symbols in context – with interfering stokes – in addition to the isolated recognition.

This approach has the main advantage that there is no need to find regions of interest, which is a hard problem for line drawings, moreover, there is no need for developing complex feature descriptors or for training.

I incorporated those techniques in a symbol recognition/spotting system [1, 2]. The system has been applied on the datasets of GREC'05, GREC'11 symbol recognition contests and a dataset of real images of electronic circuits, and it performed significantly better than other statistical or structural methods.

2- Statistical segmentation of symbols – identifying patterns – [4]:

The biggest problem in symbol spotting is the well known dilemma: correct recognition-in-context requires good and reliable segmentation, but good and reliable segmentation needs information provided by the recognition process.

Rusinol and Llados investigated avoiding this problem by identifying sets of regions containing useful information. Then spotting can be done without having to fully segment or fully recognize the contents of line drawings [5].

This means, the local regions of interest in a line drawing need to be reliably and precisely located, so that the later steps of describing and indexing them would achieve better results.

Motivated by this problem, I have used statistical grouping for partitioning line drawings into shapes, those shapes represent meaningful parts of the symbols that constitute the line drawings. The usefulness of this grouping method is twofold, first, when used as segmentation method, it makes isolated recognition methods applicable for spotting symbols in context. Second, when used to identify regions of interest, it makes symbol spotting methods perform faster and more accurately.

The grouping method is based on finding salient convex groups of geometric primitives [6], followed by combining certain found convex groups together. In my work [4], I also have shown how such grouping can be used for symbol spotting.

Additionally, my groping method has a great potential as a content analysis step, this will be shown in the next section, where the grouping is used to identify symbol patterns.

3- Clustering repeating patterns – build a symbol library [7]:

Assuming we have solved the problems of isolated symbol recognition and symbol spotting, we still need to develop techniques for symbol retrieval in large databases. Symbol retrieval should be fast in order to be used for search in digital libraries for example.

In text and image search engines, people use content analysis techniques for large databases, for example a database of images is processed and its contents are represented by a relatively small indexable set of descriptors.

Inspired by the concepts in research in text and image retrieval, I developed a novel approach for representing a collection of line drawings as a library of symbols. This would have the direct effect of performing symbol retrieval efficiently.

This symbol library is a compact representation of the line drawings. Given a set of images of line drawings, I find the ***repeating*** patterns that exist in those images in a scale and rotation invariant way, those patterns are seen to correspond to meaningful parts of the symbols in the line drawings. The approach first identifies the candidate patterns in all images using the above

mentioned grouping method [4], and then it applies a geometric-based clustering algorithm, where the set of patterns from all images is divided into clusters of similar patterns. The clusters form a library of symbols [7]. This symbol library can be used in fast symbol retrieval.

Proposed future plans

- Speeding up the geometric matching search algorithm that I use as a basic matching step.
- Improving the grouping method by exploiting application domain information.
- Putting my developed techniques together to do practical symbol retrieval for a certain application.

References

[1] N. Nayef and T. M. Breuel, “Graphical symbol retrieval using a branch and bound algorithm,” in ICIP, 2010, pp. 2153– 2156.

[2] N. Nayef and T. M. Breuel, “On the Use of Geometric Matching for Both: Isolated Symbol Recognition and Symbol Spotting ”, in GREC, Accepted for publication – 2011.

[3] T. M. Breuel, “Implementation techniques for geometric branch-and-bound matching methods,” Computer Vision and Image Understanding (CVIU), vol. 90, no. 3, pp. 258–294, 2003.

[4] N. Nayef and T. M. Breuel, “Statistical grouping for segmenting symbols parts from line drawings, with application to symbol spotting,” in ICDAR, Accepted for publication – 2011.

[5] M. Rusinol and J. Llados, “Symbol Spotting in Digital Libraries: Focused Retrieval over Graphic-rich Document Collections”, Springer Publishing Company, Incorporated, first edition, 2010.

[6] D. W. Jacobs, “Robust and efficient detection of salient convex groups,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 18, no. 1, pp. 23–37, 1996.

[7] N. Nayef and T. M. Breuel, “Building a Symbol Library from Technical Drawings by Identifying Repeating Patterns ”, in GREC, Accepted for publication – 2011.

Short CV

Name: **Nibal Nayef**

E-mail: **nnayef@iupr.com**

website: **https://sites.google.com/a/iupr.com/nayef/**

Career Objectives

- Conducting world-class research in computer vision and pattern recognition.
- Pursuing an academic career as a professor and having an excellent level in teaching.

Education

Oct. 2008 – present (Expected date of graduation: Oct. 2012)
Technical University of Kaiserslautern (TU KL) - Germany
Computer Science Department - Image Understanding and Pattern Recognition Group (IUPR)
PhD student - Full DAAD scholarship

Oct. 2004 – Jan. 2007
Jordan University of Science and Technology (JUST) - Jordan
M.Sc. in computer engineering (Grade: Excellent, GPA 90.3 %) - Full DAAD scholarship
Thesis: A Vision Based System for Arabic Sign Language Recognition at Word / phrase Levels.

Sep. 1998 – Jun. 2003
Islamic University of Gaza (IUG) - Palestine
B.Sc. in computer engineering (Grade: Excellent (First rank in class), GPA 90.13 %)
Project: Distributed Electronic Health Care System.

Work Experience (Academic and professional)

Sep. 2007 – Jul. 2008
Community College of Applied Science and Technology (CCAST) - Palestine
Information Technology Department
Job title: Instructor
Role: Teaching practical computer science courses.

May 2007 – Aug. 2007
Company of Altariq Systems and Projects – Palestine
Job title: Software developer
Role: Developing desktop applications (Csharp and SQL server).

Sep. 2003- Sep. 2004
Islamic University of Gaza (IUG) – Palestine
Computer Engineering Department
Job title: Teaching and Research Assistant
Role: Teaching courses labs, co-preparing new courses.

Skills

- **Personal:**
    - Very good team work skills.
    - High Organizational skills.
    - Good social interaction across different cultures.
- **Technical:**
    - Programming in Python, Matlab, C++, Java, VisualStudio.Net (C#, C++), SQL.
    - Office software.

Research Interests

- Pattern recognition and machine learning.
- Graphics recognition, spotting and retrieval.
- Shape representation.
- Geometric matching.
- Hand gesture recognition / sign language recognition from images and videos.

Publications

[1] Nibal Nayef and Thomas M. Breuel, "Statistical Grouping For Segmenting Symbols Parts From Line Drawings, With Application To Symbol Spotting", ICDAR, 2011, accepted for publication.
[2] Nibal Nayef and Thomas M. Breuel, "Building a Symbol Library from Technical Drawings by Identifying Repeating Patterns", GREC 2011, accepted for publication.
[3] Nibal Nayef and Thomas M. Breuel, "On the Use of Geometric Matching for Both: Isolated Symbol Recognition and Symbol Spotting", GREC 2011, accepted for publication.
[4] Nibal Nayef and Thomas M. Breuel, “Graphical Symbol Retrieval Using a Branch and Bound Algorithm”, Int. Conf. on image processing (ICIP), 2010, pp.2153–2156.
[5] Mohammed Alrousan, Omar Aljarrah and Nibal Nayef, “Neural Networks Based Recognition System for Isolated Arabic Sign Language”, proceedings of the 3rd International Conference on Information Technology (ICIT), 2007.

Languages

- Arabic (mother tongue)
- English (excellent)
- German (good)

Hobbies and Interests

Hiking, Traveling, Reading, Movies, Photography, Learning foreign languages.

Copyright Protection of Manga Using Content-based Image Retrieval Methods

Weihan Sun
Osaka Prefecture University,
1-1 Gakuen-cho, Naka, Sakai, Osaka 599-8531, Japan.
Koichi Kise (Advisor)

Background

Manga is a kind of narrative artwork expressed by sequential comics typically printed in black-and-white. Objects in manga are mainly drawn with lines, but tones and word balloons elaborating the story lines are also employed. In its short history, manga has developed quickly to become one of the most popular types of image publication in the world. Especially in Japan, manga occupies a pivotal position in the publishing industry. From a report by the AJPEA (The all Japan Magazine and Book publishers' and Editors Association) in 2007, manga accounts for 36.7% of all publications [1] in Japan. Recently, the developing digital technique drums up another booming business: e-manga (digital manga magazine) which can be easily downloaded and viewed by digital terminals such as PCs and cell phones. However, the problem of illegal copies is threatening the manga industry. Therefore, there is a great interest in protecting the related copyrights.

In reality, the determination of what constitutes an illegal copy is a controversial problem and always depends on the judgment of professionals. However, a huge volume of manga publications require copyright protection, so that it is impossible for human beings to check every manga page manually. The purpose of our research is to apply computer techniques to detect candidate images for professionals' further judgment.

Problems

For copying manga, illegal users not only duplicate whole manga pages directly, but also focus on certain interesting parts to make partial copies, and apply them to their own drawings. Therefore, we must consider the detection of partial copies from unknown complex backgrounds. In addition, because of the simple compositions and abstract expressions applied in manga, as shown in Fig. 1, hand-drawn copies created by tracing and similar copies that infringe copyrights of manga characters can also be made, which challenge the detection. Because manga generally have minimal color information, it is difficult to embed a digital watermark without being perceived. For the illegal copy detection using image retrieval methods, commonly used features like SIFT [2], which have good performance for color photos lost their effectiveness, since hand-drawn copies and similar copies contain many detailed changes to the originals.

(a) Oringial part. (b) Hand-drawn partial copy. (c) Similar partial copy. (d) Similar partial copy.

Figure 1: Examples of hand-drawn copy and similar copy.

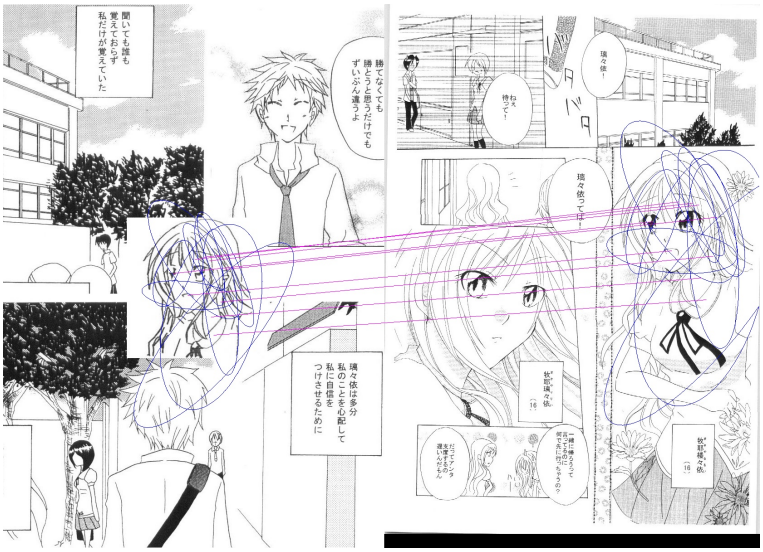
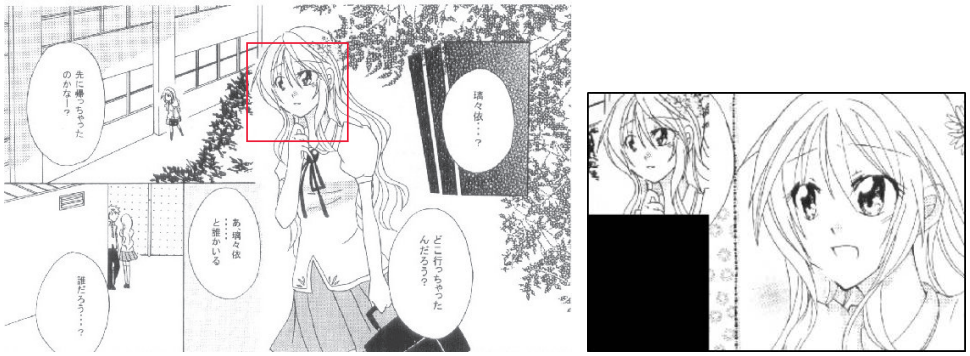


Figure 2: Example of hand-drawn partial copy detection from a complex background. (Left image is an illegal copy with a hand-drawn partial copy in the middle. Right part is the original image detected from the database of copyrighted manga pages. The ellipses are local feature regions and the matched pairs are connect by lines.)

Proposed solutions

To detect partial copies of line drawings: we have proposed applying the technique of content-based image retrieval: copyrighted images are collected in a database; suspicious images are treated as queries; the parts copied by suspicious images are reported as results. For printed and hand-drawn partial copies, we proposed a local feature matching method [3]. As shown in Fig. 2, the partial copy is marked and connected with its corresponding regions. The method are also effective for rotation and scale transformation within a certain range. In addition, a hash table method was applied to speed up the detection [4]. For similar partial copies, the regions of manga characters are detected and matched by using the features extracted from these regions [5]. As shown in Fig. 3, the character in the manga page was detected (Fig. 3(a)) and matched with a similar character in the database (Fig. 3(b)). We also proposed a model to increase the recognition accuracy of similar characters [6].

These results reported by our system can help human being search various kinds of illegal copies and offer the evidences in the same time. In addition, the method do not need to make any changes to the manga pages or depend on special printing. Therefore, it is easy to apply in manga publications nowadays.



(a) Manga character detection. (b) Manga character recognition.

Figure 3: Examples of similar partial copy detection from a complex background.

ICDAR2011

Recently, we did a research about the style of different titles of manga, which will be presented in ICDAR2011. We propose a bag-of-features method using visual words based on regions of interest (ROIs) for similar manga retrieval. For ROIs, face ROIs (regions around the faces of manga characters) and generic ROIs are applied. From the experimental results, we can see that despite of the complexity of manga, there still some discriminative patterns applied in a specific title of manga.

Expected target

Although the previous methods have the effectiveness for detecting various kinds of illegal copies, the database size will grow with the increase of database images. It can be used for a database with about 10 thousand manga pages as we applied in our experiments, but the number is far from the manga pages required copyright protection. According to a report by AJPEA [1], in Japan there are $10,965$ independent manga books published in 2006 (the number of manga pages for each independent book is over 200). With the increase of database, the problem of memory, detection time and accuracy will occur. Therefore, our work focus on the scalability of the detection method and try to test the method based on a database with 3 million manga pages (almost equals to the number of manga pages of independent books published for one year in Japan).

References

[1] "2007 Publication Index Annual Report," The all Japan Magazine and Book publishers' and Editors Association, the Research Institute for Publications, 2007.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[3] W. Sun and K. Kise, "Detecting Printed and Handwritten Partial Copies of Line Drawings Embedded in Complex Backgrounds," International Conference on Document Analysis and Recognition, pp. 909–919, 2009.

[4] W. Sun, K. Kise, "Speeding up the Detection of Line Drawings Using a Hash Table", in Proceedings of the 1nd China-Japan-Korea Joint Workshop on Pattern Recognition, vol.2, pp. 896-900, 2009.

[5] W. Sun and K. Kise, "Similar Partial Copy Detection of Line Drawings Using a Cascade Classifier and Feature Matching," International Workshop on Computational Forensics, pp. 121–132, 2010.

[6] W. Sun, K. Kise, "Similar Partial Copy Recognition for Line Drawings Using Concentric Multi-Region Histograms of Oriented Gradients", in Proceedings of the 12th IAPR Conference on Machine Vision Applications, pp. 71-74, 2011.

Curriculum Vitae

Personal Information

- Name : Weihan Sun
- Birthday : 1982/02/10
- Gender : Male
- Nationality : Chinese
- Affiliation : Intelligent Media Processing Laboratory, Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Japan.
- Phone : +81-72-254-9613 Fax: +81-72-254-8291
- Email : sunweihan@m.cs.osakafu-u.ac.jp

- Expected graduation date : 2013/03/31
- Thesis research : Illegal copy detection for copyright protection of line drawings.

Biography

Weihan Sun received BIM. degree in Economics Information Management from Tianjin University of Finance and Economics, Tianjin, China, in 2004. From 2004 to 2007, he worked as a software engineer at Tianjin Applo Info Tech Co., China. In 2007, he joined Intelligent Media Processing Laboratory of Osaka Prefecture University as a research student. He received M.E. degree in computer engineering from Osaka Prefecture University, Japan, in 2010, and studying as a Ph.D student in the same University. His research interests include image processing, pattern recognition, image retrieval, copyright protection.

Publications

- W. Sun, K. Kise, “Partial copy detection for copyright protection of line drawings”, Trans. IEICE, Vol.J93-D No.6 pp.909-919, 2009. (in Japanese)
- W. Sun, K. Kise, “Detecting Printed and Handwritten Partial Copies of Line Drawings Detecting Printed and Handwritten Partial Copies of Line Drawings”, in Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), pp. 341-345, 2009.

- W. Sun, K. Kise, “Speeding up the Detection of Line Drawings Using a Hash Table”, in Proceedings of the 1nd China-Japan-Korea Joint Workshop on Pattern Recognition (CJKPR2009), vol.2, pp. 896-900, 2009.
- W. Sun, K. Kise, “A Method for Copyright Protection of Line Drawings”, in Proceedings of the 2009 International Conference on Multimedia, Information Technology and its applications (MITA2009), 167-168, 2009.
- W. Sun, K. Kise, “Copyright Protection of Line Drawings: Copyrighted Part Detection Using Cascade Classifiers”, in Proceedings of the 2nd China-Japan-Korea Joint Workshop on Pattern Recognition (CJKPR2010), pp. 9-10, 2010.
- W. Sun, K. Kise, “Similar Partial Copy Detection of Line Drawings Using a Cascade Classifier and Feature Matching”, in International Workshop on Computational Forensics (IWCF2010), pp. 121-132, 2010.
- W. Sun, K. Kise, “Similar Partial Copy Recognition for Line Drawings Using Concentric Multi-Region Histograms of Oriented Gradients”, in Proceedings of the 12th IAPR Conference on Machine Vision Applications (MVA2011), pp. 71-74, 2011.
- W. Sun, K. Kise, “Similar Manga Retrieval Using Visual Vocabulary Based on Regions of Interest”, in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), 2011.

Document image quality evaluation

Rabeux Vincent
Advisors : J.P Domenger, N. Journet
rabeux@labri.fr, domenger@labri.fr, journet@labri.fr

1 Research statement

My PhD. takes place in the research context of quality evaluation of very old document images. In scanning and indexing document chains, quality evaluation of a digitized document is a concerning issue. One could need to evaluate the image quality in order to : readjust the scanner parameters, predict the **O**ptical **C**haracter **R**ecognition (**OCR**) error rate, annotate a document image of quality meta-datas. Studies [2, 3, 4, 1] were made on quality image evaluation in which metrics on characters degradations were introduced. These studies were made on modern documents suffering from holes in characters, broken and touching characters, and background speckle noise. Image quality evaluation did not account for ancients document images. The goal of my PhD is to be able to evaluate the quality of ancients documents ancients documents by analyzing their specific defects such as the bleed through defect for example (the verso's ink is visible on the recto side of a page).

1.1 Proposed Plan

My PhD can be divided in several major steps :

- At first we need to define what are the main defects that affects the overall document quality. Indeed, there is so many kinds of defects that it is impossible to analyze them all. We first have to select defects that will affect post treatment algorithm such as the OCR. Even in the context of human quality perception defects do not have the same influence to the human eye.

- Then, each selected defect will have to be segmented and analyzed using various algorithms and technics. The main goal of this step is to provide several measures, meta-data characterizing the document quality.
- At last, we will create prototypes using these quality meta-data in order to predict different usages such as the OCR error rate or human eye readability.

1.2 Progress to date

Up to now, I have been working on mainly the bleed-through defect. Bleed-through is a defect from which a lot of ancient documents suffer and can be characterized by the possibility to see the verso's ink through the recto side of a page. The bleed-through defect can be explained by two combined reasons : the light reflected by the scanner backing, and the ink of the verso side which is diffusing into the paper. The analysis of the bleed-through defect can be divided in two mains issues : the bleed-through pixels identification and the definition of measures characterizing at their best the quantity of bleed-through on a page.

In order to precisely locate the bleed-through pixels in a page, the only useable information is the verso's ink pixels their selves. But in order to use this information, the recto side and the verso side must be registered (aligned). Our work on the registration of a recto verso pair ended up with a new method that is as much accurate but a lot faster than the state of the art.

The second part of our work on bleed-through analysis enabled us to provide six different measures that characterize the bleed through of a digitized document on several aspects such as the intensity, the location and the quantity of bleed-through components. This six different measures where tested by creating an OCR error rate prediction model. This OCR prediction model is very accurate on simple structured documents where the main defect is bleed-through.

References

[1] M Cannon, P Kelly, and S Iyengar. An automated system for numerically rating document image quality. *Proceedings 1997 SPIE Conference on Electronic Imaging*, Jan 1997.

[2] Michael Cannon, Judith Hochberg, and Patrick Kelly. Quality assessment and restoration of typewritten document images. *International Journal on Document Analysis and Recognition*, 2(2):80–89, 1999.

[3] L Junichi, J Kanai, TA Nartker, and Juan Gonzalez. Prediction of ocr accuracy. *Symposium on Document Analysis and Information Retrieval (SDAIR)*, Jan 1995.

[4] S Rice, J Kanai, and T Nartker. An evaluation of ocr accuracy. *Information Science Research Institute Research Institute*, Jan 1993.

2 Resume

Vincent Rabeux

☎ 0698861020
✉ vincent.rabeux@gmail.com
202 avenue de Thouars, Bat. B5, appt. 142
33400 Talence
Age : 25 ans

PhD. student in computer science

ICDAR Doctoral Consortium 2011

Studies :

2009 – 2012	**PhD.** - *Quality evaluation of digitized documents*, theme : Image Analysis and Structuration (SAI) -LaBRI
2007 – 2009	**Master** - Software engineering specialized in project management, with honors (score > 80%) - Université Bordeaux 1
2006 – 2007	**Bachelor degree in computer science** - with honors (score > 70%) - Université Bordeaux 1
2003 – 2006	**DUT in computer science** - rank 12/80 - Université Bordeaux 1
2003	**Baccalauréat**, Equivalent High school diploma in sciences - HoChiMinhCity - Vietnam

Experiences:

2006 | 3 month	**Training period :**	**LaBRI** - Image and sound – Development of a (C++) library for reading, writing and analysis High Definition images. – Installation of a global visualization and acquisition process with HD cameras.
2007 | 1 month	**Technical writer :**	**LaBRI** - Visualization Team – User and developer documentation (english) of the software Tulip. (Tulip is a well known software for graph visualization.)
2009 | 6month	**Training period :**	**Capgemini** (Software engineer in Java EE) - Continuous integration, performance and unit tests. - Java, JEE and Struts programming.
2010 | 1 year	**Teaching :**	Object Oriented programming in C++ 4 hours a week to second year students. Supervision of the associated project.

Skills :

Computer Science Theory :	Graph algorithms Calculation models Natural Language Analysis Language Theory Formal Methods	
Programming :	C/C++ (QT, GTK, Visual C++)	Advanced
	C# (Framework Mono 2.0 et .NET)	School Level
	Java (JNI, JUnit)	Mastered
	Web : PHP (Jelix), JavaEE, Silverlight	Mastered
	Objective-C (iPhone, Mac OS X)	Mastered
Image Analysis :	OpenGL, ImageMagick, OpenCV, GIRL	
Operating Systems :	Linux ; Windows XP ; Mac OS X	
Languages :	English	Advanced
	Spanish	School Level
	Vietnamese (Speaking)	Beginner

Interests :

Programming :	Since I started programming in high school a created a lot of personal projects in PHP, Java EE and C++ (QT). I am also very interested in technologies and frameworks such as Qt (my every day tool), Jelix (http ://jelix.org/ a PHP framework in which i participated before my PhD) and Apple and Linux technologies.
Video Editing :	I created a lot of promotional videos for associations and hotels.
South-est Asia :	I lived 8 years in Vietnam where i developed a real interest for asian culture.

Research on Part-Based Method of Character Recognition

Wang Song

Kyushu University, Japan
Advisor: Prof. Seiichi Uchida

Introduction

Part-based methods have been proposed for object recognition. In those methods, a query image is first decomposed into parts, each of which will be described by a description method as a descriptor. The recognition result is determined by the comparison between the descriptors. This research aims to use the part-based method in handwritten character recognition. Most researchers believe that global features are essential for representing characters, but in the part-based method the global structure information is totally discarded. Without usage of global structure, the part-based method is supposed to be robust against deformation of the character. However, there are also several problems with the part-based method, for example, the low running speed and recognition rate. Therefore some research should be done in order to improve the part-based method of character recognition.

Problems

1) How should the part-based method be applied to character recognition?
2) How to improve the recognition rate and running speed?
3) With the number of character classes increases, can we still expect high recognition rate of part-based method?
4) Does the part-based method perform better than the other character recognition methods with severely deformed characters?
5) How to build mathematical explanation for part-based method?

Proposed Plan

1) First, several part-based methods should be designed and evaluated by experiments. The MNIST is selected as the test bed. According to [1], the part-based method of simple structure achieved about 93% recognition rate. In their experiments, for each category, first 1,000 samples of MINIST training set were used as the training set, and next 1,000 samples were used as test set. However, this recognition rate is lower than most of the

non-part-based methods. We will use more complicated structure of part-based method to improve the performance of part-based method.

2) Second, the SURF is used as the description method of image part. In SURF, each part is descripted by a 128-dimension vector, and this vector is called a keypoint. The similarity of keypoints is measured by Euclidean distance. In [1], since about 60 keypoints are extracted form a sample image in average, the database of the part-based method is 60 times the size of the database of non-part-based method. For each query sample, in part-based method, because there are 60 query keypoints, the recognition process should be done 60 times. Consequently, the time consumption of part-based method is 3,600 times of the time consumption of non-part-based method. From the observing of the distribution of keypoints in 128-dimension space, some improvement may be made.
3) The third step is to apply the part-based method to test sets with more classes (letters, Chinese characters, etc.). Also, some deformed character database may be used to test the performance.
4) The forth step is to use more description methods in part-based recognition.

Progress to Date

Up to now, three part-based methods have been compared [3]: first one is from [1] and called single voting; second one is a new proposed method called multiple voting and it can be seen as an extended version of the single voting; the third one is from [2] called class distance. The difference of the three methods is the voting process. In part-based method the voting process determines the class that the query image belongs to. The single voting has the simplest voting process in which each reference keypoint stands for one single vote; the multiple voting has more complicate voting process, in which each reference keypoint stands for multiple votes; the class distance has the most complicated voting process, in which each vote is represented by distances of different classes. The more complicated the keypoint is, the more information the vote contains. Two groups of experiments which used different size of training sets were done and the results are shown in table 1. From the results of first row we can see that the more information the vote contains the higher recognition rate the method has; in the meantime, from the second row of the table we can find that the multiple voting has the robustness against the decrease of the size of training set.

Table 1 Recognition rate of three methods (%)

	Size of training set	Single Voting	Multiple voting	Class distance
Recognition Rate	1000/class	93.57	94.92	97.91
	50/class	86.11	93.40	92.80

Another trial was made to observe the distribution of parts [4]. The part-based method in [1] employed the nearest neighbor method to recognize each part of the image. In detail, each query part (or to say keypoint) from query image will find its nearest neighbor keypoint in the database by calculating the Euclidean distance, and the class of its nearest neighbor keypoint will be the final recognition result of this query keypoint. Some of the reference keypoints in the database always generate the correct result if they are selected as the nearest neighbor keypoint. In the meantime, some other keypoints in the database always generate the wrong result. If we select the "good" keypoints to create the database, we may find a way to reduce the size of the database to enhance the speed, or to improve the recognition rate. Table 2 shows some experiment results of selection strategies. Strategy 1 uses only the selected or unselected keypoints to create a new database; strategy 2 only disabled selected or unselected keypoints in the voting process, the disabled keypoints still can be selected as nearest neighbor but they cannot contribute votes. From the table we can see that the selected and unselected keypoints had obviously different performances. Although the recognition rate was not improved, the selection process may still help with the long runtime of the part-based method.

Table 2 Selection

	Database	Recognition rate
No selection	All keypoints	86.11%
Strategy 1	Selected keypoints	84.37%
	Unselected keypoints	52.92%
Strategy 2	Selected keypoints have votes	85.62%
	Unselected keypoints have votes	39.19%

Reference

1. S. Uchida and M. Liwicki, "Part-Based Recognition of Handwritten Characters," Proc. ICFHR, pp. 545-550, 2010.
2. O. Boiman, E. Shechtman, and M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," Proc. CVPR, 2008.
3. S. Wang, S. Uchida and M. Liwicki, "Comparative Study of Part-Based Handwritten Character Recognition Methods," ICDAR 2011.
4. S. Wang, S. Uchida and M. Liwicki, "Look Inside the World of Parts of Handwritten Characters," ICDAR 2011.

Mr. Wang, Song

Human Interface Laboratory, Department of Advanced Information Technology,
Faculty of Information Science and Electrical Engineering, Kyushu University, Japan
E-mail: wangsong@human.ait.kyushu-u.ac.jp

Education

2010-present

Kyushu University
Fukuoka, Japan
Ph.D, information science
Expected graduation date: October, 2013

2008-2010

Huazhong University of Science and Technology (HUST)
Wuhan, China
MS, computer science

2004-2008

Hebei University
Baoding, China
BS, physics

Publications

Look Inside the World of Parts of Handwritten Characters, accepted by ICDAR 2011.

Comparative Study of Part-Based Handwritten Character Recognition Methods, accepted by ICADR 2011

Present Research

The part-based method for handwritten character recognition

Past Projects

Query-by-humming MP3 search

Recognition of score table in test paper

Segmentation and Recognition of Touching Characters in Offline Unconstrained Chinese Handwriting

Liang Xu
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Advisor: Dr. Cheng-Lin Liu

1 Overview

Handwritten Chinese text recognition has been an active research area recently. There are usually isolated character, broken character, and touching (or connected) characters in a handwritten string. Recognizing touching characters is one of the most difficult steps in an OCR (Optical Character Recognition) system. Though several efforts have been made, it is still not well solved by the researchers. We propose methods and techniques to segment and recognize touching characters in offline Chinese handwriting, which is not very cursive and can be recognized by human beings. This research summary will include the following parts, which are the core of my dissertation work.

a) Preparation of touching characters database
b) Separation of touching pattern
c) Learning-based separation points filtering
d) Recognition of touching characters

2 Background

Chinese character is used by the largest population in the world. And Chinese handwriting recognition has a lot of applications, such as postal address reading, business form reading, bank check reading and so on. Touching characters are often met in these situations. Fig. 1 shows two examples of touching characters.

(a) (b) (c) (d)

Fig. 1: Touching characters and corresponding touching points. (a-b) Single-connected pattern, meaning “refer to”. (c-d) Double-connected pattern, meaning “repair”.

Compared to Latin-based scripts and numerals, Chinese character usually has a highly compound structure, made up of several components. Besides, no extra space exists between Chinese words. So it seems unlikely to segment Chinese string without using recognition. Currently, “over-segmenting strategy” is often applied to segment Chinese characters. In the pre-segmentation step, some patterns are over-segmented with the hope that the correct touching boundaries are contained in the cut positions.

Then the segments are grouped into characters by character recognition and contextual information. The string recognition performance highly relies on the correct separation of the touching pattern. However, the separation of touching characters is a great challenge due to the variability of touching structures in Chinese handwriting. Fig. 2 shows the flowchart of our proposed system for touching handwritten Chinese string recognition.

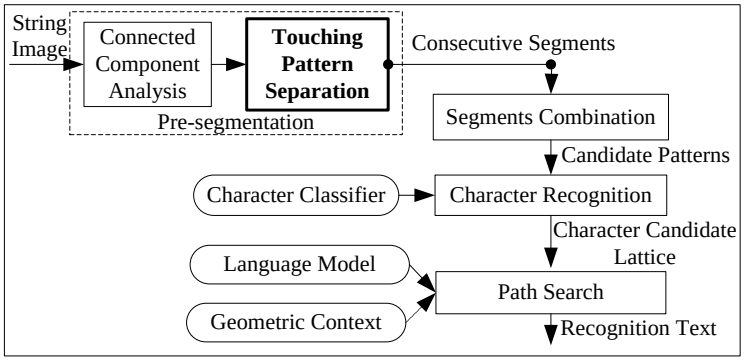


Fig. 2: The flowchart of touching handwritten string recognition system.

3 Motivation

We have briefly classified the touching string (two characters) into two main types: single-connected pattern and double-connected pattern, based on the number of touching points. Fig. 1 shows an example of both types. According to our examination, most of touching strings belong to single-connected pattern. For the remaining touching strings, most are double-connected pattern. As a result, we will firstly focus on the separation of single-connected pattern, which is also easier than double-connected pattern. Then, we will extend the separation algorithm to double-connected pattern.

In general, separation methods for touching pattern can be divided into three categories depending on the features utilized: foreground based methods (including projection analysis, contour analysis, and skeleton analysis), background based methods, and combined foreground and background methods. We have proposed two separation methods using foreground information (i.e., contour and skeleton), and we find that projection analysis is prone to result in errors through experiments. In the future, we plan to explore to use the background information together.

4 Finished work

Preparation of Touching String Database

We are using the Chinese handwritten text database from Chinese Academy of Sciences-Institute of Automation (CASIA-HWDB). This database contains about 5,000 handwritten documents with about 2,700 character classes. We have extracted about 10,000 touching strings from a part of this database according to its ground-truth. These touching strings will be used to evaluate our proposed separation

algorithms.

Separation of Single-connected Pattern

Two separation algorithms are proposed on this problem as following.

- ***Contour Analysis with DTW.*** Touching points usually lie at the corners on the contour of a pattern, except the ligature-type connection which can be treated in an extra process. It is therefore very likely to detect a corner point on the upper or lower contour around the touching location. A separation line can be formed by a corner point together with a corresponding point at the opposite contour side. This correspondence is implemented by DTW (Dynamic Time Warping) matching between upper and lower contour. Several heuristic rules are applied to remove some redundant separation lines. Compared with a typical method based on partial contour analysis, our method can detect more touching points, at a cost of more redundant separation points.
- ***Visibility-based Foreground Analysis.*** In order to reduce the redundancy of separation points, we attempt to combine contour analysis with skeleton analysis for separation of touching pattern. Skeleton analysis provides direct clues of strokes and fork points, which can help detect touching point precisely. Moreover, we find that most of touching points lie in or near the visible area (i.e., top and bottom profile of the touching pattern) according to our observation. A heuristic rule based on the visibility is applied to filter out redundant separating points efficiently.

5 Ongoing work

1) **Learning-based Separation Points Filtering.** Instead of heuristic rules, it is necessary to find a more principled approach to discriminate the correct separation point with redundant separation point. We will attempt to compute the structure features of a separating point from heuristic rules. Also, we will try to extract some local structure properties in the neighborhood of a separating point (e.g., background, foreground features). Then we will investigate traditional statistical methods and machine learning techniques (e.g., linear discriminate function (LDF) and SVM) to use these structure features in the sampled data.
2) **Separation of Double-connected Pattern.** Skeleton analysis and contour analysis (e.g., outer contour, hole information) will be adapted from the single-connected pattern separation.
3) **Recognition of Touching String.** Current over-segmenting strategy assumes that all the touching regions should be separated before recognition. However, for complicated touching case, it is very difficult to separate correctly without recognition, or it will bring a lot of redundancy. Can we add recognition information into the separation process? Can we utilize the separation information (e.g., confidence of a separation point) into the string recognition process, since the over-segmenting strategy isolates these two parts? We will try to find another feasible framework for string recognition (e.g., probabilistic framework using HMM).

Mailing Address:
Room 1020, Automation Building.
95 Zhongguancun East Road, Haidian District.
Beijing, P.R. China. 100190

Tel: +86-10-62632251
E-mail: lxu@nlpr.ia.ac.cn

Liang Xu

Research Interests

- Document analysis and recognition
- Digital image processing
- Statistical machine learning

Research Experience

July 2009 -- Present **Touching Character String Separation project**

- Propose a new algorithm based on contour analysis with DTW.
- Propose a new algorithm based on the combination of contour analysis and skeleton analysis.
  - Better performance than the previous one.

Sept. 2008 -- June 2009 **Binarization project**

- Propose a new algorithm based on a cascade of two linear classifiers to extract handwritten text on color document image, against printed characters and background noises.
  - Use the features of the RGB values of each pixel and average RGB values of its neighborhood, which are calculated rapidly from integral image.
  - The algorithm has been successfully used to collect a large offline handwriting database-CASIA-HWDB.

Education

Sept.2007-present Institute of Automation, Chinese Academy of Sciences Beijing, China

- Ph.D. candidate in the Pattern Analysis and Learning Group in the National Laboratory of Pattern Recognition (NLPR)
- Expected graduation date: July 2013

Sept. 2003-July 2007 Zhongshan (Sun Yat-sen) University Guangzhou, China

- B.E. in Electrical Engineering

Publications

1. **Liang Xu,** Fei Yin, Qiu-Feng Wang, Cheng-Lin Liu, "Touching Character Separation in Chinese Handwriting Using Visibility-Based Foreground Analysis," *ICDAR2011*, accepted.
2. **Liang Xu,** Fei Yin, Cheng-Lin Liu, "Touching Character Splitting of Chinese Handwriting using Contour Analysis and DTW," *Proc. 2010 Chinese Conference on Pattern Recognition (CCPR)*, Chongqing, China, Oct. 2010, pp.814-818.
3. **Liang Xu,** Fei Yin, Yi-Feng Pan, Cheng-Lin Liu, "An Efficient Color Document Binarization Method," *2009 PhD Candidates Academic Conference-Computer Vision & Artificial Intelligence.* Tianjin, China, Oct. 2009. (In Chinese)

Professional Activities

Reviewer for ICDAR2011, Asian Conference of Pattern Recognition (ACPR2011), International Workshop on Machine Learning for Signal Processing (MLSP 2011).
Volunteer for Chinese Conference of Pattern Recognition (CCPR 2007).

Skills

C/C++ developer
Matlab

Selected Courses

Pattern Recognition, Digital Image Analysis, Introduction to Algorithm

References

Dr. Cheng-Lin Liu (liucl@nlpr.ia.ac.cn)