

Doctoral Dissertations May 2014 - current

Off-campus UMass Amherst users: To download campus access dissertations, please use the following link to [log into our proxy server](#) with your UMass Amherst user name and password.

Non-UMass Amherst users: Please talk to your librarian about requesting this dissertation through interlibrary loan.

Dissertations that have an embargo placed on them will not be available to anyone until the embargo expires.

Efficient Representation and Matching of Texts and Images in Scanned Book Collections

[Download](#)

[Ismet Zeki Yalniz, University of Massachusetts - Amherst](#)

[Follow](#)

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

Year Degree Awarded
2014

First Advisor
R. Manmatha


Second Advisor
James Allan

Third Advisor
W. Bruce Croft

Keywords
Digital libraries, Information retrieval, Scanned book collections

Subject Categories
Computer Sciences

Abstract
Millions of books from public libraries and private collections have been scanned by various organizations in the last decade. The motivation is to preserve the written human heritage in electronic format for durable storage and efficient access. The information buried in these large book collections has always been of major interest for scholars from various

 **Included in**
[Computer Sciences Commons](#)



ncluded in
[nputer Sciences Commons](#)

[SHARE](#)

Enter search terms:

in this series

[Advanced Search](#)

 [Notify me via email or RSS](#)

[Browse](#)

[Collections](#)

[Disciplines](#)

[Authors](#)

[Author Corner](#)

[Author FAQ](#)

[Submit Dissertation](#)



disciplines. Several interesting research problems can be defined over large collections of scanned books given their corresponding optical character recognition (OCR) outputs. At the highest level, one can view the entire collection as a whole and discover interesting contextual relationships or linkages between the books. A more traditional approach is to consider each scanned book separately and perform information search and mining at the book level. Here we also show that one can view each book as a whole composed of chapters, sections, paragraphs, sentences, words or even characters positioned in a particular sequential order sharing the same global context. The information inherent in the entire context of the book is referred to as global information and it is demonstrated by addressing a number of research questions defined for scanned book collections.

The global sequence information is one of the different types of global information available in textual documents. It is useful for discovering content overlap and similarity across books. Each book has a specific flow of ideas and events which distinguishes it from other books. If this global order is changed, then the flow of events and consequently the story changes completely. This argument is true across document translations as well. Although the local order of words in a sentence might not be preserved after translation, sentences, paragraphs, sections and chapters are likely to follow the same global order. Otherwise the two texts are not considered to be translations of each other.

A global sequence alignment approach is therefore proposed to discover the contextual similarity between the books. The problem is that conventional sequence alignment algorithms are slow and not robust for book length documents especially with OCR errors, additional or missing content. Here we propose a general framework which can be used to efficiently align and compare the textual content of the books at various coarseness levels and even across languages. In a nut-shell, the framework uses the sequence of words which appear only once in the entire book (referred to as "the sequence of unique words") to represent the text. This representation is compact and it is highly descriptive of the content along with the global word sequence information. It is shown to be more accurate compared to the state of the art for efficiently i) detecting which books are partial duplicates in large scanned book collections (DUPNIQ), and, ii) finding which books are translations of each other without explicitly translating the entire texts using statistical machine translation approaches (TRANSNIQ).

Using the global order of unique words and their corresponding positions in the text, one can also generate the complete text alignment efficiently using a recursive approach. The Recursive Text Alignment Scheme (RETAS) is several orders of magnitude faster than the conventional sequence alignment approaches for long texts and it is later used for iii) the automatic evaluation of OCR accuracy of books given the OCR outputs and the corresponding electronic versions, iv) mapping the corresponding portions of the two books which are known to be partial duplicates, and finally it is generalized for v) aligning long noisy texts across languages (Recursive Translation Alignment - RTA).

Another example of the global information is that books are mostly printed in a single global font type. Here we demonstrate that the global font feature along with the letter sequence information can be used for facilitating and/or improving text search in noisy page images. There are two contributions in this area: (vi) an efficient word spotting framework for searching text in noisy document images, and, (vii) a state of the art dependence model approach to resolve arbitrary text queries using visual features. The effectiveness of these approaches is demonstrated for books printed in different scripts for which there is no OCR engine available or the recognition accuracy is low.

Recommended Citation

Yalniz, Ismet Zeki, "Efficient Representation and Matching of Texts and Images in Scanned Book Collections" (2014). *Doctoral Dissertations May 2014 - current*. Paper 43.
http://scholarworks.umass.edu/dissertations_2/43

This page is sponsored by the [University Libraries](#).

© 2009 [University of Massachusetts Amherst](#) • [Site Policies](#)