A Framework for Intelligent Navigation of Scanned Documents

Akash Dutta, Amit Kumar Das, Manas Hira
Department of Computer Science and Technology
Bengal Engineering and Science University, Shibpur, Howrah, India
E-mail: akashdutta1802@gmail.com, akash.dutta@students.becs.ac.in, {amit, manas}@cs.becs.ac.in

***Abstract*—Navigational facilities of the available e-document readers do not offer enough support for non-stereotype navigation. The present investigation aims to develop a framework for intelligent navigation of scanned documents through retrieval of within-document entities of desired types, e.g., drawings, equations, tables, etc. The framework has been implemented as a fully functional web service by integrating four modules, namely, e-book archive, coordinate extractor, database and Graphical User Interface. This navigation framework has also been implemented as a standalone application intended for personal computers and e-book readers.**

***Keywords*-within document retrieval; navigation system; digital document.**

I. Introduction

Typical weakness of the present day electronic document readers is their lack of flexibility in furnishing the navigation facilities. They are particularly limited to maneuvering with respect to page, chapter, index or embedded tags from the present position. Surprisingly, the framework of all the available document readers strongly suggests a serious deficiency in considering the need of intelligent and more human-like navigation of electronic documents. For example, an avid reader may be interested in viewing only those pages of an e-book which have colourful photographs, while others may be interested only in those pages containing tables or equations or graphs. Evidently, there is pressing need to incorporate intelligent navigation facilities which will address such non-stereotype navigational needs of the reader.

Digitized documents may be classified as electronic documents and scanned documents. Electronic documents are created using electronic input devices and the entire document is produced in a digital environment from scratch. They are indexed and tagged in structured vectorized formats. Consequently, identification and retrieval of document components such as, texts, diagrams, tables, equations etc. from structured documents like PDF and Post Script format are well known practices.

On the other hand, a large number of digitized documents are widely accessible as scanned image versions of their original hard copies, in which each page of the documents are converted into digitized image formats. These scanned documents are not structured and hence identification, extraction and retrieval of different document components require specialized approaches. In the last two decades, numerous research endavours [1]–[5] have been directed towards developing methodologies for solving the above mentioned problem and have been met with fair amount of success. However, construction of an intelligent navigation framework by utilizing the already developed methodologies is yet to be formulated. The present investigation envisages development of a framework for intelligent navigation of digital documents with an aim to retrieval of desired entities (e.g., text paragraphs, drawings, figures, photographs, tables, references, etc.) from within the scanned documents through a graphical interface.

II. Related Work

Various image processing methodologies have been formulated over the years for identifying different components of a scanned digitized document. Researchers have thrived on the identification, extraction [1] and retrieval [2] of different document components using document image analysis. Being a relatively exigent task, component extraction from complex background is recently encouraging various kind of approaches to deal with this problem. Chen [3] has proposed a knowledge-based approach for extraction and identification of text-lines from complex real-life background. Whereas, Sun et. al. [4] have insinuated a method using wavelet transform and support vector machines to contend with the same problem. Tong et. al. [5] have employed state-of-the-art edge detection and line identification algorithms to ascertain transmission lines in natural complex background. The background work discussed so far has been only used in certain modules of our proposed framework. However, no work on developing a framework such as the present one has been proposed yet.

III. Adopted Framework and Implementation

The framework has been implemented as a web service. Users of the web service are allowed to login, browse different e-books in the e-book archive, and navigate through the contents of an e-book of their choice using the navigation options incorporated in the system.

Primarily our web service has four different interacting modules which work as a single functional unit (see Fig. 1). The first module constitutes of the E-book Archive. It can be conceptualized as a Digital Library having only scanned

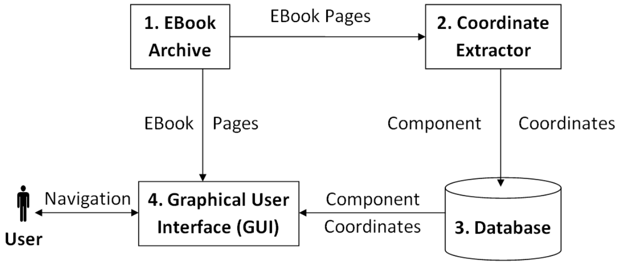
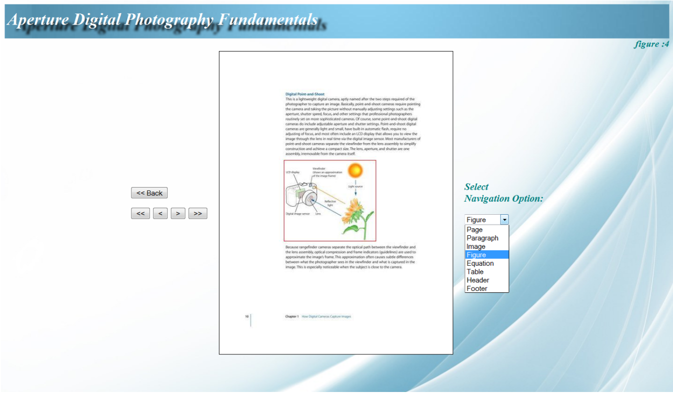


Figure 1. Functional block diagram of the proposed framework.


Aperture Digital Photography Fundamentals
figure :4
<< Back
<<
<
>
>>
Select Navigation Option:
Figure
Page
Paragraph
Image
Figure
Equation
Table
Header
Footer
LCD display
Viewfinder (shows an approximation of the image frame)
Light source
Reflective light
Digital image sensor
Lens
10
Chapter 1 How Digital Cameras Capture Images

Digital Point-and-Shoot

This is a lightweight digital camera, aptly named after the two steps required of the photographer to capture an image. Basically, point-and-shoot cameras require pointing the camera and taking the picture without manually adjusting settings such as the aperture, shutter speed, focus, and other settings that professional photographers routinely set on more sophisticated cameras. Of course, some point-and-shoot digital cameras do include adjustable aperture and shutter settings. Point-and-shoot digital cameras are generally light and small, have built-in automatic flash, require no adjusting of focus, and most often include an LCD display that allows you to view the image through the lens in real time via the digital image sensor. Most manufacturers of point-and-shoot cameras separate the viewfinder from the lens assembly to simplify construction and achieve a compact size. The lens, aperture, and shutter are one assembly, irremovable from the camera itself.

Because rangefinder cameras separate the optical path between the viewfinder and the lens assembly, optical compression and frame indicators (guidelines) are used to approximate the image's frame. This approximation often causes subtle differences between what the photographer sees in the viewfinder and what is captured in the image. This is especially noticeable when the subject is close to the camera.

Figure 2. Screen shot of the GUI displaying an e-book page with a red box bounding a figure when navigation option selected as ‘figure’.

PDF documents. PDF has been chosen only because of its abundant availability with respect to other e-book formats.

The second module, the Component Identifier and Coordinate Extractor, is purely a program code which takes as an input any e-book in its PDF, converts each of its pages into corresponding image format (JPEG), implements different image processing techniques to identify different components of an image (e.g., text paragraphs, equations, tables, line graphics, photographs, headers, footers, etc.) and finally extracts the coordinate information of respective components and stores them into corresponding database tables.

The third module in our web service is the database which stores the coordinate information for each component of each document of the archive as identified in the second module.

The web-based Graphical User Interface (GUI) module allows an interactive session with the user, wherein, depending on the user's choice, an e-book can be navigated component-wise.

The web-service has been implemented using JSP which interacts with the MySQL DBMS and loads the e-book pages from the e-book archive and navigates only to those pages containing the component of interest bounded in a red box. We have used state-of-the-art methods to segment each page of the e-books into different components and saved the coordinate information (i.e. left-top point coordinate of the bounding box and its height and width) into tables of a MySQL database. Fig. 2 shows a typical screen shot of the GUI of the implemented framework. The screen shot of the implemented framework shows that after choosing a specific e-book to navigate, the user can select any one of the component types from the right-hand side drop down menu and navigate through the e-book, viewing only the pages having that particular component. The user can use the next, previous, first and last buttons provided on the left hand side to browse through particular components. This navigation framework has also been implemented as a standalone application, specially designed for desktops, laptops, e-book readers etc. In its standalone version, the framework works on a single e-book as selected by a user. The user interface part of the application has been implemented using Java and with the help of embedded helper applications; the latter one accomplishes the component identification and coordinate extraction.

IV. Conclusion

The intelligent navigation framework has been developed with the sole intention of assisting the reader with the navigation tools which will give an experience of flexibility of document reading. The adapted framework, if integrated with the existing ones, may serve as a universal e-document reader with intelligent navigation. This framework may also be conceived and implemented as a cloud application for internet users.

References

[1] Yen-Lin Chen, Bing-Fei Wu, “A multi-plane approach for text segmentation of complex document images,” *Pattern Recognition*, vol. 42, pp. 1419-1444, 2009.

[2] A. J. Kass, “An interchange standard and system for navigation digital documents,” M.S. thesis, Dept. Electrical Eng. and Computer Sci., Massachusetts Inst. of Tech., MA, 1995.

[3] Chen Yen-Lin, “A knowledge-based approach for Textual Information Extraction from Mixed Text/Graphics Complex Document Images,” *IEEE Int. Conf. Syst. Man and Cybern.*, pp. 3270-3277, 2010.

[4] Sun Hongxing, Zhao Nannan and Xu Xinhe, “Extraction of Text under Complex Background Using Wavelet Transform and Support Vector Machine,” *IEEE Int. Conf. Mechatronics and Automation*, pp. 1493-1497, 2006.

[5] Tong Wei-Guo, Li Bao-Shu, Yuan Jin-Sha and Zhao Shu-Tao, “Transmission Line Extraction and Recognition from Natural Complex Background,” *IEEE Int. Conf. Mach. Learning and Cybern.*, pp. 2473-2477, 2009.