Quality Assurance in High Volume Document Digitization: A Survey

Xiaofan Lin
Digital Printing and Imaging Laboratory
HP Laboratories Palo Alto
HPL-2006-18
February 1, 2006*

quality assurance, document image analysis, OCR, digital library

Quality assurance (QA) plays a critical role in high volume document digitization projects by making sure that the specified quality standard is reached under cost and time constraints. This paper takes a systematic view on this issue by summarizing and abstracting related existing work: quality bottlenecks and technical solutions throughout the whole processing pipeline, including cataloging, capture, image analysis and recognition, and error cascading; various strategies to conduct cost-effective QA, such as combination of auto-QA and manual QA, batch QA, special QA user interface, and open source QA.

* Internal Accession Date Only
Published in the 2nd International Conference on Document Image Analysis for Libraries, 27-28 April 2006, Lyon, France
Approved for External Publication

Quality Assurance in High Volume Document Digitization: A Survey

Xiaofan Lin
Hewlett-Packard Laboratories
1501 Page Mill Rd MS 1203
Palo Alto, CA 94304, USA
E-mail: xiaofan.lin@hp.com

Abstract

Quality assurance (QA) plays a critical role in high volume document digitization projects by making sure that the specified quality standard is reached under cost and time constraints. This paper takes a systematic view on this issue by summarizing and abstracting related existing work: quality bottlenecks and technical solutions throughout the whole processing pipeline, including cataloging, capture, image analysis and recognition, and error cascading; various strategies to conduct cost-effective QA, such as combination of auto-QA and manual QA, batch QA, special QA user interface, and open source QA.

1. Introduction

An important channel of creating digital libraries is to digitize vast amount of existing paper documents. Throughout this paper, "digitize" is used in the broad sense and means the complete processing that covers capture, image analysis, text recognition, and so on. A number of document digitization projects have been in existence for a while [1][2][3][4][5]. Recent high-profile efforts, such as Google Print [6], Open Content Alliance [7], Internet Archive [8], and Amazon's "Search inside Book" [9] have really pushed mass book digitization (in terms of millions of books) to the attention of general public. After decades of research on document image analysis (DIA), we can find most underlying algorithms through commercial-off-the-shelf (COTS) software, open-source/public-domain software, complemented with necessary in-house development. Baird presented an extensive overview on various DIA techniques for creating digital libraries [10]. However, due to many factors, such as the imperfect nature of DIA, human errors, hardware limitations or failures, and software bugs, it is still challenging to maintain the required quality standard in high volume document digitization projects, especially with cost and time also in the equation. With a number of large scale digital library projects going on, the quality assurance (QA) issue is attracting more and more attention. Kelly et al. discussed the general QA procedures in digital library programs and their focus is on the necessary documentations, appropriate standards and best practices related to the software engineering aspects [11]. Yacoub concentrated on the QA processes of a particular document understanding system [12]. QA is also touched upon in a number of previous papers [1][3][4][13][14]. This paper takes a systematic view on the QA issues by summarizing and abstracting from related existing work. The purpose of this paper is to make the digitization system architects fully aware of the various QA issues and existing solutions. The DIA researchers can also take them into consideration when designing new algorithms. Section 2 presents the role of QA in high volume document digitization. Section 3 discusses the different QA problems and techniques throughout the whole processing pipeline, including cataloging, capture, image analysis and recognition. Section 4 generalizes various QA strategies to conduct cost-effective QA, such as combination of auto-QA and manual QA, batch QA, special QA user interface, and open source QA. Section 5 gives a summary.

2. Role of QA in high volume document digitization

In ISO 9000:2000 Standard [15], quality assurance is defined as "a set of activities whose purpose is to demonstrate that an entity meets all quality requirements." This general definition also applies to the QA in document digitization. However, document digitization has a couple of unique characteristics: First, the core DIA software is imperfect *by default* because the nature of DIA is to imitate human's cognitive capabilities, which are still unmatched by today's computer algorithms and systems. Conventional QA methods such as redundancy and failover may prevent DIA software from crashing, but not from making mistakes. Second, manual processing is a double-sided sword. On the one hand, human intervention is indispensable for the success of a document digitization project: Operators control the imaging device to capture paper documents and they also detect and/or correct mistakes made by DIA software in order to satisfy the set quality requirement. On the other hand, human intervention introduces extra errors, limits the system's throughput, and increases the project's cost. This is especially true when huge number of documents need to be digitized. Because of the above factors, QA is critical to the success of a high volume document digitization project and it strikes the subtle balance among the various factors that interplay in a document digitization project: quality requirement, schedule, budget, number and nature of documents, and available technologies (capture, image analysis and recognition, etc.)

3. QA throughout the digitization pipeline

Generally speaking, a document digitization pipeline can be divided into several steps: cataloging, image capture, image analysis and recognition. Post-digitization applications such as information retrieval, repurposing and publishing are another subject and will not be covered in this paper. Each step in the pipeline has its own quality problems and corresponding QA solutions, which are summarized in Table 1. The next subsections devote to individual steps.

Table 1: Quality problems and QA methods in each step of digitization

Stages	Quality problems	Problem types	Problem causes	QA methods	Refs
Cataloging	Incorrect metadata	Manual entry	Metadata not available in existing catalogs	Automatic database scan, manual correction	[19][20]
Image capture	Missing /duplicated pages	Human operation errors	Manual page flipping	1. Automatic analysis 2. Manual correction through UI	[4] [10] [21]
		Mechanic problems	Automatic page feeding or turning		
	Poor imaging quality	Uneven lighting, distortion, out of focus	Camera-based imaging	1. Calibration and auto-detection 2. Adaptive thresholding, auto-focus, perspective correction, etc.	[3][25] [26]
		Curling pages	Thick document	Dewarping algorithm	[28][30] [31]
		Junk regions	Mismatch between capture region and page region	1. Auto cropping 2. Manual cropping	[4][24]
		Skew	Page placement, source document	1. Auto deskew 2. Manual deskew	[34][35]
	Poor document quality	Low contrast, bleed-through, faded background	Aging documents, printing defects	Digital image enhancement	[36][37]
Image analysis and recognition	Incorrect image segmentation	Incorrect region types or ranges	Zoning	1. Correction through UI 2. Learning from manual correction	[47][48] [51][52]

Stages	Quality problems	Problem types	Problem causes	QA methods	Refs
		Incorrect pixel allocation	Layering	Reprocessing with different parameters	[43][46]
	Text recognition errors	Miscellaneous	Miscellaneous	1. Combination 2. Manual correction through UI	[14][29] [39]
	Document structure analysis errors	Miscellaneous	Complex document structure	Manual correction through UI	[21][29] [40][41]
Error cascading	Current stage errors	Miscellaneous	Preceding stage errors	Sensitivity analysis and feedback	

3.1. Cataloging

The purpose of cataloging is to define the necessary metadata for a document. Although this step itself does not involve image capture and processing, those metadata have direct impact on subsequent digitization steps (discussed later in subsection "Error cascading") and information retrieval (IR) applications. Common metadata can include title, author(s), subject(s), publishing date, number of pages, language(s) used, ISSN/ISBN, page size, and so on. An effective way of cataloging is to automatically import data from existing catalogs, such as Library of Congress Catalog [16], RLG Union Catalog [17], and OCLC Union Catalog [18]. Assuming the source catalog has been verified and in use for long time, catalog importing can guarantee a high accuracy of cataloging data. However, not all documents are cataloged and the available catalogs may not provide all the required metadata. Thus, manual cataloging usually cannot be completely avoided and may introduce errors into the catalog data. The library community has established procedures to reduce cataloging errors. For example, OCLC [19] combines automatic database scan with manual correction (end user feedback [20] through Internet is a very effective way).

3.2. Image capture

This is the initial step where the paper documents are converted to the raw electronic images. Many quality problems can emerge from this step:

Missing pages: No matter if an operator manually flips pages or an automatic page feeder injects pages, some pages may be left out. This problem has been observed by many people [4][10]. Internet Archive's Scribe [23] has an interface displaying all the captured page numbers on a screen to let the operator decide if any pages are missing. To reduce manual work, Lin and Xiong introduced a page-association algorithm to decide the page number of each page and then use the sequence pattern to detect potential missing pages [21].

Artifacts of camera-based capture: Scanner and digital camera are the two most common imaging devices in document digitization. Scanner can easily keep consistent lighting both across pages and within a page. But it cannot handle thick (unless dissembled into pages) or fragile books well, which are not uncommon in large-scale projects. This weakness of scanners and dramatic image quality improvement of digital cameras make digital camera a better choice as the imaging device in many projects. Examples include the BookScanner [22] by PARC and the more recent Scribe [23] by Internet Archive. However, it is challenging to control the lighting condition and color accuracy when using a digital camera to capture document images. To minimize artifacts such as uneven lighting, geometric distortion, and out-of-focus, the imaging device usually needs to be calibrated (for example, using a reference page) before capturing a book. To automatically detect remaining problems, Simske et al. [3] proposed inserting a barcode page at the beginning of each book and then measuring the MTF (modulation transfer function) to decide the scanning quality. Barney [25] used corners in bi-level images to estimate scanning characteristics. After the capture quality problems have been detected, a wide range of image processing techniques surveyed by Dormann et al. [26] can be employed to remedy them.

Page warping: Pages can warp up in a thick book. Many algorithms based on text line distortion models [28][30][31] are proposed to deal with the nonlinear warping. However, warping can be so serious that part of the page is missing or the squeezed contents become unreadable and then no dewarping algorithm can recover the lost information. In this case, the only solution is to detect such situation either manually or automatically (algorithms yet to be seen) and recapture the page.

Junk regions: Because the actual capture area is usually larger than the page of interest, junk regions can be present in the forms of black/shaded borders, portions of the facing page, or parts of the supporting surface. COTS OCR software [32][33] has limited capability of discarding junk regions, but such built-in processing is seldom sufficient due to lack of knowledge on the capturing conditions. So special algorithms are designed to remove junk regions. Fan et al. [24] combined two cropping algorithms, one based on line detection and the other on text region growing, to achieve robust cropping. Bourgeois et al. [4] proposed a morphology-based algorithm to detect and remove line frames in medieval manuscripts.

Skew: Image skew is very common in capture caused by either the way the page is placed on the imaging surface or printing defects (especially in old books). It has been researched for long time [34][35]. Many skew detection algorithms work accurately if the assumed models, mostly on page edges and text lines, are valid. The remaining challenge is to handle exceptions when the assumptions do not hold, for example, a page without rectangular or linear elements. At least the deskew algorithm should alert the operator on such corner cases.

Poor document quality: An aging document may have very low contrast and faded background, making it difficult to capture high quality images. Since we cannot change the physical pages, digital image enhancement techniques have to be applied to improve the image's visual quality. For example, Nishida et al. [36] introduced a multi-scale algorithm to reduce bleed-through (contents printed on one side of a page shows up on the other side). In a large scale document digitization project, the challenge is to automatically decide "when to apply which enhancement" because image enhancement techniques have adversely affect an image's quality if applied to the wrong image. Boutros [37] described a prototype that can automate the enhancement process.

3.3. Image analysis and recognition

Text recognition errors: QA requirements can vary a lot on the text recognition. A common approach is to hide the recognized text behind the processed document image and only use the text for information retrieval purposes such as searching and indexing. This approach is supported by popular electronic book formats such as PDF and DjVu and is adopted by the leading online book browsing/reading services, including Google Print [6], Amazon [9], and Internet Archive [8]. As Taghva analyzed in [38], on normal quality documents state-of-the-art OCR engines can satisfy the quality required for information retrieval without human intervention. However, other applications want to reuse the contents or reassemble the pages from recognized text and image objects and thus require a specified level of recognition rate [14][29]. In this case, automatic QA techniques like classifier combination (see the comprehensive review by Rahman [39]) and manual correction [29] have to be applied.

Document structure analysis errors: As surveyed by Mao et al. in [40], after decades of research, page layout analysis still has a lot to be desired, especially in terms of formal models, quantitative measurements, and performance on complex documents. Higher-level logical structure analysis, such reading order detection and document-wide structure understanding, is even more challenging. For example, [21] and [41] reviewed and proposed techniques to automatically extract and analyze the table of contents of a document, which is a common task in book digitization. Generally speaking, the existing document structure analysis algorithms cannot reliably handle the variety of documents common in a high volume digitization project. So usually page-by-page verification through some graphic user interface is necessary to guarantee the accuracy [29].

Image segmentation errors: Each page is captured as a flat homogenous image and the size is normally large in order to retain enough details. Although the raw image can serve archiving purpose well so that it can be potentially reprocessed by the next generation DIA techniques, its large size makes it unsuitable as the final delivery format for the end user. To reduce the file size while keeping fidelity, the raw image has to be segmented into different types of parts, which are compressed using different techniques. There are two alternative image segmentation strategies: zoning and layering. With the zoning method, an image is decomposed into non-overlapping regions of different types, such as text, graphics, and picture using page layout analysis algorithms discussed earlier. All versions of PDF specifications [42] support this method. With the layering method, an image is separated into different layers, including a foreground layer, a background layer, and a mask layer that allocates each pixel to either foreground or background. DiVu [43] is one of the earliest implementations to support the layering approach. Both research [24] and commercial [44][45] systems have been developed to compress images into layered PDF, which is not backward compatible with earlier versions of PDF specification.

The two methods both carry risks and may make some regions out of place or even illegible in the worst case. The zoning method may classify one text block as picture region and render it in gray scale while rendering the other text blocks on the same page as bi-level. It may also classify one picture block as text region and render it as bi-level, thus losing significant information in the picture. The layering method can allocate sharp areas in a picture into the foreground layer, or even allocate some characters into the background layer making them unreadable ([46] presented some low contrast pages where DjVu fails). Errors of the layering method are usually less visible than those of the zoning method because they are distributed over the whole page rather than concentrating in certain regions. However, this error dilution feature also makes it almost impossible to manually correct image segmentation errors since the layer allocation is conducted at the pixel level. Reprocessing the image with different parameters may be the only practical way to get better but still imperfect layering result. Fixing zoning errors is possible with an interactive graphic user interface, but this means increased labor and time cost. Achieving consistent visual quality under cost constraints remains an open problem in large scale document digitization projects. So far the digitized books hosted on most web sites exhibit the above mentioned artifacts to some extent.

3.4. Error cascading

The errors in upstream processing steps can directly lead to quality problems in the downstream processing steps. It is easy to understand that poorly captured raw images will pose challenges for further image processing and recognition. Less obviously, even cataloging step can also have big impact on later DIA steps. For example, as mentioned in [24], Internet Archive hosts books in a wide range of languages and sometimes a single book contains more than one language. Because most state-of-the-art OCR engines only support limited language detection, usually within several languages, the language metadata are passed on to the OCR engine as parameters in order to efficiently and accurately recognize the text. Consequently, incorrect or incomplete language metadata for a book can result in poor or even unusable OCR results. In order to control error cascading, sensitivity analysis should be carried out to decide the tolerated error rates for individual stages. In practice, this is very challenging since some DIA steps do not have reliability measurements at all. With the presence of error cascading, the processing pipeline needs to have feedback mechanism in place so that an error can be traced to the real source. The author has seen little existing research in this direction.

4. General QA strategies

In the previous section we have surveyed the quality problems and corresponding technical solutions for each step. This section generalizes the QA strategies applicable to different stages.

4.1. Combination of automatic QA and manual QA

As shown in many of earlier examples, manual check cannot be completely avoided because there are always corner cases automatic QA will fail. Meanwhile, from the perspective of reducing cost and improving speed, manual check should be reduced as much as possible. Thus, a recurring strategy is to combine automatic QA and manual QA (see Figure 1). Automatic QA first attempts to correct the quality problem. If auto-QA is regarded reliable, manual QA can be bypassed. In the ideal case, most samples should directly pass through auto-QA, leaving only a small percentage of samples to be manually checked upon. Displayed as the dotted line, an interesting path is to use manual QA to boost auto-QA, thus reducing future manual QA. Several commercial OCR software packages have incorporated this concept in the proofing tool: If the user has corrected one character, the system can automatically correct similar characters upon the user's confirmation. This concept has also been applied to layout analysis. As described by Malerba et al. in [47] and Ma et al. in [48], the systems use machine learning to automatically correct some zoning errors based on a few manual corrections.

To implement this strategy, it is critical to estimate the auto-QA's reliability on individual samples and decide when to invoke manual operation. There is some existing work on estimating the accuracy of text recognition without knowing the ground truth. Sarkar et al. [27] applied latent conditional independence models to OCR's conference scores in order to decide if a page should be routed for manual check. Lin et al. [50] used adaptive confidence transform to predict OCR recognition rate. On zoning-based image segmentation, Simske et al. [3] compared the re-rendered page with the original image and used the difference after image registration as an indicator of zoning reliability. Many auto-QA algorithms have some assumptions on the input. Then we can also measure how well those assumptions are satisfied and use this as an indicator of reliability. For example, if a dewarping algorithm depends on the text lines to construct the model, it should output a low reliability score if very few text lines are located. Besides, quantitative reliability metric is preferred over simple Boolean metric because it allows more flexible tradeoff between quality and cost by just adjusting the reliability threshold.

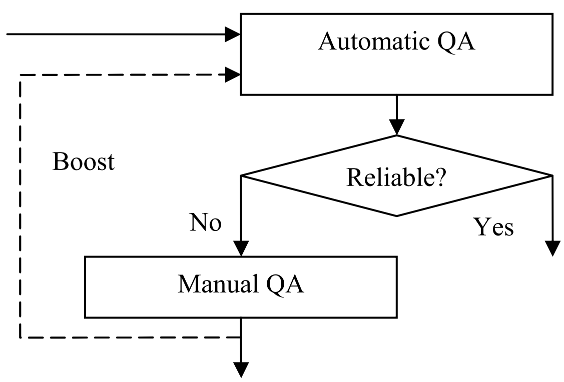

Figure 1: Combination of manual and automatic QA

4.2. Batch QA

One characteristic of high volume document digitization is that a document can contain multiple pages similar in many aspects (font, layout, style, etc.) and contextually related. This can potentially benefit QA process. For example, multiple pages are used for frame cropping by exploiting frame consensus [5][24]; missing pages are identified through considering page numbers on continuous pages [21]. The downsides of batch QA include extra complexity (problematic pages detected by batch QA have to be reinserted into the workflow) and storage (the raw images have to be stored to do batch QA) in the processing.

4.3. Usability of QA system

Besides the various algorithmic aspects surveyed so far, a practical issue is how to make the QA system easy to use and efficient. As discussed earlier, the current layout and structure understanding methods are still very fragile and manual correction is usually necessary. The natural question is to how to design a good correction user interface. Commercial OCR software packages have been providing generic interactive editing and proofing tool for long time. In the research community, we have also seen a lot of related work in layout analysis performance evaluation [51][52] and specific digitization projects [29].

4.4. Open source QA

Leveraging Internet as a world-wide collaboration medium, some large digitization projects resort to the "Open source QA" model, in which the volunteers around the world can participate in the QA process, especially the time-consuming manual QA part. In theory, this model can really push the envelope of the quality standard in document digitization. On the other hand, this QA has drawbacks similar to open source software: lack of control on the progress, intellectual property issues, et al. So far, this model is mainly used in non-commercial projects that process out-of-copyright documents for the benefit of general public. Good examples include the Project Gutenberg [2], the Million Microfilm Project [5], the Bookshare Project [53], and the French Archives Project [54]. It would be interesting to see how commercial digitization projects can adopt this model.

5. Conclusions

This paper attempts to abstract from existing work common QA issues and solutions in high volume document digitization. The key observations include:

- QA plays an important role in document digitization to deal with the imperfect DIA components and the human aspects of the processing.
- Quality issues together with corresponding solutions exist throughout the major DIA steps in the digitization system.
- Quality problems can cascade through the whole pipeline and feedback mechanism is needed to trace the source of the problem.
- Combining auto-QA and manual QA is effective to satisfy both the quality standard and cost

constraints. Confidence reporting from individual DIA components will greatly facilitate QA.

This paper is not intended to be exhaustive. We concentrate on the QA special to DIA rather than the general software and hardware QA aspects (monitoring, automatic recovery, data integrity, etc.). In addition, some "boutique" type of digitization projects target special categories of documents (for example, historical documents [55]) and may have unique QA needs and methods.

5. Acknowledgements

The author would like to thank Jian Fan, Steven Simske, Steven Rosenberg, Dan Tretter, Richard Schedler, Brewster Khale, and many others for the collaboration and discussions on book digitization projects, which motivate the author to write this survey.

6. References

[1] G. Thoma and G. Ford, "Automated Data Entry System: Performance Issues," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, San Jose, 2002, pp. 181-190.

[2] Project Gutenberg website, http://www.gutenberg.org.

[3] S. Simske and X. Lin, "Creating Digital Libraries: Content Generation and Re-mastering," *Proc. International Workshop on Document Image Analysis for Libraries*, Palo Alto, January 2004, pp. 33-45.

[4] F. Le Bourgeois, E. Trinh, et al., "Document Image Analysis Solutions for Digital Libraries," *Proc. International Workshop on Document Image Analysis for Libraries*, Palo Alto, January 2004, pp. 2-24.

[5] W. Barret, L. Hutchison, et al., "Digital Mountain: From Granite Archive to Global Access," *Proc. of International Workshop on Document Image Analysis for Libraries*, Palo Alto, January 2004, pp. 104-121.

[6] Google Print website, http://books.google.com

[7] Open Content Alliance website, http://www.opencontentalliance.org

[8] Internet Archive website, http://www.archive.org

[9] Amazon "Search inside Book" announcement, http://www.amazon.com/exec/obidos/tg/feature/-/507108/104-7825001-2871961

[10] H. S. Baird, "Difficult and Urgent Open Problems in Document Image Analysis for Libraries," *Proc. of International Workshop on Document Image Analysis for Libraries*, Palo Alto, January 2004, pp. 25-32.

[11] B. Kelly, A. Williamson, and A. Dawson, "Deployment of Quality Assurance Procedures for Digital Library Programmes," *Proc. the IADIS International Conference WWW/Internet*, July 2003.

[12] S. Yacoub, "Automated Quality Assurance for Document Understanding Systems," *IEEE Software*, May/June 2003, pp. 76-82.

[13] D. Doermann, H. Ma, et al., "Translation Lexicon Acquisition from Bilingual Dictionaries," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, San Jose, 2002, pp. 37-48.

[14] A. Belaid and L. Pierron, "A Generic Approach for OCR Performance Evaluation," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, San Jose, 2002, pp. 203-215.

[15] ISO homepage, http://www.iso.org

[16] Library of Congress website, http://catalog.loc.gov/

[17] RLG Union Catalog website, http://www.rlg.org/libres.html

[18] OCLC Union Catalog website, http://www.oclc.org/worldcat/default.htm

[19] OCLC QA procedure, http://www.oclc.org/bibformats/en/quality/

[20] LOC catalog error reporting, http://www.loc.gov/help/contact-libarch-report.html

[21] X. Lin and Y. Xiong, "Detection and Analysis of Table of Contents Based on Content Association," to appear in *International Journal on Document Analysis and Recognition*

[22] S. Ready and R. Street, "The PARC Bookscanner," http://www2.parc.com/emdl/members/ready/bookscan.pdf

[23] B. Khale, "The Open Library," http://www.openlibrary.org/details/openlibrary

[24] J. Fan, X. Lin, and S. Simske, "A Comprehensive Image Processing Suite for Book re-mastering," *Proc. ICDAR 2005*, Seoul, South Korea, pp. 447-451.

[25] E. B. Smith, "Estimating Scanning Characteristics from Corners in Bilevel Images," *Proc. SPIE Conference on Document Recognition and Retrieval VIII*, San Jose, 2001, pp. 176-183.

[26] D. Doermann, J. Liang, and H. Li, "Progress in Camera-based Document Image Analysis," *Proc. Seventh International Conference on Document Analysis and Recognition*, pp. 606-610, 2003.

[27] P. Sarkar and H. Baird, "Triage of OCR Results Using 'Confidence' Scores," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 216-222, San Jose, 2002.

[28] A. Ulges, C. H. Lampert, and T. M. Breuel, "Document Image Dewarping Using Robust Estimation of Curled Text

Lines," *Proc. of International Conference on Document Analysis and Recognition*, Seoul, 2005, pp. 1001-1005.

[29] S. Yacoub, V. Saxena, and S. N. Sami, "PerfectDoc: A Ground Truthing Environment for Complex Documents," *Proc. International Conference on Document Analysis and Recognition*, pp. 452-456, Seoul, South Korea, 2005.

[30] L. Zhang and C. L. Tan, "Warped Image Restoration with Applications to Digital Libraries," *Proc. of International Conference on Document Analysis and Recognition*, Seoul, 2005, pp. 192-196.

[31] H. Cao, X. Ding, and Changsong Liu, "Rectifying the Bound Document Image Captured by the Camera: A Model Based Approach," *Proc. Seventh International Conference on Document Analysis and Recognition*, 2003, pp. 71-75.

[32] ABBYY company website, http://www.abbyy.com

[33] ScanSoft company website, http://www.scansoft.com

[34] H. Baird, "The Skew Angle of Printed Documents," *Proc. Conf. of the Society of Photographic Scientists & Engineers*, Rochester, New York, May, 1987.

[35] D. Bloomberg, G. Kopec, and L. Dasari, "Measuring Document Image Skew and Orientation," *Proc. SPIE Conference on Document Recognition II*, 1995, pp. 302-316.

[36] H. Nishida and T. Suzuki, "A Multiscale Approach to Restoring Scanned Color Document Images with Show-through Effects," *Proc. Seventh International Conference on Document Analysis and Recognition*, 2003, pp. 584-588.

[37] G. Boutros, "Automating Degraded Image Enhancement Processing (DIEP)," *2005 Symposium on Document Image Understanding Technology*, College Park, Maryland, Nov. 2005.

[38] K. Taghva, J. Borsack, and A. Condit, "Evaluation of Model-based Retrieval Effectiveness with OCR Text," *ACM Transactions on Information Systems*, vol 14, January 1996, pp. 64-93.

[39] A.F.R. Rahman and M.C. Fairhurst, "Multiple Classifier Decision Combination Strategies for Character Recognition: A Review," *International Journal of Document Analysis and Recognition,* vol 5, 2003, pp.166-194.

[40] S. Mao, A. Rosenfeld, and T. Kanungo, "Document Structure Analysis Algorithms: A Literature Survey," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, Santa Clara, 2003, pp. 197-207.

[41] H. Déjean and J.-L. Meunier, "Structuring Documents According to Their Table of Contents," *Proc. ACM Conference on Document Engineering*, Bristol, UK, 2005.

[42] Adobe PDF specification website, http://partners. adobe.com/public/developer/pdf/index_reference.html

[43] P. Haffner, Y. L. Cun, L. Bottou, P. Howard, and P. Vincent. "Color Documents on the Web with DjVu," *IEEE Int. Conf. on Img. Proc*., vol 1, Kobe, Japan, October 1999, pp. 239-243.

[44] LuraTech company website, http://www.luratech.com

[45] CVISION company website, http://www. cvisiontech.com

[46] U. Garain, T. Paquet, and L. Heutte, "On Foreground-background Separation in Low Quality Color Document Images," *Proc. Seventh International Conference on Document Analysis and Recognition*, 2005, pp. 585-589.

[47] D. Malerba, F. Esposito, and O. Altamura, "Correcting the Document Layout: A Machine Learning Approach," *Proc. Seventh International Conference on Document Analysis and Recognition*, 2003, pp. 97-102.

[48] H. Ma and D. Doermann, "Bootstrapping Structured Page Segmentation," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, Santa Clara, 2003, pp. 179-188.

[49] M. Mirmehdi, P. L. Palmer, and J. Kittler, "Towards Optimal Zoom for Automatic Target Recognition," *Proc. of 10th SCIA*, 1997, pp. 447-453.

[50] X. Lin, X. Ding, Y. Chen, J. Liu, and Y. Wu, "Evaluation and Application of Recognition Confidence in OCR," *Proc. of ACCV'98*, Hongkong, Jan. 1998.

[51] B. Yanikoglu and L. Vincent, "Pink Panther: A Complete Environment for Ground-truthing and Benchmarking Document Page Segmentation," *Pattern Recognition*, vol 31, Sept 1998, pp. 1191-1204.

[52] S. Mao and T. Kanungo, "PSET: A Page Segmentation Evaluation Toolkit," *Fourth IAPR International Workshop on Document Analysis Systems*, Rio de Janeiro, Brazil, Dec 2000.

[53] Benetech's Bookshare Project website, http://www .bookshare.org

[54] B. Couasnon, I. Leplumey, "A Generic Recognition System for Making Archives Documents Accessible to Public," *Proc. Seventh International Conference on Document Analysis and Recognition*, 2003, pp. 228-232.

[55] A. Antonacopoulos and D. Karatzas, "Document Image Analysis for World War II Personal Records," *Proc. International Workshop on Document Image Analysis for Libraries*, Palo Alto, January 2004, pp. 336-341.