IJDAR (2004)
Digital Object Identifier (DOI) 10.1007/s10032-004-0124-5

International Journal on IJDAR Document Analysis and Recognition

Reports of the DAS02 working groups

Elisa Barney Smith[1], David Monn[2], Harsha Veeramachaneni[3], Koichi Kise[4], Alessio Malizia[5], Leon Todoran[6], Adnan El-Nasan[3], Rolf Ingold[7]

[1] Department of Electrical and Computer Engineering, Boise State University, Boise, ID, USA
[2] Center for Communications Research, Princeton, NJ, USA
[3] Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[4] Osaka Prefecture University, Osaka, Japan
[5] Department of Information Science, University "La Sapienza" of Rome, Rome, Italy
[6] Intelligent Sensory Information Systems, University of Amsterdam, Amsterdam, The Netherlands
[7] DIUF, University of Fribourg, Fribourg, Switzerland

Published online: ♣ 2004 –

Abstract. This document is a collection of four working group reports in the areas of digital libraries, document image retrieval, layout analysis, and Web document analysis. These reports were the outcome of discussions by participants at the Fifth IAPR International Workshop on Document Analysis Systems held in Princeton, NJ on 19–21 August 2002.

1 Introduction

Document image analysis and understanding has been a fertile research area supported by an active international community for many years. With the advent of the World Wide Web and digital libraries and the exponential growth in documents accessible online, new challenges are arising every day. As a means of identifying the numerous problems that will likely need to be solved over the next decade, a working group session was organized as an integral part of the Fifth IAPR International Workshop on Document Analysis Systems (DAS02).

There were four parallel sessions devoted to important topics in document analysis and with implications for the design and implementation of such systems. The participants in the workshop were allowed to join the group that best matched their research interests. These four groups were focused on (a) digital libraries, (b) document image retrieval, (c) layout analysis, and (d) Web document analysis, although the discussion often strayed outside the predetermined boundaries. The structure was designed to encourage free-form "brainstorming" with participants expressing their viewpoints and speculating on the four themes, as chosen by the chair of the working group session, Henry Baird, to reflect current trends and challenges in document analysis systems.

In each group, one person was assigned to be moderator and another given the role of scribe, whose job it was to capture the essence of the discussion. It is through their efforts that we are able to present the following reports.

Dan Lopresti, Jianying Hu, and Ram Kashi

2 Working Group on Digital Libraries and Antique Documents

This section summarizes the discussions of the Working Group on Digital Libraries and the Analysis of Antique Documents. Eight researchers from two countries participated: B. Agüera y Arcas, H. Baird, E. Barney Smith (scribe), A. Dengel, D. Lopresti, D. Monn (moderator), J. Uchill, and L. Vincent. The participants represented a mixture of both private industry and universities.

The working group recognized that there are already a number of well-known digital libraries available online today, including the Making of America Collection [1], the U.S. Library of Congress [2], and other specialized collections [3].

Still, despite the obvious potential synergies between document analysis research and digital libraries, there has not been much interaction between the two communities. Digital libraries are typically built using off-the-shelf commercial OCR systems, oblivious to the more advanced document analysis techniques under development in our field. On the other hand, most document analysis researchers are not aware of the special problems that arise when building digital libraries, nor do they regard the vast collections of scanned document images now accessible on the Web as a resource that could be invaluable in their work.

The group first identified what it felt were the challenges facing digital libraries. We then discussed several of the features we thought would be good for digital libraries to have. The remaining time was spent discussing what our community could provide to libraries and institutions that are trying to create digital libraries.

2.1 Challenges of digital libraries

Digital libraries are emerging as a supplement to traditional libraries. Still, their growth is in its beginning stages. Two goals in particular are of concern in constructing libraries of textual material. The first is providing digital images of sufficient quality for use by those who wish to view the documents in their original form, whether it be for reading or for examining features such as printing style, text layout, marginalia, or nontextual elements such as pictures and graphics. The second is providing an accurate transcription of the text, not only for searching but also for ease of reading and printing when content, and not necessarily how the text originally appeared, is the main focus. This second goal usually depends on the first to the extent that the scanned images must be of sufficient quality to achieve high accuracy in the transcription by an OCR engine.

In order for libraries to meet these goals efficiently when converting significant portions of their collections to digital form, the current processes could likely benefit from increased automation. For example, the scanning process alone currently requires a significant amount of clerical support, such as identifying poorly scanned pages and rotating upside-down text. Certain documents, especially those that are very old, may require specialized imaging techniques.

Since it is often desirable to be able to read the transcription of a document, there are several issues that must be addressed beyond simply having very high accuracy. Determining the proper reading order for both columns and footnotes poses an enormous challenge. Understanding where in the text each footnote is referenced may require the recognition of special characters (daggers, etc.) that may not even be part of the OCR character set. Some documents may also contain other special characters such as diacritical markers or section and paragraph symbols (pilcrows). Once one has an understanding of the footnotes and references and where they occur, there is the issue of how to present them to the reader. Perhaps Web-based libraries of the future will be able to provide hyperlinked references and pop-up windows containing the footnotes.

The OCR of mathematical equations continues to be a problem, both in recognizing the symbols and interpreting the equations for proper viewing and printing. The same is true for understanding tables. In technical documents, equations, tables, and figures are often numbered and referenced elsewhere. Recognizing these labels and matching them up with the references requires a special understanding.

A lack of funding is the most commonly cited reason for why there are not more digital library projects under way currently. New books that are published through digital technology are not immediately contributed to digital libraries, even though the major portion of the cost, which is doing the digitization, would not be incurred. Part of the reason for this was attributed to copyright issues and worries that publishing online would decrease hardcopy sales.

Some publishers have found that simultaneously publishing books in both paper and electronic form has actually led to an increase in the sales of the paper version. Amazon.com has many sample pages of books available on the Web now as a tool to increase their sales. Many workshop and conference proceedings, including the proceedings of the DAS02 workshop [8], are also available electronically.

There are several very nice digital library projects currently under way, but they are not interconnected. Should they be? What effect would a large centralized digital library have on smaller digital library projects or smaller paper libraries? What should the architecture of a global digital library be?

2.2 What features does the DIA community wish to see?

There were many additions that were felt would increase the value of a digital library. It seems possible that advanced document analysis techniques could open up new options for the delivery of content to users, thereby increasing the perceived value of the information. Decisions on how to best deliver the content to users need to be made. Configuring digital libraries so that the contents can be easily viewed on a PDA was the first feature working group members suggested be added to future digital libraries. Having multiple layers in the document to represent the image, the OCRed text, hyperlinks, highlights, notes, etc. would also add value to the library.

To make the content of greater use, the libraries must be easily searchable. The search engine can focus on the text data, the interpreted content represented by that text, or the structure of the document. Members agreed that integrating these would enable digital libraries to go beyond a simple search.

There was also discussion about the fact that when a digital library is created, it would benefit our community and other users if the metainformation on the collection were included. It was felt that most DIA researchers were not aware of all existing digital library corpuses or how much of what types of data is in each one. This information could make the digital libraries a useful source of data for DIA research.

The concept of personal digital libraries arose. Are there tools already out there, or would it be reasonable to develop tools so that people could create their own personal digital library from their own resources? Scanner hardware that was less burdensome on users so that people could digitize a document page by page as they read it and a digital camera mounted on eyeglass frames were two proposed ideas. The availability of scanning hardware that was less harmful to antique books was another issue brought forth.

2.3 What could the DIA community provide?

Members agreed that the DIA community could use its experience to make recommendations to libraries as they begin a digitization project.

One topic of interest to members was whether libraries were using the best scanning resolution when doing their digitization. The required resolution depends on the type of input document, particularly where antique documents are concerned. Older documents should not be rescanned often, so starting with the correct resolution is important. For these documents, the details about the printing and paper are often as important as the textual content, making higer-resolution scans more important for such documents than for a recent publication. Some guidelines suggested by the group were that documents printed before the year 1600 should be scanned at 2,000 dpi, documents from 1600–1800 at 600 dpi, and documents printed more recently than 1800 at 400 dpi. These were all heuristics, and developing some better reasonings for these guidelines is a possible direction for growth. The decision on whether to scan in color vs. grayscale vs. bilevel was also touched upon as an area where our community could help guide libraries involved in these projects.

When the digitized document is viewed, a decision needs to be made about whether to keep the original digital image or just the converted OCRed text. Enhancement of the original images might be necessary for improved legibility or to improve OCR. These two criteria are not always equivalent. It was agreed that when enhancement is done, the original image should still be saved for future use.

It was mentioned that the value of the library could be increased by expanding it through annotations, corrections of the OCR layer, addition of knowledge, etc. and that the users of a digital library could contribute to this if the proper framework were developed.

The discussion moved on to what software our community could provide to libraries. Should we encourage our software to be embedded on their digital library site or make the software external to the site in the same vein as a search engine like Google is external to other Web sites but can be used to search a local site?

2.4 Summary of the working group

The working group concluded that digital libraries were of interest to DIA researchers and members looked forward to the increase in size and number of digital libraries. The concluding thoughts revolved around the question, What DIA technology is applicable to what aspects of the problem? We have developed many useful tools as parts of our research, but integrating them and making them available to other researchers and implementers of digital libraries could be a place for improvement. What fundamental technologies should we be focusing on to help digital libraries expand?

Elisa Barney Smith and David Monn

3 Working Group on Document Image Retrieval

This working group consisted of the following members, who actually expressed interest in a wide variety of topics, including document image retrieval, classifier combination, handwriting recognition, and learning: K. Kise (moderator), G. Nagy, A. Bagdanov, H. Veeramachaneni (scribe), E. Ishidera, S. Jaeger, and J. Wnek.

The moderator started the discussion with his interest, document image retrieval. However, because of the range of interests present, the session consisted mostly of brainstorming, and no definite conclusions were drawn. The following are the list of items we discussed.

3.1 Open problems

Document image retrieval may offer exciting new applications that will attract researchers to the field of document analysis. What interesting applications can be envisioned?

What is document image retrieval? The definition depends on the definitions of the terms “document”, “query”, and “database” with respect to the particular problem. How can you pose the question of text retrieval/image retrieval?

Can retrieval of images from a database that are similar to a query image be considered as document image retrieval? No, because similarity between general images is subjective. Therefore, it is necessary to precisely define the scope of the term “document”.

Can queries be images or do they have to be words (or symbols)? If the queries are images, do they have to be symbolically represented? If so, should we limit the problem to annotated image databases? How can queries be described? If queries are symbolically represented images (the symbols may be the features that are extracted), should the user of the system agree with the features that are used? What if the user is not a human but a machine that is using the retrieved images for some postprocessing? Can we phrase the definitions and problems in terms of the downstream application?

How can image retrieval systems be evaluated? What criteria represent the validity of the features used and the overall accuracy in the retrieval system?

3.2 Summary of the working group

Through the discussion we were surprised to learn that there is no clear consensus on a relatively simple term, “document image retrieval”. The main discussion was on the types of queries that characterize retrieval methods as well as databases.

As we noted at the beginning of the section, this working group was comprised of participants with varying interests. Hence the moderator thought that it would be a good idea not to dwell on a specific issue but to consider generic (or metalevel) topics such as: What is the most important problem and why? What is the future of our technology? Is the field of document analysis going to be extinguished by growing use of electronic documents and XML? These topics are open for future discussion.

Harsha Veeramachaneni and Koichi Kise

4 Working Group on Layout Analysis

This section summarizes the discussions of the Working Group on Layout Analysis. Nine researchers from different countries participated: F. De Rosa, M. Bilderbeek, P. kok Loo, K. Hadjar, E. Bodansky, T. Breuel, Y. Zheng, L. Todoran (moderator), and A. Malizia (scribe).

The participants were divided almost evenly between private industry and universities. The group first spent a few minutes identifying main research topics like representation schemes and image formats and successively discussing major applications and evaluation of results, concluding with suggestions on future discussions.

4.1 Document layout analysis

Starting with the definition of layout analysis, we have tried to underline the main characteristics of this field. Layout analysis extracts the geometric structure of a scanned document image. It is hard to find a general-purpose method that can achieve high-precision results on different kinds of documents (Sect. 4.3); therefore user feedback is needed to get improved performance.

4.2 Methods

After reviewing the main layout analysis points, we discussed the most common techniques used in this field. While document layout analysis is a simpler problem than general image segmentation, it still raises challenging issues in geometric algorithms and image statistics.

A wide variety of algorithms have been proposed for layout analysis [9,6,5]. Among them are:

- Analysis of connected components
- Projection (recursive X-Y) cuts
- Split and merge
- Quad-tree techniques
- Analysis of the background structure
- Texture based analysis
- Morphology-based approaches
- Local vs. global geometric feature

4.2.1 White space analysis. White space analysis consists of finding a cover of the background white space of a document in terms of maximal empty polygons (usually rectangles). There are also granulometry approaches based on rectangles of varying size and aspect ratio. These rectangular granulometries are used to probe the layout structure of document images, and the rectangular size distributions derived from them are used as descriptors for document images.

4.2.2 Matching-based methods. The matching-based methods may be automatic or manual depending on a set of parameters that can be tuned to evaluate the segmentation and classification results. Using these parameters (which could also be set by the user), the method can decide which kind of matching phase to perform – manual or automatic.

Manual indexing will require a user to data-entry information from forms. Even if using automatic recognition, we will have a cost savings in terms of indexed documents per second. Moreover, using both automatic and manual segmentation, semiautomatic indexing can be performed, which could help in validation for manual data entry.

4.2.3 Machine learning. Understanding documents is a relatively easy task for an intelligent human reader most of the time. This is due to the fact that the documents are prepared using some common assumption about structuring them, and authors intend to convey information in ways that allow readers accurate and efficient interpretation. Some methods use inductive learning from examples of documents with identified data elements, with high automation and minimal user input. Other methods use incremental learning in an interactive environment, where the classification is driven by a model that contains a static as well as a dynamic part and evolves through use.

In summary, the goal of the combined approach is to automatically "learn" complex document structures, store them in general templates, and utilize them in a data extraction process.

4.3 Image formats

One of the most important issues in document analysis is document image formats. In fact, the knowledge (or conversion) of the data format is fundamental for selecting the correct types of filtering and operators used in layout analysis.

4.3.1 Black and white. Due to the prevelance of simple, black-and-white documents, and for computational reasons, most layout analysis techniques were developed for bitonal images. These methods are less expensive, in terms of computation, geometric, and stochastic operators, than those applied to gray-level or color images. Mean, variance, and connected components are only some of the possible operators.

The classification and segmentation strategy changes depending on the types of filters that could be used on a document image. We also must recall that many algorithms for segmenting document images perform their preprocessing phases on a bitonal version of the original gray-level image.

The major drawback of black-and-white techniques comes from their simplicity: they cannot handle the more complex document images that are becoming more common nowadays.

4.3.2 Gray level. The scanning phase performed in order to acquire document images usually outputs gray-level images. For simple documents, a binarization step

(using an adaptive threshold, for instance) is often used to convert gray images to black-and-white images. Then all processing is done with bitonal techniques.

For certain documents, however, the binarization process removes essential information. For instance, some layer structure induced in documents by using gray shading is lost in binary images, where everything is either background or foreground; intermediate levels are not possible. An example here is text written on images. Furthermore, feeding gray-level images directly into OCR can improve OCR results. Also, text/picture separation is more accurate in gray images. Lately, an increasing interest is being shown in processing gray images.

4.3.3 Colors. Color now plays an important role in publishing everything from scientific journals, newspapers, and magazines to advertisements. Besides the aesthetic reasons, the use of color in printing allows the publisher to convey more logical information to the reader. The nature of documents in current applications is therefore rapidly shifting from simple black-and-white documents to complex color documents. The layered structure, already present in primitive form in some gray documents, is now far more complex, and the use of text printed on color background images is a common practice. The color of the text is no longer uniform. It is clear that these complex color documents cannot be processed as gray or binary images. New techniques need to be developed to tackle this problem.

However, whereas document analysis for black-and-white documents is mature, color document analysis is still in its infancy. Some tools have been developed to achieve color-based analysis:

- Color OCR,
- Color document compression,
- Color string localization.

But the problem is still challenging because of the fragmentation of the color background and text pixels, which affects these kinds of documents. Furthermore, until recently, no standard dataset existed for color document analysis (Sect. 4.5). As a consequence, each developer used his/her own dataset for evaluation. New ground truth datasets are starting to be proposed in order to make the evaluation process more objective.

4.4 Layout representation

The segmented information obtained by the various methods on different kinds of image documents is then organized structurally. This is useful in managing the layout analysis results (geometrical, statistical, templates). Commonly, a list of attributes is used to represent the data obtained from layout analysis. This list of attributes could be a set of pairs made of attributes and values. Even complex data structures could be found at this level, but generally we see an ordered list of document regions and their attributes.

Another typical approach is based on hierarchical organization of segmented information. This approach uses trees to build a compact and manageable representation of the document region's layouts. Using hierarchical structures, layout information could be subdivided into regions that could be connected by an arc if they shared a relationship based on the chosen classification and layout detection algorithm. Indeed, both lists-of-attributes and hierarchical approaches could use the XML standard format to represent data structures. In fact, in hierarchical approaches, a DTD (document type definition) could help to generate a structured template of the information trees.

More complex methodologies use a structured approach based on graph models. These graphs represent the relationships between layouts of document regions in a multipart information structure. For instance, geometric information concerning a region could be used to compute weights; those weights could be assigned to arcs that represent a proximity relationship between two different regions corresponding to nodes.

4.5 Datasets

Standard datasets are fundamental in evaluating layout analysis results and making those results available to and testable by the scientific community. We list below the well-known datasets available today.

First, in order to let a dataset be representative of the problem in question, it is important to quantify the complexity of a document in the collection prior to the evaluation phase. For instance, the UW series from the University of Washington is useful in standard journal evaluations, but not for generic complex documents like those obtained from magazines. Furthermore, if the complexity of the documents in a dataset is known and well defined, the complexity measures can be used to weigh the evaluation results leading to evaluation independent of page difficulty. We must point out that many large collections of document images are now becoming available online as part of digital library initiatives published on the World Wide Web. But these collections rarely include the ground truth information needed for evaluation.

Some of the existing datasets currently available are:

- UW-III [7] (University of Washington)
- MTDB [10] (University of Oulu, Finland)
- UNLV [11] (Univ. of Nevada Las Vegas)
- MOA (Making Of America)
- UvA-CDD[1] (University of Amsterdam)

The dataset discussion led us to questions about universal formats for documents in such datasets. Whereas the image format used is mostly TIFF or JPEG, for representation of the ground truth each dataset uses a different format. It was suggested that an XML (DTD based) representation could be used to store all of the information needed for ground truth. An effort should be

[1] Universiteit van Amsterdam Color Document Dataset: http://www.science.uva.nl/UvA-CDD/

made toward the standardization of existing and future datasets.

4.6 Major applications

The document layout analysis field is concerned with the automatic segmentation and interpretation of regions found in paper documents. We discuss below the main areas where such automation systems have been used:

1. Indexing and retrieval (e.g., position of the searched information on a page). Document image retrieval systems are of particular interest in some application areas such as the batch acquisition of paper documents. Given an example image as a query, a document image retrieval system should return a ranked list of visually similar documents from an indexed collection. In document collections, automatic conversion of documents is often expensive or impossible. In such cases, image retrieval may be the only feasible means of providing access to the document database.
2. Automatic genre classification, which is useful for grouping documents for routing through office workflows as well as for identifying the type of document before applying class-specific strategies for document understanding.
3. Online access to complex compound documents with client-side search and browsing capability is one of the key requirements for effective content management systems. Enterprise applications, including corporate intranet usage and internal workflow management systems, require rapid transmission and feature-rich viewing that enable users to quickly access and browse important documents.
4. Geometrical and logical structure information extracted from layout analysis could also be used in document type conversion and compression (e.g., bmp2pdf) and automatic linking of semantic information extracted from a document image (articles spread over various boxes on a newspaper page could be transformed into an A4 compact version, which is more human-readable).

4.7 Summary of the working group

After a discussion on layout analysis that lead from methods and algorithms to image formats and major application fields, we briefly summarize our conclusions with a sketch of a typical document layout analysis system as discussed during our meeting. There are four main components in a typical document layout recognition environment: the recognition module, the indexing module, the data-entry annotation for human feedback, and the query module. Moreover, we can also define an archive release module if we want to export indexes and images to other applications.

In conclusion, a proposed topic for future discussion has been raised. We refer to it as "automatic vs. interactive". The automatic recognition phase is based on the segmentation and classification methods, taking as input a paper document page, which is then presented to the user if manual indexing was decided (by the user or by a certain system parameter) with a visual interface for the data entry. A data-entry GUI could be developed to let the users evaluate the automatic classification performed by the system and edit the segmentation for refining the results. Since document layout analysis is a very specific task, it is hard to find a general-purpose method that can achieve high-precision results on different kinds of documents. Thus, a more extended discussion on the relationship between automatic and interactive techniques is needed to build document layout recognition systems that effectively work on a wide range of documents with good accuracy grades.

Some suggestions could be: techniques for improving interactive human verification of results and algorithms providing a method for evaluating their own results, for example using a range of values for the acceptance test of segmented regions. In fact, if the values are acceptable, the next phase of the system will be automatic indexing, while for those documents where the values are out of range, it should be data-entry annotation and verification by the user.

Alessio Malizia and Leon Todoran

5 Working Group on Web Document Analysis

Participants in this working group included: A. Antanacopoulos, A. El-Nasan (scribe), J. Hu, R. Ingold (moderator), R. Kashi, D. Karatzas, and J.E.B. Santos.

In the early stage of the discussion, two fundamental questions were raised: first, What is a Web document? and second, Is document image analysis still of interest? For the former question, it was observed that almost any kind of information that is displayable on a screen can be considered a Web document. It was also said that since, in the future, any document (in the classical sense) should become available on the Internet, Web document analysis may eliminate the need for one of these. However, it was also agreed that some specific new research topics are bound to arise.

The second question, about the utility of DIA in the context of growing collections of electronic documents, received a confident positive answer. Attacking the argument that the Semantic Web based on huge ontologies will provide all necessary high-level information in an explicit form, thereby eliminating the need for DIA, it was observed that, on the contrary, DIA tools may be very helpful in producing the necessary annotations. The image can be regarded as the only standard representation for Web documents.

The discussion also included tools, many of which exist in practice. Beyond text extraction, which is quite simple, it seems that very few tools are actually widely available. One example is the table extraction method presented by Wang and Hu at the workshop [12]. Although it was observed that similar approaches should also work on lists, it was also argued that some improve-

ments are necessary to deal with complex nested structures.

Finally, the working group came up with a list of research topics we consider to be the most important open issues. Among this broad list, the following topics were mentioned:

- Categorization of Web documents
- Reverse engineering tools working on dynamically generated documents
- Device-independent Web authoring tools, especially for the mobile Web
- Semantic analysis to augment the Semantic Web
- Text extraction from nontext data, such as images or animations
- Security checks via a variation on the well-known Turing test (as presented by Baird and Popat [4]).

Adnan El-Nasan and Rolf Ingold

6 Summary

Although significant progress has been made in research on document analysis systems, it is clear from the conclusions of the DAS02 working groups that, if anything, the variety of interesting and challenging problems ahead of us is growing more rapidly than at any time in the past and should prove sufficient to keep the community actively involved for years to come.

References

1. Cornell University Library. The making of America collection. http://moa.cit.cornell.edu/
2. American Memory from the Library of Congress. http://memory.loc.gov/
3. Princeton University Library Papyrus Home Page. http://www.princeton.edu/papyrus/
4. Baird H, Popat K (2002) Human interactive proofs and document image analysis. In: Proceedings of the 5th international workshop on document analysis systems. Lecture notes in computer science, vol 2423. Springer, Berlin Heidelberg New York, pp 507–518
5. Cattoni R, Coianiz T, Messelodi S, Modena CM (1998) Geometric layout analysis techniques for document image understanding: a review. Technical Report 9703-09/1998, ICT-IRST
6. Doermann D (1998) The indexing and retrieval of document images: a survey. Comput Vision Image Understand 3(70):287–298
7. Liang J, Rogers R, Haralick R, Phillips I (1997) UW-ISL document image analysis toolbox: an experimental environment. In: Proceedings of the 4th international conference on document analysis and recognition, Ulm, Germany, August 1997, pp 984–988
8. Lopresti D, Hu J, Kashi R (eds) (202) In: Proceedings of the 5th international workshop on document analysis systems. Lecture notes in computer science, vol 2423. Springer, Berlin Heidelberg New York. `http://link.springer.de/link/service/series/0558/tocs/t2423.htm`
9. Nagy G (2000) Twenty years of document image analysis in PAMI. IEEE Trans Patt Anal Mach Intell 1(22):38–62
10. Sauvola J, Kauniskangas H (1998) MediaTeam document database II. CD-ROM collection of document images. University of Oulu, Finland. `http://www.mediateam.oulu.fi/MTDB/index.html`
11. Taghva K, Nartker T, Borsack J, Condit A (1999) UNLV-ISRI document collection for research in OCR and information retrieval. Technical Report 99-01/1999, Information Science Research Institute, University of Nevada, Las Vegas
12. Wang Y, Hu J (2002) Detecting tables in HTML documents. In: Proceedings of the 5th internationl workshop on document analysis systems. Lecture notes in computer science, vol 2423. Springer, Berlin Heidelberg New York, pp 249–260