Content Recognition and Indexing in the LiveMemory Platform

Rafael Dueire Lins, Gabriel Torreão, and Gabriel Pereira e Silva

Universidade Federal de Pernambuco, Recife - PE, Brazil
rdl@ufpe.br, gabrieltorreao@gmail.com, gfps@cin.ufpe.br

Abstract. The proceedings of many technical events in different areas of knowledge witness the history of the development of that area. LiveMemory is a user friendly tool developed to generate digital libraries of event proceedings. This paper describes the module designed to perform content recognition in LiveMemory.

Keywords: Digital libraries, image indexing, content extraction.

1 Introduction

LiveMemory is a software platform designed to generate digital libraries from proceedings of technical events. Until today, only very few prestigious events have proceedings printed and widely distributed by international publishing houses. Thus, copies of the proceedings are restricted to those who attended the event. In this case, past proceedings are difficult to obtain and very often disappear; bringing gaps into the history of the evolution of events and even research areas. The digital version of proceedings, which started to appear at the end of the 1990's, possibly made things even worse. Only conference attendees were able to obtain copies of the CDs of the proceedings. LiveMemory was used to generate a digital library released in a DVD containing the whole history of the 25 years of the proceedings of the Symposium of the Brazilian Telecommunications Society, the most relevant academic event in the area in Latin America. The problems faced in the generation of the SBrT digital library ranged from compensating paper aging effects, filtering back-to-front noise [5], correcting page orientation and skew during scanning, to image binarization and compression. LiveMemory merges together proceedings that were scanned and volumes that were already in digital form. The SBrT'2008 digital library was organized per year of the event.

This paper outlines the functionality of the LiveMemory platform in general and addresses the way it recognizes the contents of the pages, making possible general indexing of documents and better access to the information in the library. This module works by getting information from two different sources. The first one is the image of the pages of the "Table of Contents" of the volume. The second one is each paper page image. Besides those pages there are introductory pages such as the history of the event, the address of the volume editor, etc. There may also be track or session separation pages, remissive index, etc. Pages are segmented to find the block areas which correspond to the information and then transcribed via OCR. The

J.-M. Ogier, W. Liu, and J. Lladós (Eds.): GREC 2009, LNCS 6020, pp. 220–230, 2010.

transcription of the blocks of the Table of Contents and headings of papers are cross analyzed to generate the entries of the navigation index (hyperlinks) in the digital library. It is important to remark that the volumes of SBrT varied widely in layout from one year to another, or even within the same volume, as most of those volumes were typewritten according to "loose" requirements stated that each year editor at a time there were no word processors. Even the page numbering systems adopted varied from one year to another. Some volumes are numbered with Indo-Arabic numerals throughout, some others use Roman numerals in introductory pages, there are volumes that are split into "sessions" or "tracks" and each paper gets a numbering according to its position in there. The title and page number segmentation process was developed in MatLab© and correctly spotted the required information in almost 100% of times. In the cross reference system, that information was checked against the transcription of the pages of the Table of Contents and in case of inconsistent information the priority is given to the index in the calculus of page attributes.

This paper is organized as follows. In the next section one provides a brief overview of the features of the LiveMemory platform. Section 3 details the page content functionality of the platform. The information cross-reference modulo is described in Section 4. The concluding section details the results obtained for the content detection module in LiveMemory in the development of the SBrT Digital Library, presents the conclusions and draws lines for further work.

2 LiveMemory Image Pre-processing Routines

The top-level interface of the LiveMemory platform allows the user to generate the opening screen of the proceedings to be generated. In that screen, the user provides the information of the number of volumes to be inserted. The LiveMemory environment automatically builds the hierarchy of directories for the different volumes. The user may also provide a wallpaper image to the screen and an opening soundtrack to be played when the library is accessed. The user must provide information of which volumes are already in digital form and which volumes are originally in paper. In the previous version of LiveMemory the only entry to the library is through the top menu that provides buttons to volumes. To improve that situation a few difficulties need to be overcome. The volumes that were originally in digital form use several different technologies. Some volumes are one large pdf file where all pages/articles appear one after another. Some others are structured/browsable pdf files where each article has an entry in the index. Some volumes have some search and indexing software that point at pdf files. Some other volumes are encapsulated Flash or database protected files. Being able to "unstructured" all the available data to generate a global library index or re-index by author or keywords them is far from being a trivial task, which is considered out of the scope of this paper.

This section outlines the image processing functionalities in LiveMemory. All printed proceedings are scanned in true color with a resolution of 200 dpi and stored in uncompressed bmp file format. The scanned images are loaded in a directory that corresponds to the year of the event. LiveMemory is targeted at non-experts in image processing, thus the image processing part is as automated as possible and asks for no

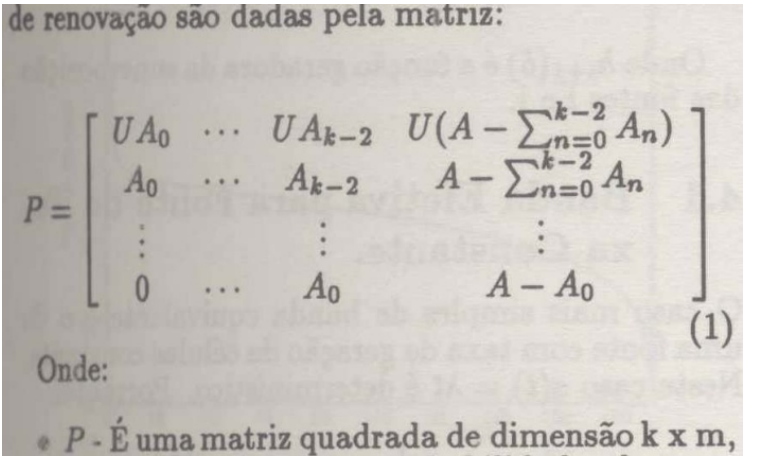
parameter input. The set of tools to suitably filter images encompasses the following routines:

- content identification,
- image binarization,
- noise border removal,
- orientation and skew correction,
- page size normalization,
- salt-and-pepper filtering, and
- image compression in Tiff_G4 file format.

Content identification for index generation is explained in the next section. The most important image processing routines are outlined below. LiveMemory makes use of some of the functionalities of BigBatch [4] a platform to process monochromatic documents. Similarly, to BigBatch, the document process interface may work in user driven or batch modes.

2.1 Image Binarization

Monochromatic images claim much less space than their color equivalent, are much faster loaded for visualization, need less toner for printing, etc. Most proceedings were printed in black-and-white. Thus, it is advantageous to have the pages in their monochromatic version, whenever possible. One phenomenon observed in several of the proceedings digitized by the authors to the SBrT Digital Library is that several volumes exhibit a light back-to-front interference [5], also known as bleeding or showtrough. Fig.1 zooms into a part of a page of a volume of SBrT with such noise. To minimize such phenomenon, LiveMemory successfully uses an entropy based binarization algorithm that was designed to remove back-to-front interference in historical documents [5].

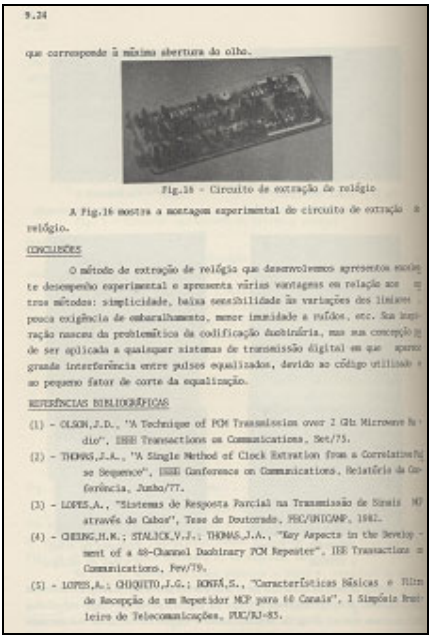

de renovação são dadas pela matriz:

$$P = \begin{bmatrix} UA_0 & \cdots & UA_{k-2} & U(A - \sum_{n=0}^{k-2} A_n) \\ A_0 & \cdots & A_{k-2} & A - \sum_{n=0}^{k-2} A_n \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & A_0 & A - A_0 \end{bmatrix} \quad (1)$$

Onde:

- P - É uma matriz quadrada de dimensão k x m,

Fig. 1. Part of a document with light back-to-front noise



9.24

que corresponde à máxima abertura do olho.

Fig.15 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração do relógio.

CONCLUSÕES

O método de extração de relógio que desenvolvemos apresentou excelente desempenho experimental e apresenta várias vantagens em relação aos outros métodos: simplicidade, baixa sensibilidade às variações dos limiares, pouca exigência de casamento, menor imunidade a ruídos, etc. Sua inspiração nasceu da problemática da codificação duobinária, mas sua concepção pode ser aplicada a qualquer sistemas de transmissão digital em que apareça grande interferência entre pulses equalizados, devido ao código utilizado e ao pequeno fator de corte da equalização.

REFERÊNCIAS BIBLIOGRÁFICAS

(1) - OLSON,J.D., "A Technique of PCM Transmission over 2 Gh: Microwave Radio", IEEE Transactions on Communications, Set/75.

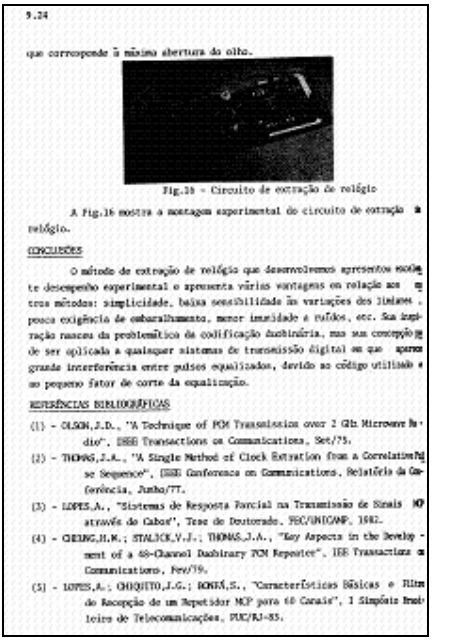
(2) - THOMAS,J.A., "A Single Method of Clock Extration from a Correlative Pulse Sequence", IEEE Conference on Communications, Relatório da Conferência, Junho/77.

(3) - LOPES,A., "Sistemas de Resposta Parcial na Transmissão de Sinais PCM através do Cabo", Tese de Doutorado, FEC/UNICAMP, 1982.

(4) - CHEUNG,H.M.; STALICK,V.J.; THOMAS,J.A., "Key Aspects in the Development of a 48-Channel Duobinary PCM Repeater", IEE Transactions on Communications, Fev/79.

(5) - LOPES,A.; CHIQUITO,J.G.; ROMÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Fig. 2. Proceedings page with photo in true-color Size: 431 kB - JPG, 435kB - pdf



9.24

que corresponde à máxima abertura do olho.

Fig.15 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração do relógio.

CONCLUSÕES

O método de extração de relógio que desenvolvemos apresentou excelente desempenho experimental e apresenta várias vantagens em relação aos outros métodos: simplicidade, baixa sensibilidade às variações dos limiares, pouca exigência de casamento, menor imunidade a ruídos, etc. Sua inspiração nasceu da problemática da codificação duobinária, mas sua concepção pode ser aplicada a qualquer sistemas de transmissão digital em que apareça grande interferência entre pulses equalizados, devido ao código utilizado e ao pequeno fator de corte da equalização.

REFERÊNCIAS BIBLIOGRÁFICAS

(1) - OLSON,J.D., "A Technique of PCM Transmission over 2 Gh: Microwave Radio", IEEE Transactions on Communications, Set/75.

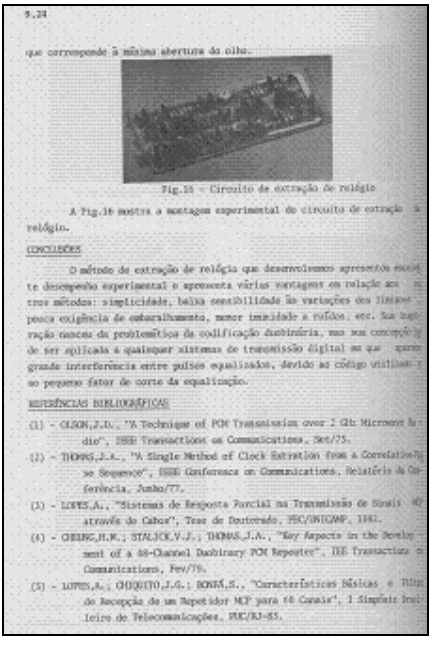
(2) - THOMAS,J.A., "A Single Method of Clock Extration from a Correlative Pulse Sequence", IEEE Conference on Communications, Relatório da Conferência, Junho/77.

(3) - LOPES,A., "Sistemas de Resposta Parcial na Transmissão de Sinais PCM através do Cabo", Tese de Doutorado, FEC/UNICAMP, 1982.

(4) - CHEUNG,H.M.; STALICK,V.J.; THOMAS,J.A., "Key Aspects in the Development of a 48-Channel Duobinary PCM Repeater", IEE Transactions on Communications, Fev/79.

(5) - LOPES,A.; CHIQUITO,J.G.; ROMÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Fig. 3. Monochromatic version of Figure 02 Size: 122kB - Tiff, 351kB-pdf



9.24

que corresponde à máxima abertura do olho.

Fig.15 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração do relógio.

CONCLUSÕES

O método de extração de relógio que desenvolvemos apresentou excelente desempenho experimental e apresenta várias vantagens em relação aos outros métodos: simplicidade, baixa sensibilidade às variações dos limiares, pouca exigência de casamento, menor imunidade a ruídos, etc. Sua inspiração nasceu da problemática da codificação duobinária, mas sua concepção pode ser aplicada a qualquer sistemas de transmissão digital em que apareça grande interferência entre pulses equalizados, devido ao código utilizado e ao pequeno fator de corte da equalização.

REFERÊNCIAS BIBLIOGRÁFICAS

(1) - OLSON,J.D., "A Technique of PCM Transmission over 2 Gh: Microwave Radio", IEEE Transactions on Communications, Set/75.

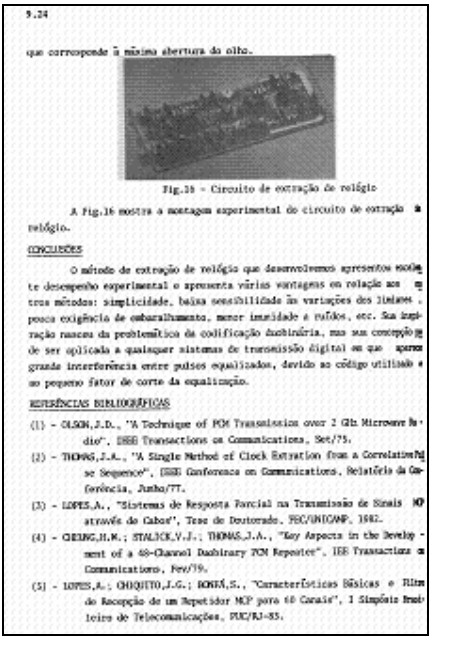
(2) - THOMAS,J.A., "A Single Method of Clock Extration from a Correlative Pulse Sequence", IEEE Conference on Communications, Relatório da Conferência, Junho/77.

(3) - LOPES,A., "Sistemas de Resposta Parcial na Transmissão de Sinais PCM através do Cabo", Tese de Doutorado, FEC/UNICAMP, 1982.

(4) - CHEUNG,H.M.; STALICK,V.J.; THOMAS,J.A., "Key Aspects in the Development of a 48-Channel Duobinary PCM Repeater", IEE Transactions on Communications, Fev/79.

(5) - LOPES,A.; CHIQUITO,J.G.; ROMÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Fig. 4. Versions of **Fig.3.** Size: 3.03 kB - Tiff and 230 kB – pdf.



9.24

que corresponde à máxima abertura do olho.

Fig.15 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração do relógio.

CONCLUSÕES

O método de extração de relógio que desenvolvemos apresentou excelente desempenho experimental e apresenta várias vantagens em relação aos outros métodos: simplicidade, baixa sensibilidade às variações dos limiares, pouca exigência de casamento, menor imunidade a ruídos, etc. Sua inspiração nasceu da problemática da codificação duobinária, mas sua concepção pode ser aplicada a qualquer sistemas de transmissão digital em que apareça grande interferência entre pulses equalizados, devido ao código utilizado e ao pequeno fator de corte da equalização.

REFERÊNCIAS BIBLIOGRÁFICAS

(1) - OLSON,J.D., "A Technique of PCM Transmission over 2 Gh: Microwave Radio", IEEE Transactions on Communications, Set/75.

(2) - THOMAS,J.A., "A Single Method of Clock Extration from a Correlative Pulse Sequence", IEEE Conference on Communications, Relatório da Conferência, Junho/77.

(3) - LOPES,A., "Sistemas de Resposta Parcial na Transmissão de Sinais PCM através do Cabo", Tese de Doutorado, FEC/UNICAMP, 1982.

(4) - CHEUNG,H.M.; STALICK,V.J.; THOMAS,J.A., "Key Aspects in the Development of a 48-Channel Duobinary PCM Repeater", IEE Transactions on Communications, Fev/79.

(5) - LOPES,A.; CHIQUITO,J.G.; ROMÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Fig. 5. LiveMemory Versions of **Fig.5.** Size: 3.03 kB - Tiff and 343 kB - pdf.

Very often, paper pages incorporate graphical elements such as photos, figures, and graphs that are printed using dithering techniques in such a way that resemble gray scale images, although printed in black and white. Figure 2 provides an example of such a page, also with some back-to-front noise. The direct binarization of such pages does not yield satisfactory graphical results as may be observed in Figure 3. The conversion of page with photos, figures and graphs into gray scale provides a reasonable alternative in size, but introduces non-uniform pages into the volume as the majority of pages are monochromatic for the sake of space and readability. LiveMemory image processing module automatically sweeps the directory of scanned images from a volume looking for pages that encompass graphical elements. These pages are found by using projection profile both in the horizontal and vertical directions. Pages whose projection presents large contiguous areas indicate the presence of graphical elements. The projections allow splitting pages into blocks, which are tagged. Similar blocks are merged together. In such way, LiveMemory decomposes pages into text and graphical elements. Text areas are binarized. The graphical elements are converted from true color into gray scale. Figure 5 provides an example of such synthetic image which, although it brings no gain in space, if compared with gray scale, it is uniform to the reader as there is no difference in the text areas from the other pages in the volume. Layout analysis is performed in the different kinds of paper pages to identify the fields of interest with the aid of an OCR platform.

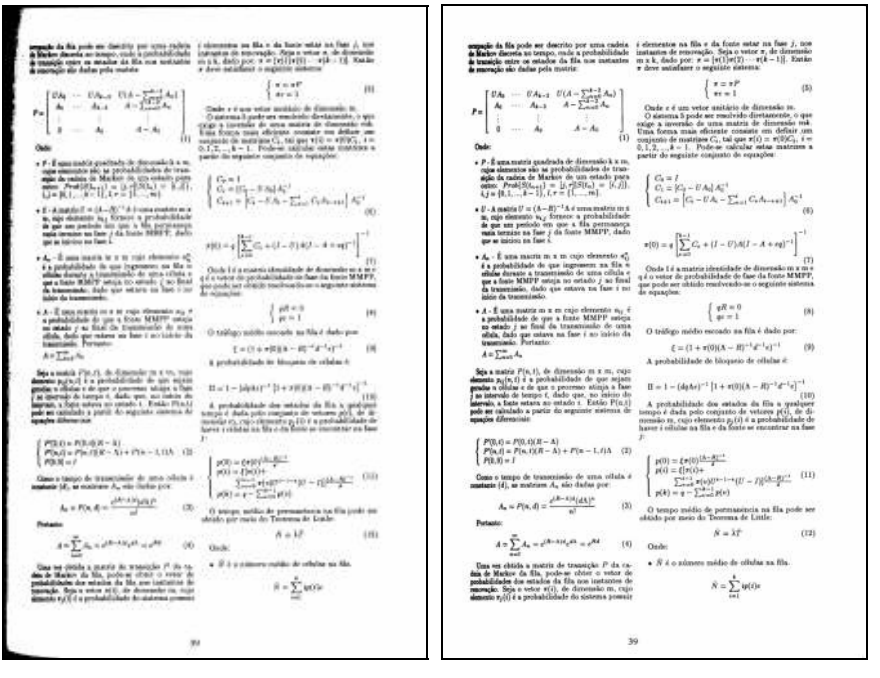


Fig. 6. Page with and without black border

2.2 Black Border Removal

As one may observe in the case of the page shown on the left hand side of Figure 6, the monochromatic version of the document exhibits a black border on its left margin. This border is the result of the uneven illumination of the scanning process due to volume binding. The same phenomenon appears, for different reasons whenever the volume of the proceedings is unbound and the loose pages are scanned using a production line automatically fed monochromatic scanner. The difference between the two cases aforementioned is that in the former the black border is within the document area, while in the latter case of automatically fed monochromatic scanners the noise surrounds the document. The right hand side of Figure 6 presents the same document of the left hand side with the black noisy border removed. The algorithm used in LiveMemory for black border removal is described in reference [1].

3 Paper Preparation

The experience with the digital volumes integrated into the SBrT digital library showed that, in general, there are standard layouts in the articles in one proceeding volume and that editors were careful enough to include headings with title and data of the authors. This information may be used for indexing articles and volumes in a similar way to the one proposed in reference [6].

A volume of proceedings has a somehow standard format that may be split into four parts:

- Volume presentation.
- Table of Contents.
- Papers.
- Remissive Index (optional).

The volume presentation frequently encompasses a title page, a forward (or preface) by the conference chairperson, the list of people on the program committee and other optional items. The Table of Contents is a list of authors, paper title, and page numbers. In general, roman numerals are used for page numbering the Volume presentation and the Table of Contents parts. Some conferences that use the Track format structure their proceedings differently, as:

- Volume presentation.
- Table of Contents.
- Track *1* (Track presentation+Papers)...Track *n* (Track presentation+Papers).
- Remissive Index (optional).

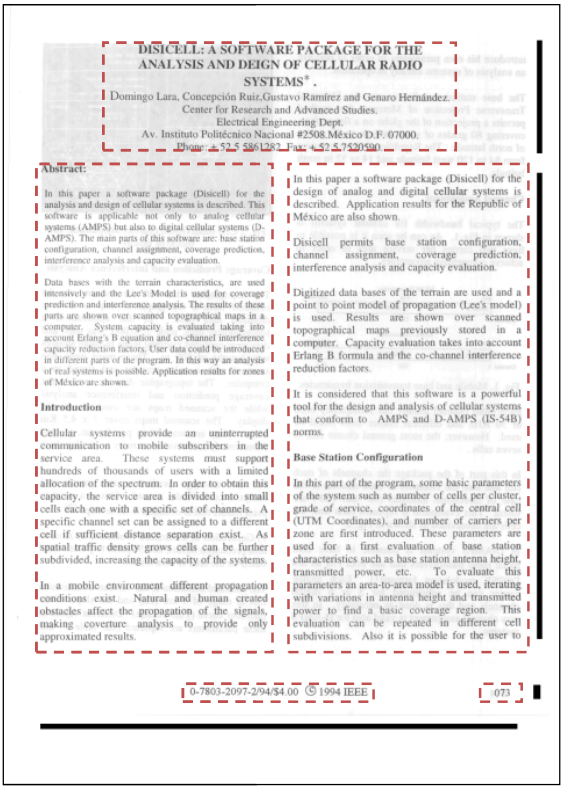
In this version of LiveMemory, the user provides information of the kind of structuring used in each volume. The papers themelves encompass front or title and content pages. The front pages of papers include:

- Paper Title.
- List of authors and affiliation.
- Abstract (or summary).
- Abstract in a foreign language (optional).
- List of keywords (optional).
- Classification indices (optional).

Identifying all these elements allows a complete navigation in the contents of papers.

3.1 Block Image Segmentation and Classification

LiveMemory segmentation algorithm uses projection profile to iteratively split the page in blocks in a top-down fashion. At first, one takes the projection profile of the whole page as the example shown in Figure 7, where one finds information blocks.


DISICELL: A SOFTWARE PACKAGE FOR THE ANALYSIS AND DEIGN OF CELLULAR RADIO SYSTEMS*.

Domingo Lara, Concepción Ruiz,Gustavo Ramírez and Genaro Hernández.
Center for Research and Advanced Studies.
Electrical Engineering Dept.
Av. Instituto Politécnico Nacional #2508,México D.F. 07000.
Phone: +52 5 5061282 Fax: +52 5 7420900

Abstract:

In this paper a software package (Disicell) for the analysis and design of cellular systems is described. This software is applicable not only to analog cellular systems (AMPS) but also to digital cellular systems (D-AMPS). The main parts of this software are: base station configuration, channel assignment, coverage prediction, interference analysis and capacity evaluation.

Data bases with the terrain characteristics, are used intensively and the Lee's Model is used for coverage prediction and interference analysis. The results of these parts are shown over scanned topographical maps in a computer. System capacity is evaluated taking into account Erlang's B equation and co-channel interference capacity reduction factors. User data could be introduced in different parts of the program. In this way an analysis of real systems is possible. Application results for zones of México are shown.

Introduction

Cellular systems provide an uninterrupted communication to mobile subscribers in the service area. These systems must support hundreds of thousands of users with a limited allocation of the spectrum. In order to obtain this capacity, the service area is divided into small cells each one with a specific set of channels. A specific channel set can be assigned to a different cell if sufficient distance separation exist. As spatial traffic density grows cells can be further subdivided, increasing the capacity of the systems.

In a mobile environment different propagation conditions exist. Natural and human created obstacles affect the propagation of the signals, making coverture analysis to provide only approximated results.

In this paper a software package (Disicell) for the design of analog and digital cellular systems is described. Application results for the Republic of México are also shown.

Disicell permits base station configuration, channel assignment, coverage prediction, interference analysis and capacity evaluation.

Digitized data bases of the terrain are used and a point to point model of propagation (Lee's model) is used. Results are shown over scanned topographical maps previously stored in a computer. Capacity evaluation takes into account Erlang B formula and the co-channel interference reduction factors.

It is considered that this software is a powerful tool for the design and analysis of cellular systems that conform to AMPS and D-AMPS (IS-54B) norms.

Base Station Configuration

In this part of the program, some basic parameters of the system such as number of cells per cluster, grade of service, coordinates of the central cell (UTM Coordinates), and number of carriers per zone are first introduced. These parameters are used for a first evaluation of base station characteristics such as base station antenna height, transmitted power, etc. To evaluate this parameters an area-to-area model is used, iterating with variations in antenna height and transmitted power to find a basic coverage region. This evaluation can be repeated in different cell subdivisions. Also it is possible for the user to

0-7803-2097-2/94$4.00 © 1994 IEEE

073

Fig. 7. – Front page block decomposition

Each of those blocks are then recursively split into sub-blocks until reaching blocks that envelope each line. If 25% of blocks are aligned vertically then that alignment point is a column margin. The analysis of the height, width and the position within the page and in relation to the other blocks in it allows one to re-merge similar blocks and to decide about their nature. For instance, in general the block that contains the title of an article is on the top-center of a page, not aligned with the column margins. A minimum width parameter is also used, to distinguish between a title and a page numbering. In the case of the paper proceedings of SBrT, no page heading had any other information besides page number. Similarly, the blocks on the bottom of a page generally contain information about the paper and page number. In the case of the paper shown in Figure 7 the block on the

top of the page contains the paper title, authors' names and affiliations. On the bottom of the page, one finds two blocks. The central one contains "0-7803-2097-2/94/$4.00 ©1994 IEEE" (every four years the SBrT annual event is called the International Telecommunications Symposium, an IEEE event) and the rightmost one presents "073", the page number in the proceedings volume. One may also observe that the page of Figure 7 also has back-to-front interference, which is more pronounced on the left hand side of the title block.

The segmentation and classification of the information on the Table of Contents is aided by user provided information on its general layout.

4 Automatic Categorization

Automatic categorization (or classification) of textual information in pre-defined classes is a research area of rising importance due to the ever growing availability of documents in digital format, thus a greater need of organizing them. A common approach to address such problem is based on machine learning: frequently an inductive process automatically generates a classifier having as starting point a set of pre-classified documents in each of the categories of interest. The advantages of such strategy in relation to the knowledge engineering one (which consists in experts in the subject manually defining a classifier) is due to its performance once it brings a considerable economy of efforts of the experts in the field, besides providing systematic grounds for extensibility, allowing more easily to address new domains in a much faster way. The LiveMemory platform follows the machine learning approach for text categorization, but also makes use of some artifices that yield an efficiency rise of such a process.

4.1 Document Organization

Indexing documents with a controlled dictionary is an example of the more general problem of database organization. Often, several other problems relative to document filing either for personal organization or in structuring a corporative database may be solved using text categorization techniques. An example of such may be found in the interesting paper in reference [9], which addresses the problem of efficiently finding patents of different categories with a high accuracy rate. Another example is the automatic fulfilment of columns in a newspaper (such as Economics, Health, International Politics, etc.). LiveMemory attempts to group together the papers in proceedings according to its concentration area. In the case of a telecommunications conference, for instance, one has mobile communications, satellite communications, computer networks, cryptography, etc. To make searches in a proceedings database in which the paper subject has been previously annotated is far more precise and faster. Very often, papers have no keywords and have to be swept to find them in their abstract or even the paper body, with the aid of a subject dictionary.

4.2 Document Organization

Text categorization dates back to 1961 with the work reported in [10] about the probabilistic text classification. Since then it has been used in a large number of different applications. LiveMemory is a tool that aims to work with proceedings independly of the language they are written. In the specific case of the proceedings of SBrT, there are papers in English, Portuguese and Spanish. Thus one has two different outlooks to cluster the papers into: one is in the language the paper is written in and another is its subject area. In the case of the SBrT digital library, priority is given to the paper concentration area, using as guidance the information within the volume such as tables of contents, indices, separation pages, etc. Even section ordering and conference schedule that often appears in some volumes are used to infer information about the content of papers, as chairpersons tend to organize sections with related papers. All information gathered must be cross-checked with abstracts and sometimes introductory and/or concluding sections. A different research concern, but somehow related to the one reported, focuses in finding document authorship and detecting plagiarism.


Table of Contents

I Structural Approaches for Recognition and Indexing

Use of Perceptive Vision for Ruling Recognition in Ancient Documents 3
A. Lemaitre, B. Coüasnon, J. Camillerapp

Evaluation of Graph Matching Measures for Documents Retrieval 13
S. Jouili, S. Tabbone, E. Valveny

Employing fuzzy intervals and loop-based methodology for designing structural signature: an application to symbol recognition 22
MM. Luqman, M. Delalandre, T. Brouard, JY. Ramel, J. Lladós

Interactive Conversion of Large Web Tables 32
R. Padmanabhan, RC. Jandhyala, M. Krishnamoorthy, G. Nagy, S. Seth, W. Silversmith

Embedding labeled graphs into occurence matrix 44
N. Sidère, P. Héroux, JY. Ramel

Web Document Generation via XML based Magazine Structure Analysis 51
AY. Kim, J. Park, YB. Kwon

II Techniques Towards Vectorization

Detection of Circular Arcs in a Digital Image Using Chord and Sagitta Properties 59
S. Bera, P. Bhowmick, BB. Bhattacharya

GOAL: Towards understanding of Graphic Objects from Architectural to Line drawings 71
S. Pal, P. Bhowmick, A. Biswas, BB. Bhattacharya

Automatic Road Vectorization of Raster Maps 83
YY. Chiang, CA. Knoblock

Robust Circular Arc Detection 85
B. Lamiroy, Y. Gucbhas

Automatic Palette Identification of Colored Graphics 95
V. Lacroix

III Sketching Interfaces, On-line Processing

Benchmarks for Computer-based Segmentation of Sketches 103
P. Company, PAC. Varley, A. Piquer, M. Vergara, J. Sánchez-Rubio

QuickDiagram: A System for Online Sketching and Understanding Diagrams 115
L. Wenqin, Y. Wang, CY. Ho, T. Lu, Z. Sun

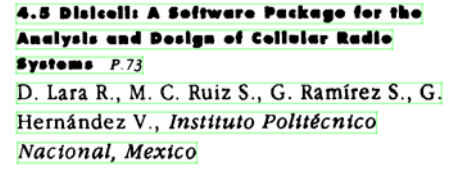
SSP: Sketching Slide Presentations, a Syntactic Approach 121
J. Mas, G. Sanchez, J. Lladós

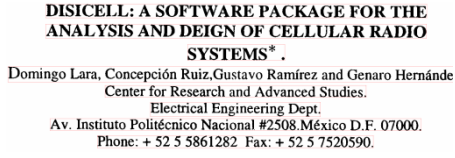
Fig. 8. Block segmentation of a page

5 The Table of Contents Generator

Having ways to fast navigate in digital libraries is mandatory. The blocks of interest spotted during the segmentation process shown above are transcribed via OCR. This information is used for indexing articles and volumes. The index generating module of LiveMemory takes the set the transcription of the Table of Contents pages from a volume and tries to match a formation rule of a regular expression to find the "page_number". The Java library for regular expression parsing was used to create the parser generator.

Each image that corresponds to a volume page is segmented to find its number and title blocks, which are transcribed using the Tesseract OCR. This information becomes attributes of the image. Figure 9 shows the results of block segmentation from the Table of Contents and of the paper title page for the article shown in Figure 8. As one may observe, the information in the two blocks are not the same. Even the title does not fully coincide as in the paper title there is a spelling mistake, corrected by the volume editor in the Table of Contents. Now, the system tries to unify the page_number information with the page attributes. The title pages are the key for the image and contents matching.


4.5 Dlslcalls: A Software Package for the Analysis and Design of Cellular Radio Systems P.73
D. Lara R., M. C. Ruiz S., G. Ramírez S., G. Hernández V., Instituto Politécnico Nacional, Mexico


DISICELL: A SOFTWARE PACKAGE FOR THE ANALYSIS AND DEIGN OF CELLULAR RADIO SYSTEMS* .
Domingo Lara, Concepción Ruiz,Gustavo Ramírez and Genaro Hernánde
Center for Research and Advanced Studies.
Electrical Engineering Dept.
Av. Instituto Politécnico Nacional #2508.México D.F. 07000.
Phone: + 52 5 5861282 Fax: + 52 5 7520590.

4.5 Dlslcalls A Sollwara Package far lha Analysls and Daslgn al Câ€¢IIÃ¤Iar lalla lyslans P. 73 D. Lara R._ M. C. Ruiz S., G. Ramirez S., G. HcmÃ©ndcz V., Instituto PolitÃ©cnico Nacional, Mexico	DISICELL: A SOFTWARE PACKAGE FOR THE ANALYSIS AND DEIGN OF CELLULAR RADIO SYSTEMS* . Domingo Lara, Concepcion Ruiz,Gustavo Ramirez and Genaro Hernandez. Center for Research and Advanced Studies. Electrical Engineering Dept. Av. Instituto PolitÃ©cnico Nacional #2508.MÃ©xico D.F. 07000. Phone: + 52 5 5861282 Fax: + 52 5 7520590.

Fig. 9. Top: (Left) Information from Table of Contents; (Right) Paper Title; Bottom – Tesseract© transcriptions

The top-left part of Figure 9 presents an image block extracted from the Table of Contents of a volume. Its automatic transcription performed by the Tesseract OCR is shown immediately below. The use of regular expressions has enabled to spot the page number in the volume. That information was used to find the corresponding image file by offsetting the list_of_filenames. The segmentation process shown in the last section is able to find the image of the paper_title_block as shown in the top of right hand side.

Another aiding element is provided by the image filenames: they follow a strict numerical order. This means that the image filenames follow a pattern such as volume_year_page_number. For instance, the first image scanned of the 1991 volume is 1991_001, the second one is 1991_002, the third page is 1991_003, etc. Then, the problem of tying hyperlinks between the table_of_contents and images becomes finding the right offset in the two lists. Unfortunately, in some volumes of the SBrT proceedings there are missing and repeated pages. This may cause unrecoverable trouble to any automatic indexing system.

6 Conclusions and Lines for Further Work

This concluding section provides some evidence of the effectiveness of the index generation scheme presented herein. A total of 613 pages of 323 papers within several volumes of the SBrT proceedings with different page, table of contents, heading and footnote layouts, typesetting and printing technologies, paper color due to aging, etc. were tested. The title blocks were correctly recognized in 96.9% of the total number of pages, while page numbers were spotted in 97.1% of cases. The linking of the "Table of Contents" to page numbers was successful in 98.3 % of the total number of papers.

LiveMemory is no doubt a valuable and user friendly platform for the generation of digital libraries of event proceedings. Its use in the case of the SBrT digital library witnesses its usability.

The automatization of the detection procedure of the layout of the pages of the Table of Contents using a classification tool such as Weka [8] is under development. Other lines for further work include automatic author and keyword searching.

Acknowledgments. Research reported herein was partly sponsored by CNPq - Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico and CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Government. The authors also express their gratitude to the SBrT – the Brazilian Telecommunications Society, for granting the permission to use the images from their proceedings.

References

1. Ávila, B.T., Lins, R.D.: A New Alg. for Removing Noisy Borders from Monochromatic Documents. In: ACM-SAC 2004, pp. 1219–1225. ACM Press, New York (2004)
2. Ávila, B.T., Lins, R.D.: A New and Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. In: ACM DocEng 2005. ACM Press, New York (2005)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice-Hall, Englewood Cliffs (2007)
4. Lins, R.D., Ávila, B.T., de Araújo Formiga, A.: BigBatch: An Environment for Processing Monochromatic Documents. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4142, pp. 886–896. Springer, Heidelberg (2006)
5. Silva, J.M.M., da Lins, R.D., da Rocha Jr., V.C.: Binarizing and Filtering Historical Documents with Back-to-Front Interference. In: ACM Symposium on Applied Computing, Dijon. Proceedings of SAC 2006, pp. 853–858. ACM Press, New York (2006)
6. van Beusekom, J., et al.: Example-Based Logical Labelling of Document Title Page Images. In: ICDAR 2007, pp. 919–924. IEEE Press, Los Alamitos (2007)
7. Tesseract, http://code.google.com/p7tesseract-ocr/
8. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
9. Larkey, L.S.: A patent search and classification system. In: Proceedings of DL 1999, 4th ACM Conference on Digital Libraries, Berkeley, CA, pp. 179–187 (1999)
10. Maron, M.: Automatic indexing: an experimental inquiry. Journal of the ACM (JACM) 8(3), 404–417 (1961)