Efficiently Generating Digital Libraries of Proceedings with The LiveMemory Platform

Rafael Dueire Lins, Gabriel de Pereira e Silva, Gabriel Torreão, Neide F. Alves
Universidade Federal de Pernambuco, Recife - PE, Brazil
rdl@ufpe.br, gabriel.psilva@ufpe.br, *gabrieltorreao@gmail.com, nfalves@uea.edu.br*

Abstract

The proceedings of many technical events in different areas of knowledge witness the history of the development of that area. This paper describes LiveMemory, a user friendly software platform designed to generate digital libraries of proceedings.

Keywords: digital libraries, proceedings, document engineering, document processing.

I. INTRODUCTION

The proceedings of technical events provide a written testimony of the development of a research area. In the past, only very few prestigious events had proceedings printed and widely distributed by international publishing houses. Thus, copies of proceedings were restricted to those who attended the event of organizing bodies. In this case, past proceedings were difficult to obtain and very often disappear; bringing gaps into the history of the evolution of events and even research areas. Digital version of proceedings made things no better, if not worse. Only conference attendees were able to obtain copies of the CDs with the proceedings. The first author of this paper pioneered the release of the proceedings of the Symposium of the Brazilian Telecommunications Society in CD in 1997. A decade later, he was the chairman of that event again and idealized to release the proceedings of SBrT´2007 in a DVD containing all proceedings since 1997. As there were no digital versions of the proceedings for 1998 and 1999, the printed versions were scanned. This was the first step for the generation of a digital library with the proceedings of the most important technical event in telecommunications of Latin America. The experience and success with the proceedings of SBrT´2007, was the basis for a bolder step: generating the proceedings of SBrT´2009 in a DVD containing the whole history of the 26 years of the event. The problems faced for the generation of this digital library that was handed for all participants of the event were huge and ranged from compensating paper aging effects, filtering back-to-front interference [5] (also known as show-trough or bleeding), and page rotation during scanning, to image compression. The experience gained in the development of the proceedings of SBrT´2009 and the volume in which this article appears, the proceedings of ITS 2010, was the seed to the development of the LiveMemory platform, described herein, a software tool developed with the aim to help to prepare digital libraries of proceedings of events.

II. DATA MODELLING

The starting point to the development of this project was to know which information is needed for the development of a database to store the information of the digital library of the proceedings. The analysis of the several volumes of the different years of the proceedings of the events of SBrT, The Brazilian Telecommunications Society, has shown the need of having five tables to store the relevant data:

- Features – contains the information about the library such as title and quantity of volumes.
- Configuration – in this table the user informs in which lines data such as paper title, authors, abstract, keywords, etc. may be found.
- Author (CodAutor) – stores the name of authors;
- Summary (CodResumo) – in this table the paper title, the publication year and keywords are stored.
- Abstract and Author – links each paper to its author(s) and abstract.

Figure 1 instantiates the relationship between the main tables, through the Entity-Relationship (ER) model.

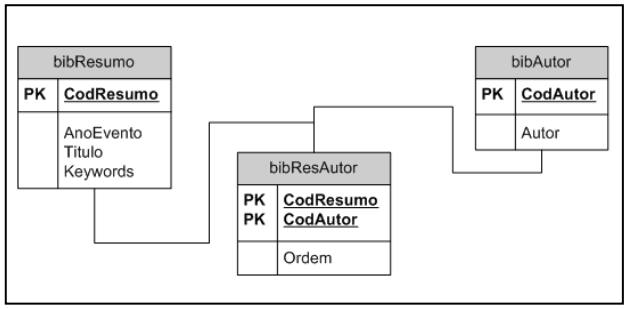


Figure 1. ER Model of an library in LiveMemory

II. THE LiveMemory TOP-LEVEL INTERFACE

Whenever LiveMemory is called the top-level interface of the platform opens the screen shown in Figure 2 and allows the user to generate a new digital library of proceedings or to carry-on the development of a already started project. If the user wants to start a new project the choice of the top button will show the screen presented in Figure 3, in which 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. Whenever the user activates LiveMemory an

editable screen such as the one presented in Figure 3 appears to be filled in.

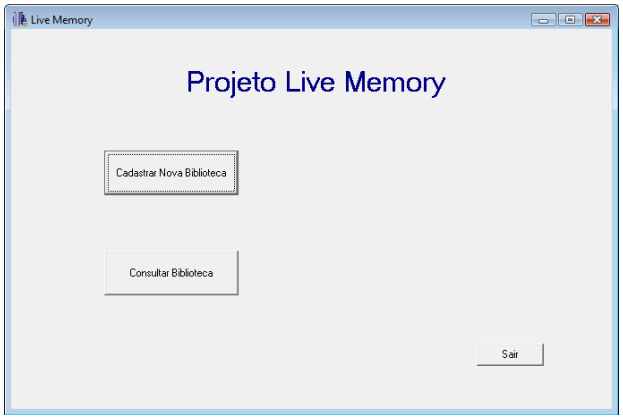


Figure 2 – LiveMemory opening screen.

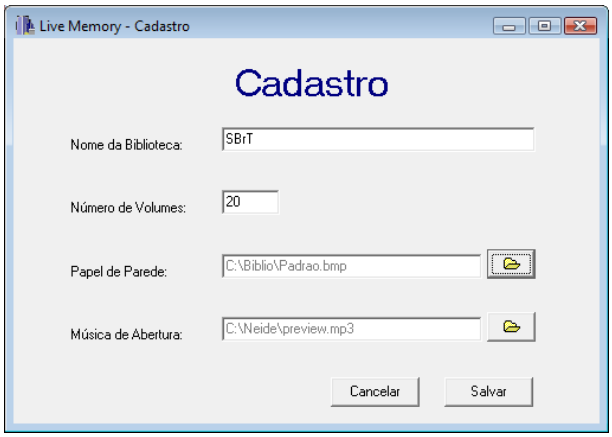


Figure 3. Live Memory opening screen.

The project may be saved by clicking on the "Salvar" button. The "Next" button generates a new screen with the data already provided by the user and asks for new data. Figure 4 shows the new screen generated by LiveMemory to be filled in by the user. For each volume the user may provide a cover image that will be inserted as a thumbnail of the active button of the volume hyperlink. If any image filtering is needed, the user may select button "Edição de Documentos".

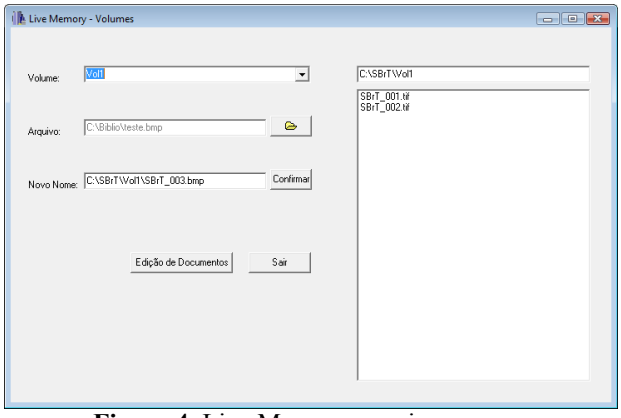


Figure 4. Live Memory opening screen.

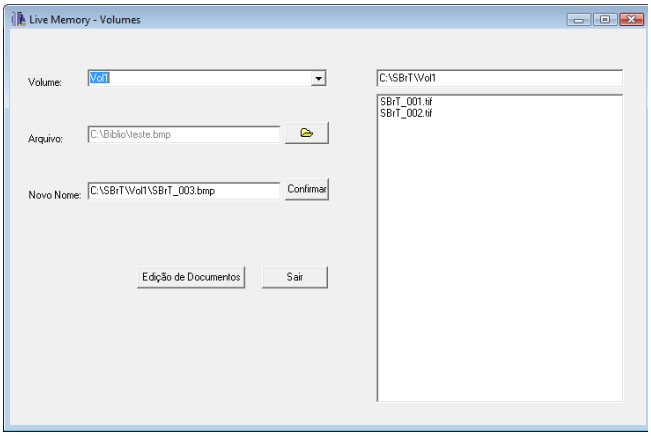


Figure 5. LiveMemory data loading customized screen.

In the case of the SBrT´09 or ITS 2010 libraries the images were either the label of the CD, if the original proceedings were on that medium, or the scanned image of the volume cover if printed. LiveMemory copies CDs of proceedings directly, keeping the original structure and features. In the case of the SBrT´2009 the proceedings of SBrT´2008 already in DVD had the proceedings of SBrT´1997 to SBrT´2008. Although the opening screens, music and video were presented as in the SBrT´2008 DVD, one needs to delete the duplicate loading of proceedings from 1997 to 2009. The current version of LiveMemory does not check for this redundancy and the user has to edit the directory structure manually. As this sort of cumulative library building becomes more frequent, LiveMemory will be later modified to make automatic directory sharing.

3. IMAGE PROCESSING ROUTINES

Whenever the proceedings are printed, pages need to be scanned and processed. All printed proceedings are scanned in true color with a resolution of 200 dpi and stored in uncompressed bmp file format. Filenames should coincide with page numbers. For instance, page 108 of the scanned volume should be named "p_108". Very often index pages receive roman numbering. Those pages should be named as "o_12", for instance. Other numbering schemes were also found. The image naming process should be performed in such a way that the list of filenames follows the same order as pages appear. Besides that, names should be consistent with the indexing scheme adopted in the volume, allowing a unique file-page identification to permit later automatic index generation. Each volume should be named in an unambiguous way. The directory with all pages of the volume is linked to the volume name and thumbnail image as shown in the screen of Figure 5. When the user finishes the loading process and clicks the "Next" button a new screen is opened as shown in Figure 6.

The "Save" button allows the user to save the current contents of the library, while "Exit" leaves the library edition process. Each loaded data volume may be seen by clicking on the "View data" button, which opens a new window. The "Process" button opens a new window with

tools to suitably filter images, they encompass the following routines: 1. Content identification; 2. Image binarization; 3. Noise border removal; 4. Skew correction; 5. Page size normalization; 6. Salt-and-pepper filtering; 7. Image compression in Tiff_G4 file format.

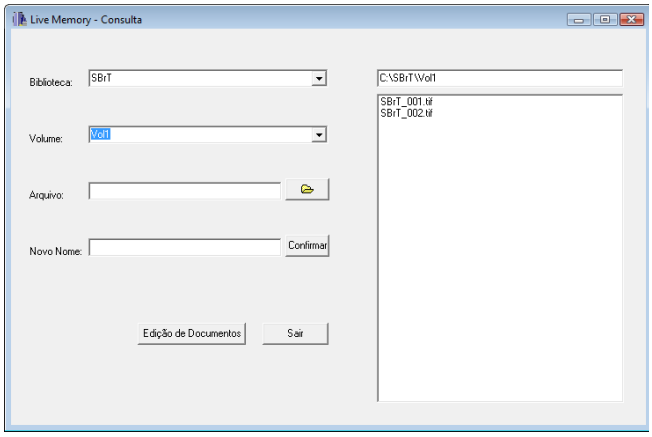


Figure 6. LiveMemory data checking screen

Content identification is explained in the next section. The other most important of those routines are detailed 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. Figure 07 presents a screen shot of the document processing window of LiveMemory working in "User Mode". The "User mode" allows the user to apply filters in any order, although some of them work only in binary documents. The user may "Revert" to the original image or call ImageJ [7] to freely process the page image.

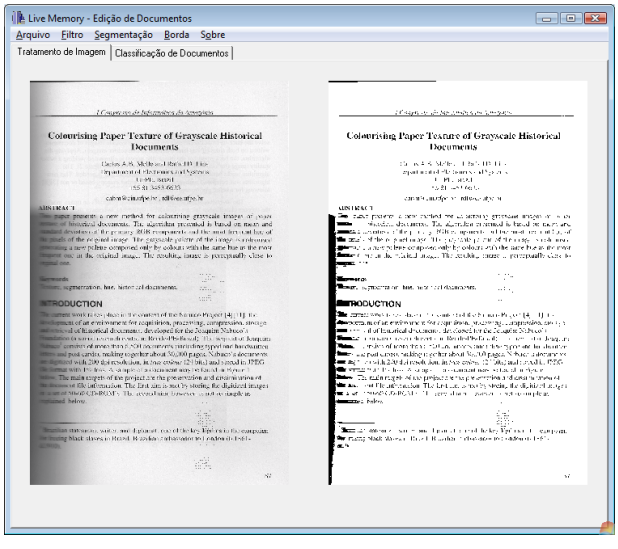
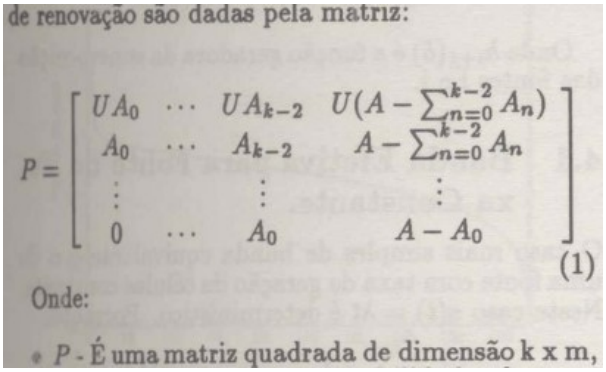


Figure 7. LiveMemory image processing screen

3.1. 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]. Figure 8 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 for 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,

Figure 8. Part of a document with back-to-front noise

3.2. Black Border Removal

As one may observe in the case of the page shown in Figure 7, 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 proceedings volume 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. Figure 9 presents a document and its version with the border removed. LiveMemory allows black border removal using the algorithm in reference [1].

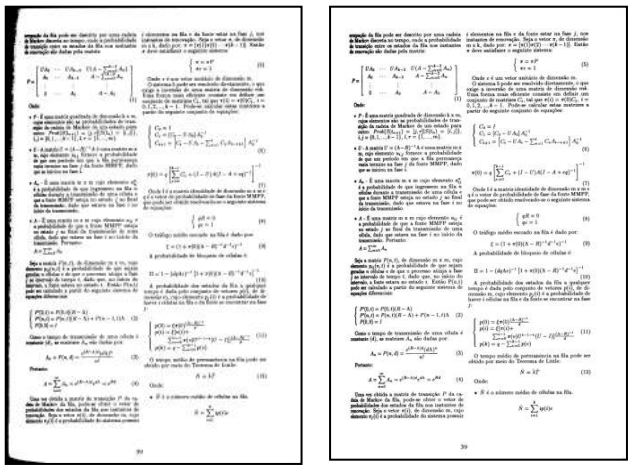


Figure 9. Page with and without black border

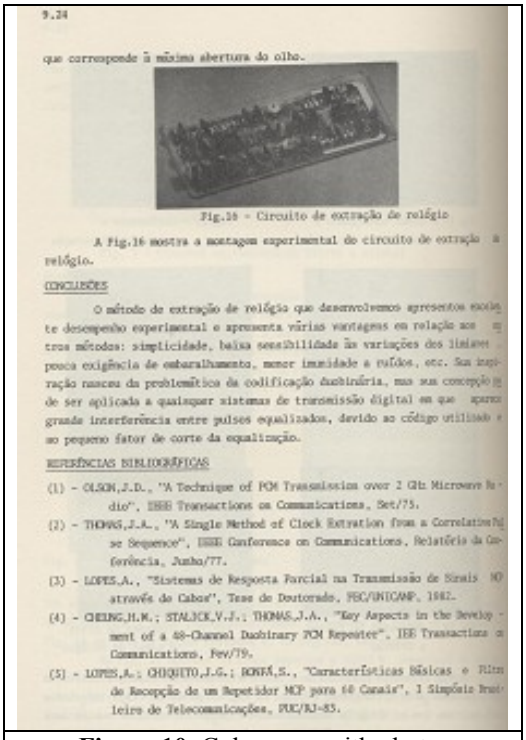
3.3. Skew Correction

Often, the scanning process performed either with automatically or manually fed scanners, yields documents with a small rotation angle that not only makes more difficult document reading, but also claims for larger storage space. The analysis of the scanning process of the SBrT library showed that the skew ranged between 1 and 3 degrees. LiveMemory uses the algorithm described in reference [2] that besides calculating the skew angle it minimizes the occurrence of uneven contours in the written parts as well as white dots in the black parts.

IV. PAGE CONTENT ANALYSIS

Very often pages from proceedings incorporate graphical elements such as photos, figures, and graphs that are printed using dithering techniques [3] in such a way that resemble grayscale, although printed in black and white. Figure 10 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 11. 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.


9.24

que corresponde à máxima abertura do olho.

Fig.16 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração a 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 embaralhamento, 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 pulsos 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.

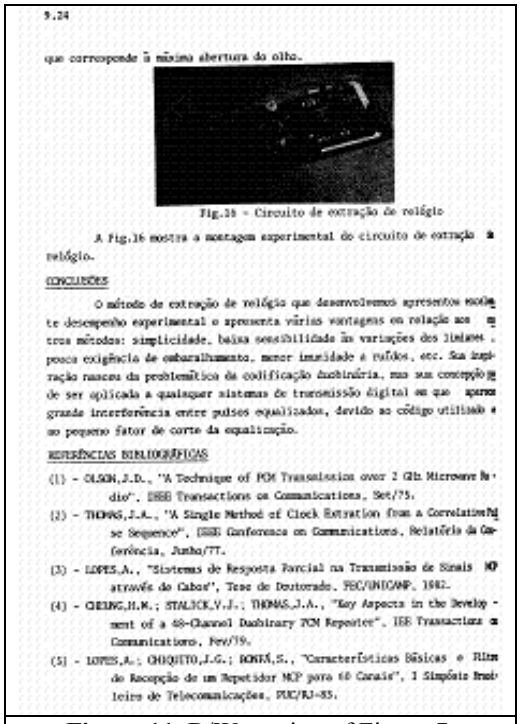
(3) - LOPES,A., "Sistemas de Resposta Parcial na Transmissão de Sinais PCM através do Cabo", Tese de Doutorado, FEC/UNICAMP, 1982.

(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.; BONFÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Figure 10. Color page with photo.
Size: 431kB – JPG, 435kB – pdf

An example of such page may be found in Figure 11. LiveMemory image processing module automatically scans 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 as explained in Section 3.1. The graphical elements are converted from true color into gray scale. Figure 10 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.


9.24

que corresponde à máxima abertura do olho.

Fig.16 - Circuito de extração do relógio

A Fig.16 mostra a montagem experimental do circuito de extração a 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 embaralhamento, 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 pulsos 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.; BONFÁ,S., "Características Básicas e Filtro de Recepção de um Repetidor MCP para 60 Canais", I Simpósio Brasileiro de Telecomunicações, PUC/RJ-83.

Figure 11. B/W version of Figure 7.
Size: 122kB – Tiff, 351kB – pdf

V. LiveMemory INDEX GENERATOR

Having ways to fast navigate in digital libraries is mandatory. The previous version of LiveMemory the only entry to the library is through the top menu that provides buttons to volumes. To improve the current situation a few difficulties need to be overcome. The volumes scanned may be transcribed via OCR. 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.

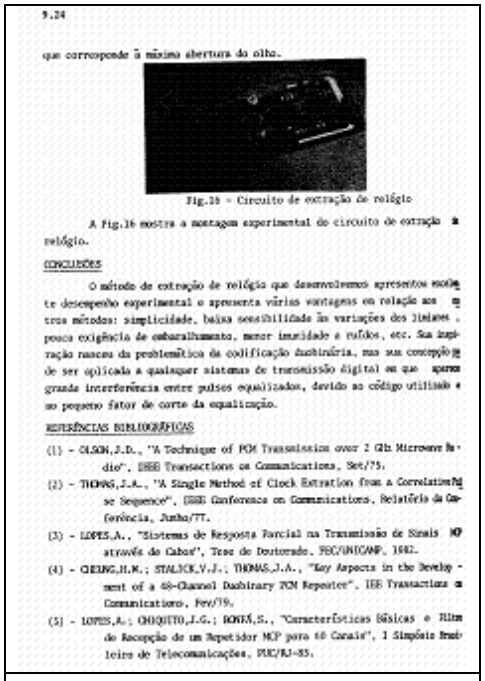


Figure 12. Gray scale version of Figure 10
Size: 3.03 kB – Tiff, 230 kB – pdf

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 [4].

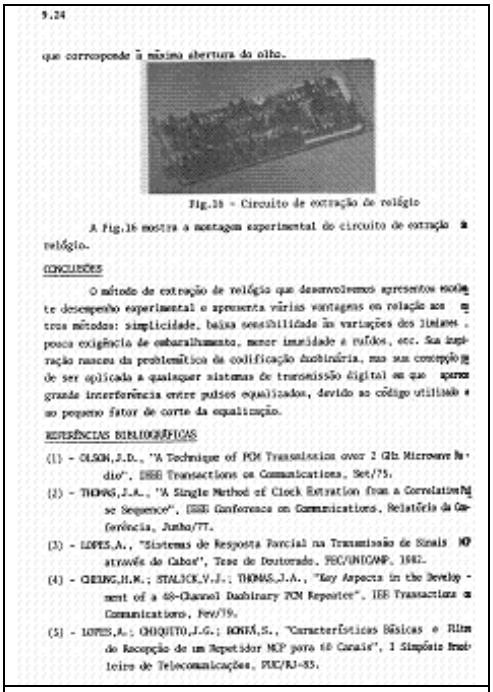


Figure 13. LiveMemory page of Figure 9.
Size: 3.03 kB – Tiff, 343 kB – pdf

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). Identifying all these elements allows a complete navigation of the content of papers. Further details of the LiveMemory Index Generator may be found in reference [9].

VI. CONCLUSIONS

This paper presents LiveMemory a simple and user-friendly software environment to generate digital libraries of proceedings of technical events. At present the top-level interface and image processing routines are implemented. The Index Generator is under development. The integration of LiveMemory with the Terassect [8] OCR Platform was made, allowing semi-automatic indexing.

The authors plan to improve the performance of the OCR by the inclusion of dictionaries specific to the library of authors, keywords, etc. Each word corrected by the user would be inserted in the dictionary, which is then used for the recognition of the subsequent papers.

Another important feature of LiveMemory not addressed in this paper is the capacity to extract information (title, page number, author names, keywords, etc) from pdf files. Such module is still under development but was already used in the generation of the version 2010 of the SBrT Digital Library distributed together with the proceedings of ITS 2010.

LiveMemory executable code will be made freely available under request.

REFERENCES

[1] B.T.Ávila and R.D.Lins, "A New Alg. for Removing Noisy Borders from Monochromatic Documents", ACM-SAC 2004, pp 1219-1225, ACM Press, 2004.

[2] B.T.Ávila and R.D.Lins, "A New and Fast Orientation and Skew Detection Alg. for Monochromatic Document Images", ACM DocEng 2005, ACM Press, 2005.

[3] R.C.Gonzalez and R.E.Woods. Digital Image Processing. Prentice-Hall, 3rd ed., 2007.

[4] R.D.Lins et al, "BigBatch: An Environment for Processing Monochromatic Documents", ICIAR 2006, LNCS 4142, pp. 886-896. Springer Verlag.

[5] J. M.da Silva et AL, "Binarizing and Filtering Historical Documents with Back-to-Front Interference", ACM-SAC 2006, Nancy, April 2006.

[6] J.van Beusekom et al, "Example-Based Logical Labelling of Document Title Page Images", ICDAR 2007, pp. 919-924, IEEE Press, 2007.

[7] ImageJ http://rsb.info.nih.gov/ij/

[8] Tesseract http://code.google.com/p/tesseract-ocr/

[9] R.D.Lins.; G.F.P.Silva; G.Torreao . Content Recognition and Indexing in the Livememory Platform GREC 2009. p. 224-230.