Automatic Selection of Visually Attractive Pages for Thumbnail Display in Document List View

Fabrice Matulic
ETH Zurich, Dept. of Computer Science
8092 Zurich, Switzerland
[First Name].[Last Name]@inf.ethz.ch

Abstract

Document summarization is a task which is difficult to perform automatically, especially if the document is only available as raw pixel data. This paper presents a technique to represent a document as a selection of its most eye-catching pages. The algorithm looks for salient features such as illustrations, diagrams, large titles, headings etc. that cause a page to stand out and ranks its conspicuousness according to the colour, size and number of such elements. A filter function can also be applied to introduce some spread in the selection process, if desired, in order to avoid cases where the extracted pages are too close to each other.

The algorithm is intended as part of a document catalogue system and user interface, in which multiple page thumbnails are shown for each document. The aim is to broaden and enrich a document's visual profile beyond the traditional front cover icon and generally to increase its appeal to potential readers during their browsing experience.

1. Introduction

A crucial aspect when marketing a book that authors and publishers pay considerable attention to is the book's appearance and presentation. A catchy title, an attractive front cover and a gripping blurb greatly contribute to arousing the interest of potential readers. Equally important is how the book is displayed in the bookstore, whether it is prominently placed face out on a shelf near the entrance or whether it is hidden behind a row of other books on the shelf of a remote back room. In the digital world of the Long Tail [1], publications benefit from a wider exposure, as a greater number and variety of items can be accessed in a couple of mouse clicks and keystrokes. But while online bookstores and library systems have virtually unlimited presentation space, the issue of how individual documents should be displayed and what information should be shown at a given time on a user's screen remains. Sites like Amazon and Google Book Search have acknowledged the importance of graphical previews for documents and so commonly include thumbnails of front covers and even inside pages in their summary views. To further increase a book's visibility, Amazon provides a "Look Inside" (now "Search Inside") feature, which allows customers to peek inside a book before purchasing it. To be fair to publishers and authors, only a portion of the book is available for preview, typically the front and back covers, the index, the table of contents and excerpts, for example the first chapter. Looking inside, however, assumes customers' interest has already been piqued as they have already singled out a book to take a closer look at. But in many cases the deciding factor is the first glance, when the book is first casually encountered at the bookstore (brick-and-mortar or virtual). As studies show, the cover plays an important role in getting the reader's attention [2]. Even when leafing through the pages of a book, the features that catch the eye are text written in large font (such as titles), figures and images. It seems therefore important to consider those visual stimuli along with the information conveyed by text when advertising a book product. A natural way to do so beyond displaying its front cover is to also show a few of its most appealing pages. But first, one needs to translate "appealing" in computational terms, if one is to automate the selection process. In this paper, a simple and fast technique to achieve this for the purpose of thumbnail display is presented.

The article is structured as follows: the next section explains the context of the present work. Section 3 details the proposed algorithm. Section 4 presents some early experimental results. Section 5 provides ideas for possible improvements that came up during evaluation. Finally, section 6 concludes the article and lays out the directions for future work.

2. Background

The graphical front-end where the sample pages of a book appear plays an important role in drawing potential readers' attention to a particular document.

It is the virtual shelf, whose layout and presentation will entice customers to pick up (click on) a book. An example of how such an interface might look like is shown in Figure 1.

Thumbnails of selected pages from each document are displayed along with the front cover (represented with a larger icon in the example) and relevant text information about the document. Such an interface can be used for any kind of application that requires user interaction with a document database or repository, may it be for an online bookstore, a digital library or simply local document access. Depending on the desired level of sophistication, the thumbnails can be made to react to user input, such as expanding and contracting on mouse-overs using the popular fish-eye effect. As an added benefit, the interface could also enable users to open documents directly to one of the listed pages by clicking on its associated icon.

Figure 1. User interface with animated fish-eye view where the thumbnails of the selected pages are shown.

Other paradigms where displaying a portion of a document makes sense are of course possible, for instance an enhanced icon view of folder contents for operating systems or a condensed virtual album containing selected pages through which the user can "flip" using a pointing device. Currently, a few mock-ups of possible UIs have been designed as proof of concept. Implementation, integration and evaluation of the interfaces are in progress.

Along with the work done on the presentation layer, effort has been expended on the automated process responsible for choosing the pages of the documents to display as thumbnails, which is the focus of this paper. The task can be seen as a special case of document summarization where the granularity is set at the page level and where attractiveness weighs more than representativeness.

Automatic document summarization has received considerable attention in the past years. The vast majority of the work focuses on text summarization (see [12] and [13] for details about the techniques involved), but some authors such as Bloomberg et al. have considered the case of document images and proposed a method to select excerpts directly from imaged text without performing OCR [3]. Beyond text, Futrelle, recognizing the importance of figures in documents, has attempted to tackle the difficult problem of summarizing diagrams [4]. Berkner et al. developed a technique to automatically generate thumbnails of document images where a so-called "SmartNail" is constructed from cropped and scaled components of the original page image and extracted text elements are remodelled to be readable in the icon [5].

More closely related to the problem addressed in this paper, McCall et al define an interestingness factor computed for each page of a document to determine how distinctive the page is in the whole document and compared to its neighbours [6]. The calculated interestingness score is used to decide which pages should be displayed when scrolling rapidly through a document and where to stop if a user wishes to jump to a particular page. The score depends on several geometrical features such as number of columns, presence of pictures, headings, tables etc. which are obtained by segmentation. The work, however, does not reveal which segmentation technique is used to obtain the components, nor does it mention how the logical labelling is performed. Furthermore, colour or luminance is not considered in the approach so that document elements of the same type but with different colour patterns contribute equally to the page score. Finally, Google Book Search uses undisclosed automated methods to analyse a book and extract information for its "About this book" page[1], including in some cases a selection of pages to be displayed as icons. Depending on the type of document and the density of text vs. images, the displayed thumbnails are limited to the table of contents and a portion of the index or they include several pages with images if the book is a photo album or a heavily illustrated report. The number of chosen pages is not fixed, as the interface does not limit the space for the page thumbnails. What is more, the fact that sometimes fairly nondescript pages appear in the results shows that attractiveness does not seem to be the main criterion for selection.

3. The Algorithm

The proposed algorithm detailed hereafter should be understood as a complement to the text summarization techniques and a faster, simpler alternative to the ones based on complex layout

[1] http://books.google.com/support/bin/answer.py?answer=53549

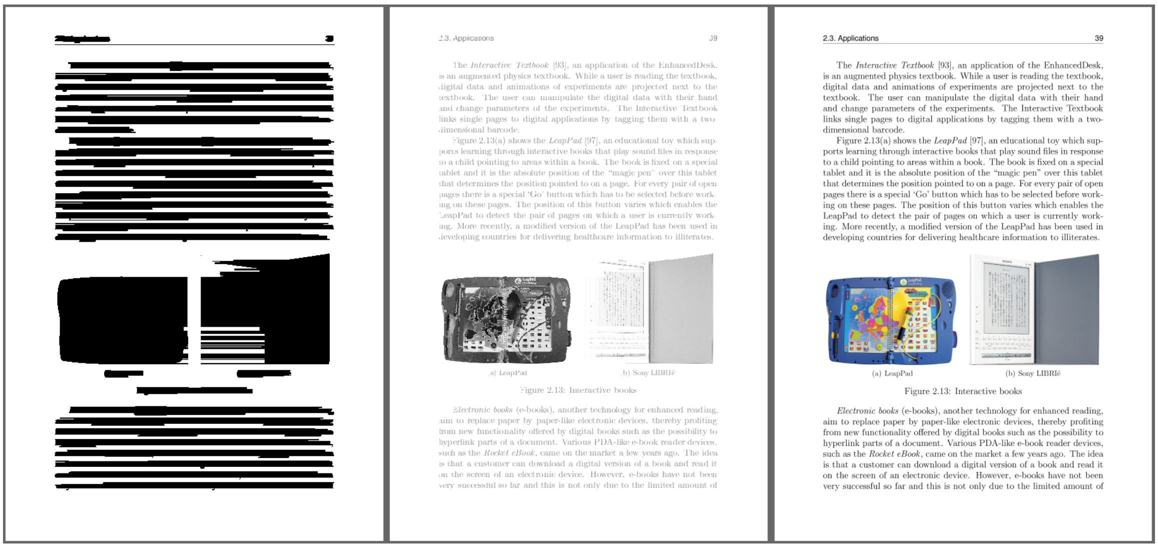

Figure 2. From left to right: a) RLSA bitmap; b) representation of tone saliency; c) original document.

analysis. The approach is tailored to the specific requirements of the intended application, i.e. a document list UI where a selection of interesting pages are shown to the user. Since the space allotted to each document preview or summary is usually constrained by the UI manager, it makes sense to fix the number of pages to be selected. The task is therefore to output the numbers or indices of the p most visually salient pages from a given n-page document (assuming, of course, p $< n$).

3.1. Unweighted Scoring

The algorithm takes a set of page images belonging to a document as input, for example as acquired by a scanner or simply from a rasterized digital document. The first step is to detect large elements in the page that are likely to be noticed by the eye after creating the thumbnail. For that, the Run-Length Smoothing Algorithm [7] with Shih et al.'s optimizations [8] is applied on the page images after they have been binarized and deskewed (if necessary). Although old and surpassed by more modern layout analysis techniques, RLSA has the advantage of being extremely fast and it is suitable enough for our purpose.

RLSA runs through the binarized page horizontally and vertically and connects or "smears" black pixels if they are below pre-defined thresholds. After performing a logical AND on the results of the two smoothing operations, blocks of the constituent elements of the page can be roughly identified. When RLSA is used for segmentation, further analysis is needed to classify the blocks into text, image or table etc. but since we are only interested in the number and size of the elements, this is not required here. Accuracy is also not as critical an issue as with layout analysis, since the blocks are not intended to be subsequently passed on to other processes, which rely on their precise extraction (e.g. OCR). A rough estimate of the size and pixel density of the largest elements is sufficient. Figure 2a illustrates the output of RLSA with high threshold values. Characters and words are merged into lines and the various parts of image elements are merged to form compact blocks.

To mark out the individual blocks obtained in the previous step, a connected component analysis is performed on the RLSA bitmaps. This is also a fast computation, as latest algorithms can determine connected components in linear time [11]. The visual significance or "appeal" score of each component is then determined by calculating a value representative of its area and the vividness or intensity of its colours. The latter attribute is specified based on the consideration that stronger, saturated hues are more visible than lighter, muted ones. This is captured by defining a "tone saliency" value for each pixel in the original page image as follows:

$$T = \begin{cases} (a-b)v + b & \text{if } v + s < 1 \\ (m-b)v + (m-a)s + a + b - m & \text{else} \end{cases}$$

where s and v are respectively saturation and value in the normalized HSV model (i.e. $s, v \in [0,1]$) and a, b and m constants such as $a < b < m$ defining the tone saliency values to assign to 100% white, black and coloured pixels at maximum saturation respectively. T is highest when v and s equal 1, that is, when the pixel colour is fully saturated and lowest when the pixel is completely white. In this simple model, therefore, brightly coloured components are deemed more "attractive" than black elements. An illustration of this concept is given in Figure 2b, where darkness of the shades reflects the tone saliency of the corresponding page (Figure 2c).

The contribution of each connected component C is determined by summing its pixels' T_i values. The squares of those results are in turn summed up over all components in the page to calculate the final score S:

$$S = \sum_{C \in P} \left(\sum_{i \in C} T_i \right)^2$$

When the scores of all pages have been computed, an array containing the sorted indices of the p pages with the highest values is returned. The indices are then used by the thumbnail generator to create the icons from the corresponding page images.

3.2. Introducing Weights

A side effect of calculating global scores is that a number of chosen pages may turn out to be close to each other in the original document. For example, if a document contains a section in which large diagrams that are similar or related to each other appear successively, chances are most or perhaps even all of them will end up in the final selection, to the detriment of other interesting, more different pages in the document. While the goal of the algorithm is not to extract the most representative pages of the document (which is why it differs from traditional summarization), a more distributed selection might be desirable to avoid the aforementioned problem.

To inject some level of spread between two consecutive chosen pages, a filter function can be introduced that artificially modulates the computed page-appeal scores depending on the index with respect to the previously selected page. The idea is to add some amount of periodicity in the selection process, by decreasing the score of pages following one that has just been selected and increasing the value of pages around the index at the pseudo-period. The simple Gaussian function below fulfils that purpose:

$$f(x) = w_\tau e^{-A(x-\tau)^2}$$

where τ is the pseudo-period (typically the total number of pages in the document divided by the number of pages to be selected), w_τ the weight to be applied at τ and A, a constant which depends on τ, w_τ and w_0, the weight at the index of the currently selected page+1.

The function is applied iteratively and shifted across the page indices every time a page is chosen. Assuming i is the index of the last selected page, say the q^{th} of the p pages to be extracted, the coordinate system is shifted so that $x = 0$ at i and values of f are computed for $x = 0..(n-i)$. The obtained values are then multiplied with the scores calculated in the first step to determine a new array of modulated scores. The indices of the pages with the q highest scores are sorted and the page with the smallest index, that is, the one closest to the previously selected page, is chosen for the next iteration. The algorithm starts with page 1, which is always selected by default, and iterates through the array until the required number of pages have been selected.

The influence of the filter function on the initial page scores can be increased or decreased by modifying its parameters. For example, the periodicity will be more strongly enforced if $w_\tau \gg w_0$.

4. Results

The page selection algorithm was tested on a number of documents containing mixed text and illustrations. Those included reports, presentations and PhD theses. A first series of experiments were carried out and the output used to adjust the parameters of the program until satisfactory results were obtained. For instance, with the 9/11 Commission Report[2] as input, the following 10 pages were selected: [1, 66, 81, 165, 255, 296, 301, 305, 329, 330] and with the filter function applied: [1, 32, 66, 81, 148, 165, 255, 305, 329, 430]. The pages in the last array are more dispersed thanks to the weights applied to the scores. But this came at the price of choosing seemingly less "interesting" pages. As shown in Figure 3, a page of pure text and a sparse diagram have taken the place of pages with denser illustrations and photos in the filtered set. This is due to the relatively low number of images and their uneven distribution in the original document.

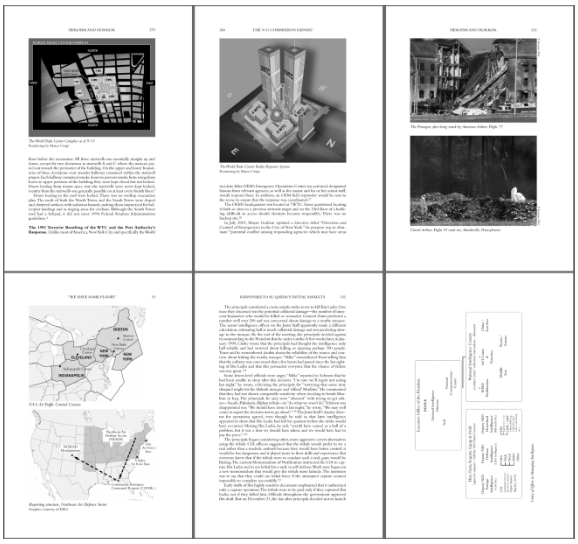

Figure 3. The 3 pages above from the original unfiltered selection were replaced by the 3 below after applying the weights

[2] http://govinfo.library.unt.edu/911/report/911Report.pdf

To assess the relevance of the results, a short subjective evaluation was conducted in which the authors of some of the documents (in the present case PhD theses) used for the tests were asked to comment on the selections made by the algorithm. Specifically, they were given thumbnail renderings of the pages of their documents and asked to choose 9 (other than the first page which is always included) that they feel would best serve to advertise their work when displayed in a catalogue UI such as the one shown in Figure 2. The authors were also asked to explain their reasons for choosing those particular pages. In a second step, they were given the output of the algorithm including unweighted and weighted results and for each page asked whether they agreed with its inclusion in the selection or not. Overall satisfaction and comments on the programme's choices were also collected.

Generally, it turned out that authors used very different criteria to make their selections. Some preferred pages with coloured images and backgrounds, whereas others wanted their choice to be more representative and hence included some text pages. Some did not consider page spread important, some did. But despite the differences in the selection approaches, there was always some overlap between the authors' first selections and the output of the program. More importantly, almost all participants deemed the choices made by the algorithm acceptable. On average, they approved 82% of the unweighted selection. The filter function was shown to have fulfilled its role as distributing and diversifying agent, albeit sometimes at the cost of selecting less appealing pages. The fact that satisfaction was slightly higher (84%) is explained by some authors who wanted more variety in the final selection and hence were happier to have one or more thumbnails of less "interesting" pages included among the ones with more pronounced features.

The dimensions of the target thumbnails played an important role in some cases. The longer dimension of the icons was first set at 120 pixels, which is slightly larger than that of the thumbnails of Amazon's book covers. Depending on the size, important chapters and section titles may or may not be readable and thus influence the participants' decisions. This was confirmed after asking the volunteers to pick 9 pages from the same documents, but this time using 200x200 icons (i.e. slightly larger than the size used in Google Book Search). At this size, chapter titles became readable and consequently some authors tended to include them in their selections. As a result, the average page-by-page satisfaction percentage fell below 70%. The algorithm, of course, does not perform any legibility analysis and so outputs the same results regardless of the resolution of the target thumbnail.

5. Possible Improvements

Through testing and discussions with the participants in the preliminary evaluation experiments, many ideas to improve the selection process (but at the cost of increased complexity) emerged. One of them, already mentioned above, is to consider text readability as a possible factor that should influence the score. If the table of contents or a list of the main chapters is not included in the document summary view, a detected chapter title or heading in the document image might be of value for the user and so should perhaps be made to influence the page score.

Another aspect which increases a page's visibility and is therefore probably worth consideration is contrast and colour distribution. The simple tone saliency feature defined above works well to cull dense, coloured elements over lighter ones but it does not make any distinction between different hues (saturated yellow for example is less visible on a white background than saturated red or blue), neither does it really reflect the attractiveness of an element as a whole. More generally, one can expect that identifying and factoring in low-level image features having an influence on humans' visual attention will bring more flexibility and make the system more robust to a wider variety of documents. Attention-driven approaches have already been successfully used for generic image retrieval (e.g. [9] and [10]) and although the techniques might not all be directly applicable to document images, it is likely that they can contribute at least in part.

As for the filter function, the weighted scores ensure there is a certain amount of distribution in the choice of pages, but they do not guarantee that the selected pages are all different, since it is possible for pages with similar visual patterns to be located at different places in the document. To eliminate or alleviate this problem, another filter function could be added that would compare the visual structure of pages to that of those which have already been selected. A page's score would be decreased accordingly if it is determined to be too similar to a previously chosen one, thereby guaranteeing more diversity without compromising interestingness.

6. Conclusion

The page-selection algorithm was designed as a quick and simple way to output a specified number of suitable candidates for page thumbnails that could be used in a document list user interface. The initial hypothesis was that this could be achieved without resorting to comprehensive layout analysis and elaborate user-attention models, which were seen as too complex and unnecessarily costly for the intended purpose. To a great extent this assumption has been substantiated, as documents with a combination of text and images yielded good results at relatively low computational costs. Besides, the penalty for selecting less outstanding pages than what authors might have chosen is minor, since the impact of showing one page icon over another is not all that consequential. Focusing on adding and refining the detection of features that affect a page's visual appeal such as visible titles and salient objects seems to offer a good compromise between increased robustness and complexity of the image processing techniques involved.

Directions for future work are therefore twofold: improvement of the algorithm by adding more features influencing the page scoring and implementation of the user interface(s). Only through the latter will it be possible to determine how far the page icons do indeed contribute to increasing a document's popularity.

7. References

[1] C. Anderson, *The Long Tail: Why the Future of Business is Selling Less of More*, Hyperion, 2006.

[2] A. D'Astous, F. Colbert and I.Mbarek, "Factors influencing readers' interest in new book releases: An experimental study", *Poetics* 34(2), 134–47, 2006.

[3] F.R. Chen and D.S. Bloomberg, "Document image summarization without OCR", *Proceedings of ICIP'96*, Lausanne, Switzerland.

[4] R. P. Futrelle, "Summarization of Diagrams in Documents", *Advances in Automated Text Summarization*, I. Mani and M. Maybury. Cambridge, MA, MIT Press., 1999

[5] K. Berkner, E. L. Schwartz and C. Marle, "SmartNails - Image and Display Dependent Thumbnails". *Proceedings of SPIE*, San Jose, 2004.

[6] K. McCall and K. Piersol, "Enhancing accuracy of jumping by incorporating interestingness estimates", US Patent 20070180355

[7] K.Y. Wong, R.G. Casey, and F.M. Wahl, "Document analysis system", *IBM Journal of Research and Development*, 26(6):647–656, November 1982.

[8] F.Y. Shih and S.S. Chen, "Adaptive document block segmentation and classification", *IEEE Trans. Syst. Man Cybernetics* Part B 26, 1996.

[9] A. Bamidele, F. W M Stentiford, F. W. M. Stentiford and J. Morphett, "An Attention-Based Approach to Content-Based Image Retrieval", *BT Technology Journal* v.22 n.3, July 2004.

[10]H. Fu, Z. Chi, D. and Feng, "Attention-driven image interpretation with application to image retrieval", *Pattern Recognition* v39 i9, 2006.

[11] F. Chang, C.-J. Chen, and C.-J. Lu. "A linear-time component-labeling algorithm using contour tracing technique", *Computer Vision Image Understanding* vol 93, 2004.

[12] I. Mani and M.T. Maybury (Eds.), *Advances in Automatic Text Summarization*, MIT Press, 1999.

[13] D.R. Radev, E.H. Hovy, and K. McKeown, "*Introduction to the Special Issue on Summarization*", Computational Linguistics, 2002.