Proceedings of SPIE, Internet Imaging II, San Jose, USA, January 2001, Vol. 4311, pp. 198–205.

Accessing Textual Information Embedded in Internet Images

A. Antonacopoulos[a], D. Karatzas[a] and J. Ortiz Lopez[b]

[a]PRImA Group, Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom
http://www.csc.liv.ac.uk/~prima

[b]Telecommunications College, Universitat Politecnica de Catalunya, Barcelona, Spain

Keywords: Text extraction, image analysis, character recognition, WWW, Internet, indexing, searching.

ABSTRACT

Indexing and searching for WWW pages is relying on analysing text. Current technology cannot process the text embedded in images on WWW pages. This paper argues that this is a significant problem as text in image form is usually semantically important (e.g. headers, titles). The results of a recent study are presented to show that the majority (76%) of words embedded in images do not appear elsewhere in the main text and that the majority (56%) of ALT tag descriptions of images are incorrect or do not exist at all. Research under way to devise tools to extract text from images based on the way humans perceive colour differences is outlined and results are presented.

1. INTRODUCTION

The message that the author of a WWW page wishes to convey is expressed not only by the textual content but also by the audiovisual setting (overall design and individual multimedia components). The use of images is by far the most common way to add visual content to an otherwise plain text document. Images are used for illustrations as well as for adding impact to a textual message. The latter is the subject of this paper. A study of the use of images and in particular of those that contain text is presented. The results demonstrate a significant need for developing new methods and tools to extract and recognise text in images.

Text remains the primary (the only one, in most cases) medium for indexing and searching for WWW pages. Search engines use the text in various circumstances on the page (e.g. title, ALT-tags, meta-tags and body text) to index, and sometimes rank, WWW pages. In the case of body-text, indexing is performed using standard information retrieval techniques. The frequency with which a term appears and its location on the page will usually rank the page higher when this term is a keyword in the search. In terms of location, the higher on the page a term is and/or it appears in a heading the higher the ranking will be.

Page titles and headers are semantically important entities; that is precisely why the text they consist of is considered more relevant for indexing a page. However, because titles and headers are very important, WWW page designers wish to add visual impact to them by creating them in image form (rather than encoding them as text).

Text in image form (or text embedded in images) is inaccessible to any automated way of indexing (e.g., search engine crawlers). In essence, this text is an important part of the document that can neither be analysed with current technology nor indeed represented in a purely textual context (e.g., browsers with images switched off or reading systems for visually impaired users).

Correspondence: A. Antonacopoulos (e-mail: aa@csc.liv.ac.uk). Work supported by equipment grant from Hewlett-Packard Co. under the Pan-European Internet Initiative. All figures were originally in colour.

Figure 1. The ALT tag text for this main page header is "Mooooo!"

HTML provides for an alternative textual description using ALT tags. However, ALT tags are not mandatory and as the study below indicates, WWW page designers do not always follow good practice in using them. In fact, a significant proportion (56%) of the ALT tag text is either inaccurate or the tag does not even exist. The problem is compounded by the dynamic nature of the WWW where pages are frequently updated and the original ALT tag text can quickly become irrelevant (it is up to the designers to verify these descriptions after changes have taken place).

Search engines (e.g. excite) do not always index ALT tags and others do not place any special relevance value on them (where they should if the tag refers to a title/header) [1]. An example of a WWW page header in image form (from the TUCOWS software site) can be seen in Figure 1. Arguably, its amusing ALT tag text "Mooooo!" is not an accurate representation of the main title of the page. In many similar situations, albeit not as entertaining, it is evident that a significant part of the message is lost without the images and appropriate alternative text.

The problem of not being able to process text in image form extends beyond text portions embedded as images on the page. Scanned documents in general (possibly as Acrobat (PDF) files) are also important examples of information ignored by search engines and invisible to text/voice browsers.

There is clearly a significant need to develop methods for extracting and recognising of text in images. The facts supporting this argument are presented in the next section. Section 3 briefly examines the characteristics and problems posed by the images in question. An overview of the research carried out by the authors for the development of a new text extraction approach and the issues involved is given in Section 4. Finally, Section 5 discusses the conclusions of the paper.

2. TEXT IN IMAGES SURVEY

To discover the extent of the presence of text in image form on WWW pages and assess the significance of the issues arising the authors carried out a survey of WWW sites over six weeks during July and August 1999. The survey shows a relatively large part of the message of a WWW page is in image form (GIF or JPEG) and that the alternative description for text in images is not reliable.

2.1. Sample and measurements

The rationale for selecting the sample was to identify types of WWW pages that an average user would be interested in browsing. The pages came from sites that one would go to purchase items (e.g., books, CDs), to read news (e.g., TV stations, newspapers), to find information (e.g., on products, services) and, in general sites, for which it is important to be found by a user (commercial, academic and other organisations). Sites dealing with leisure activities as well as work-related and other routine activities were included. The language of the pages was English, and therefore, the majority of the sites were from the United States and the United Kingdom. For the purpose of this survey, there should not be loss of generality by this choice.

In total, 200 WWW pages were manually processed. In terms of the presence of text in images the following were measured on each page:

- total number of words visible on page
- number of words in image form

- number of words in image form that do not appear elsewhere on the page

In terms of the alternative text present as ALT tags associated with the text in images, the following were measured on each page:

- number of correct descriptions (ALT tag text contains all text in image)
- number of false descriptions (ALT tag text disagrees with text in image)
- number of incomplete descriptions (ALT tag text does not contain all text in image)
- number of non-existent descriptions (there is no ALT tag text for an image containing text)

Using the above measurements, overall results as well as individual ones (for different categories of WWW sites) are obtained.

2.2. Survey results

Overall, 17% of the words visible on the WWW pages is in image form (see Figure 2). This is a significant number considering that important text is inaccessible using current technology.

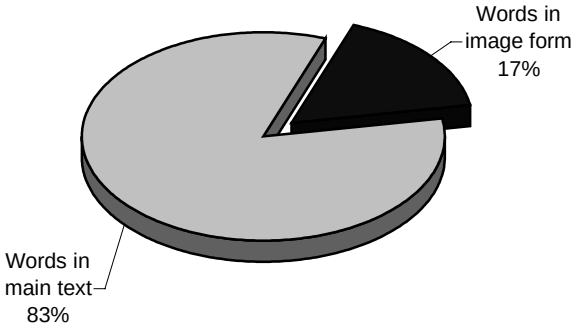

Figure 2. Percentage of words in WWW pages in image form and in the main text.

The facts on the ALT tag text description of the text in image form are not encouraging either: only 44% of the ALT text is correct (see Figure 3). The remaining 56% is either false (3%), incomplete (8%) or non-existent (45%). This means that more than half of the text in image form is totally inaccessible.

Of the total number of words in image form, 76% do not appear elsewhere in the main (visible) text (see Figure 4). This corresponds to 13% of the total visible words (and most probably important) on a page lost for indexing purposes. These figures are in broad agreement with an earlier survey of 100 WWW pages by Lopresti and Zhou [2]. This fact demonstrates that the situation is not improving. Moreover, as the volume of the information on the WWW grows the total number of unreadable material increases considerably.

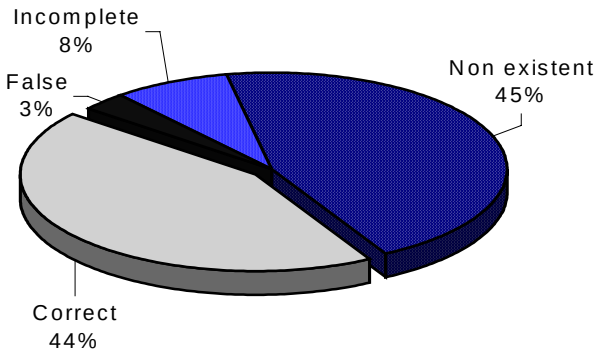

Figure 3. Percentage of correct and incorrect ALT tag descriptions.

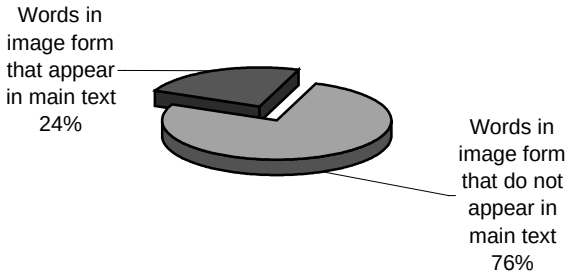

Figure 4. Percentage of words in image form not appearing in the main text.

3. TOWARD TEXT EXTRACTION AND RECOGNITION

From the above, the need is evident for the recognition of text in image form. Within the PRImA group at the University of Liverpool research is being carried out to develop methods to extract text from images and recognise it. An initial account can be found in [3] while a description of the current approach is given in [4]. Such technology can be incorporated in crawlers (for better indexing) and browsers (for more accurate alternative description). It can also function as a validating tool for ensuring the correctness of ALT tag text.

The recognition of text in images on WWW pages is by no means straightforward. In fact, it is quite more complex than usual OCR on scanned (bilevel) documents. Although there are no artefacts in WWW images due to scanning, there are a number of obstacles to overcome. The main characteristics of the text extraction and recognition problem for WWW page images are:

- the image resolution is about 75 dpi which is of very low quality compared to the usual minimum of 300 dpi required by OCR methods,
- text can be present in different colours/textures and placed on complex backgrounds (in contrast, in the vast majority of cases, OCR works on a bilevel image),
- most characters are of very small size (5 – 7 pt) compared with the characters present in traditional documents (usually at least 9 – 10 pt),

- there could be artefacts resulting from the colour quantization and antialiasing processes used in the authoring software, and
- there are serious artefacts resulting from the lossy compression (e.g., JPEG) of the images.

Considering the above idiosyncrasies of the text in image form and the fact that the WWW page designers' goal will always be to add visual impact to the text, one can appreciate the great degree of difficulty of the problem. Previous methods [2, 5] concentrate on the identification of single-coloured text, ignore very small text, and perform a global analysis of the colour information. Furthermore, the typical methods for the analysis of texture [6] can be computationally too expensive for practical application.

The new method (developed in the PRImA group) aims to improve on past approaches by avoiding the use of traditional texture analysis methods and by extending the capability to handle more difficult cases, such as gradient background and non-uniform text colour. An overview of the rationale, a description of the method and a presentation of experimental results are given in the next section.

4. THE ANTHROPOCENTRIC APPROACH

The research carried out in the PRImA group for text extraction and recognition is centred on observations of the way humans perceive colour differences. This idea stems from the fact that graphics on the WWW (including text in image form) are designed to be viewed on monitors (see Figure 5). This is an important consideration as it impacts on the design of new methods in two ways.

First, the text in image form is created in a way so that it 'stands out' from the plethora of competing visual information on a WWW page. In contrast to traditional documents, greater emphasis is placed (consciously or subconsciously) by the designer of WWW pages on the perceived background-foreground differences. Therefore, different colour similarity/separation criteria must be established for these images.

Secondly, the colour space in which the methods operate should be one that corresponds to human vision as closely as possible. The new method, in contrast to previous approaches, abandons the RGB space in favour of a colour space that distinguishes colours in terms of chromaticity and luminance. This approximates human colour perception where colours are perceived based on chromaticity (two axes: green–red, yellow–blue) and luminance (different from the commonly used (R+G+B)/3) [7]. In reality, colours that have similar distances in RGB may be perceived by humans as having greater or less difference. By working on an appropriate chromaticity-luminance colour space and by using a distance measure derived from observations of human colour perception we can identify these more subtle differences in colour.

The method consists of two main steps. The first is colour space analysis, where regions of different colours are identified in terms of hue and luminance. The second is segmentation, which results in larger (character-like) aggregate components based on spatial distance and similarity criteria. Recognition then takes place (not described here) on these components.

Figure 5. An example of a WWW page header (text in image form).

4.1. Colour space analysis

The first step involves converting the RGB data to HLS as a first reasonable approximation to a human perspective. Then the analysis starts at either the Hue or the Luminance histogram (depending on an initial analysis of colour content of the image). Regions of similar colour are identified from the peaks in the histogram. For these identified regions the other histogram is then calculated and the regions are split with finer similarity criteria. The histogram analysis alternates through Hue and Luminance until a threshold determines that regions cannot be split any further. Eventually, through this cascading analysis of alternating hue and luminance histograms the image is split into a number of layers in a tree structure [4]. Each layer contains regions of distinct (range) hue and/or luminance, with neighbouring (child or parent) layers containing regions that are closest in colour.

4.2. Segmentation

The segmentation step attempts to piece together the regions that have been separated in different layers into larger character-like aggregate regions. Each region on each of the layers is considered along with its immediate neighbours to determine whether they are 'similar enough' to be joined. This visual similarity refers to the way humans perceive colour similarity. To assess this similarity the colour information in the image is translated into wavelength data as the only objective information available on human responses to colour is in terms of wavelengths [8]. As the standard RGB or HLS colour spaces are hardware dependent, the relatively safe assumption is made that the image data (RGB) is the same as in the sRGB space [9]. This assumption allows the conversion of sRGB data to XYZ [10] for which there is a straightforward conversion to wavelengths. Using the wavelength information, a distance of colour similarity is calculated and along with an evaluation of topological and spatial criteria aggregate regions are created. The resulting regions have associated confidence values and based on those, character-like components are identified or marked for further consideration and merging with others. The final regions will be sent to the character recognition module (not described here).

4.3. Text extraction results

The resulting text-like components extracted from the image of Figure 5 can be seen in figures 6–9. In each figure, the foreground shows the result contained in a separate layer (colour). In figures 6–8 a two stage histogram analysis (Hue for the whole image, then Luminance for selected regions) was adequate. To obtain the result in Figure 9, a further consideration of the Hue histogram was necessary for the regions identified after the first two steps.

Figure 6. Result from Hue -> Luminance

Figure 7. Result from Hue -> Luminance

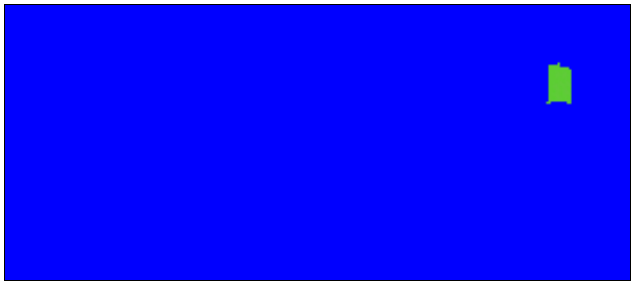

Figure 8. Result from Hue -> Luminance

Figure 9. Result from Hue -> Luminance -> Hue

5. CONCLUSIONS

This paper has presented the facts and surrounding issues involved in accessing textual information contained in images on the Internet. The case in point is the inaccessibility of semantically important information (text in image form) such as WWW page headers, titles and other information with high indexing value. The results of a recent study were reported showing that 76% of the words in image form do not appear in the main text and that 56% of the ALT tag descriptions are incorrect or missing. This fact poses a significant problem for the accuracy of both the indexing of WWW pages and for the ranking of search results. From the above it is established that there is a significant need to devise methods to extract textual information from images and represent it in a suitable encoding that can be processed and analysed by computer.

New methods and tools are being developed within the PRImA research group to extract and recognise the text embedded in images on WWW pages. Such tools can be used by search engine crawlers for improved indexing, by browsers for uniform

presentation in pure text or audio, or by specific tools to validate ALT tag descriptions. A new text extraction approach based on the way humans perceive colour differences was outlined. Preliminary results of the text extraction methods are encouraging and further work is focusing on the extraction and recognition of text in more complex situations.

REFERENCES

1. Search Engine Watch (http://www.searchenginewatch.com)
2. J. Zhou and D. Lopresti, "Extracting Text from WWW Images", Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, August, 1997
3. A. Antonacopoulos and F. Delporte, "Automated Interpretation of Visual Representations: Extracting textual Information from WWW Images", Visual Representations and Interpretations, R. Paton and I. Neilson (eds.), Springer, London, 1999.
4. A. Antonacopoulos and D. Karatzas, "An Anthropocentric Approach to Text Extraction from WWW Images", Proceedings of the 4th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro, Brazil, December 10–13, 2000.
5. M. Kopen, L. Lohmann and B. Nickolay, "An Image Consulting Framework for Document Image Analysis of Internet Graphics", Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, August, 1997.
6. M. Sonka, V. Hlavac, R. Boyle, *Image Processing, Analysis and Machine Vision*, International Thomson Computer Press, 1993.
7. R.W.G. Hunt, *Measuring Colour*, John Wiley & Sons, 1987.
8. R.E.Bedford, G.W.Wyszecki, "Wavelength Discrimination for Point Sources", *Journal Of The Optical Society Of America*, vol. 48, no. 2, February 1958.
9. M. Stokes, M. Anderson, S. Chandrasekar, R. Motta, *A standard default color space for the internet - sRGB*, 1996 (http://www.w3.org/Graphics/Color/sRGB.html)
10. G. Wyszecki and W. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2e, Wiley, New York, 1982.