Proceedings of the 4th IAPR International Workshop on Document Analysis Systems (DAS2000), Rio de Janeiro, Brazil, December 2000, pp. 515–525.

An Anthropocentric Approach to Text Extraction from WWW Images

A. Antonacopoulos and D. Karatzas
Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom

E-mail: aa@csc.liv.ac.uk
Tel: +44-151-794 3695
Fax: +44-151-794 3715
WWW: http://www.csc.liv.ac.uk/~aa

Abstract

There is a significant need to analyse the text in images on WWW pages, both for effective indexing and for presentation by non-visual means (e.g., audio). This paper argues that the extraction of text from such images benefits from an anthropocentric approach in the distinction between colour regions. The novelty of the idea is the use of a human perspective of colour perception in preference to RGB colour space analysis. This enables the extraction of text in complex situations such as in the presence of varying colour and texture (characters and background). More precisely, characters are extracted as distinct regions with separate chromaticity and/or luminance by performing a layer decomposition of the image. The method described here is the first in our systematic approach to approximate the human colour perception characteristics for the identification of character regions. In this instance, the image is decomposed by performing histogram analysis of Hue and Luminance and merging in the HLS colour space.

1 Introduction

Images are an important part of the message of any document. In the case of WWW pages, in particular, images play a crucial role in bringing visual impact to an otherwise plain text medium. WWW page designers regularly create page headers and titles as well as other semantically important textual entities in image form.

Using current technology it is not possible to analyse the text embedded in images on WWW pages. This is a significant problem as explained below.

Search engine crawlers [1] are not able to use the text in images for indexing. More often than not, this text contains key index terms, which in the majority of cases do not appear elsewhere on the page (main text). The authors have conducted a study of 200 WWW sites which were considered to have content that most users would be likely to access (e.g., information, services, commerce etc.) and originated from commercial, academic and other organisations. The results agree with the earlier study by Lopresti and Zhou [2] and show that the situation is considerably serious. Of the total number of words visible on a WWW page, 17% are in image form (most often semantically important text). Of the words in image form, 76% do not appear elsewhere in the encoded text. Furthermore, the ALT tags of the images in question are incomplete, wrong or do not exist in 56% of the cases.

Moreover, this inability to access important key terms hinders the effective ranking of search results [3].

Another important issue is that with the increasing use of multimedia content on the WWW there is no uniform representation of the content page in a form that can be analysed in an automated way. Text remains the only medium that can be readily analysed, converted to voice etc. It can be argued that it is desirable, given the richness of text expressions, to obtain a textual representation (natural language) of the content of WWW pages. This will enable, for instance, listening to information on a WWW page without looking at a monitor (with a significant application potential). The recognition of text in images is a step towards achieving this representation.

Responding to the evident need for text recognition on WWW images a small number of approaches have been proposed. Zhou and Lopresti [4] have proposed methods for both text extraction and recognition. Their method for text extraction is based on clustering in the RGB colour space and then for each cluster assessing connected components whose colour belongs to that cluster. The approach works well with GIF images (only 256 colours) and when characters are of almost uniform (constant) colour. With similar assumptions about the colour of characters, the approach of Antonacopoulos and Delporte [5] uses two alternative clustering approaches in the RGB space but works on (bit-reduced) full-colour images (JPEG) as well as GIFs. Jain and Yu [6] report a method based on decomposing an original image into a number of foreground images and a background one. The original number of colours (8-bit or 24-bit images) is dramatically reduced (to between 4 and 8 distinct colours) by bit dropping and colour quantization in the RGB space. Although this method produces good results for relatively simple images, it shares the same problems with the other methods when more complex images are encountered.

Existing approaches assume a practically constant and uniform colour for text and fail when this is not the case. In practice, there are many situations where gradient or multicolour text is present. Furthermore, the background may also be complex (in terms of colour) so that the assumption that it is the largest area of (almost) uniform colour in the image [6] does not necessarily hold.

This paper proposes a method to extract characters of non-uniform colour and in more complex situations. It argues that the RGB colour space representation is not suited to the extraction of text from WWW images and adopts an approach based on analysing differences in chromaticity and luminance instead. This approach is closer to how humans perceive distinct objects.

In the following section, the characteristics of the images on the WWW are presented. In Section 3, the text extraction method is described, each step in a separate subsection. Finally, results are presented and the main issues are discussed in Section 4.

2 Characteristics of images on WWW pages

Images on the WWW differ in many aspects from real scenes and other document images. Especially the images that this paper is concerned with (banners, headers, illustrations etc.), are optimised for viewing on a monitor screen (see Figure 1). This fact manifests itself in the following two main ways.

First, text in image form is designed to have an impact on the reader. Within the plethora of information on the WWW, this text must compete to attract the attention of the viewer and be quickly noticed. The colour of the text and its visual separation from the background is therefore chosen (consciously or subconsciously) according to how humans perceive it to ‘stand out’.

Humans perceive colour differences with certain bias and distinguish colours based on chromaticity (two axes: green–red, yellow–blue) and luminance (different from the commonly used (R+G+B)/3) [7].

The premise of this paper is that a method for text extraction in these circumstances will benefit from the analysis of colour and luminance differences as humans perceive them and not

necessarily as expressed in the RGB space. In reality, colours that have equal distances in the RGB space are perceived by humans as having unequal differences. Therefore, the difference between colours that were designed to be contrasting will be perceived by humans as greater than that between non-contrasting colours, whereas the pairwise distances in the RGB space may not be as indicative of the colour difference. Similarly, colours belonging to the same object will be designed to be perceived as more similar (irrespective of their RGB distance). We argue that the background-foreground difference is more straightforward to identify in a colour space that takes into account the perceived difference between colours as opposed to identifying it in the RGB space.

Figure 1. A heading in image form.

Secondly, from a technical aspect, WWW images have to obey size constraints as downloading speed directly depends on the data volume. As a result, these images tend to be of low resolution (just good enough for display) and tend to be compressed (JPEG standard). More specifically to the images containing text, the following can be observed: The resolution is usually just 72 dpi, and various artefacts are present due to colour quantization and lossy compression. Moreover, the font-size used for text is very small (about 5pt–7pt) [2]. These conditions clearly pose a challenge to traditional OCR, which works with 300dpi images (mostly bilevel) and character sizes of usually 10pt or larger.

3 Text extraction method

Based on the idea of exploiting the human perception of colour differences to identify and extract characters from their background, the authors have been systematically examining different approaches to differentiate between the background and character regions. The approach presented in this paper is based on the Hue-Luminance-Saturation (HLS) representation of colour as an attempt to analyse the differences in chromaticity and luminance. A second approach, based on a more detailed colour representation according to the tristimulus theory [7][8] is also under development. The work on character segmentation based on the HLS representation is described below. The method works by identifying and analysing regions that differ in chromaticity (hue in this case) and luminance.

3.1 Colour space analysis

The first step performed by the method is a conversion of the RGB data stored in the image file into the HLS representation. The HLS representation is chosen in this case as a good first approximation to the human visual system as it contains representations of both the Hue (wavelength) and the Luminance characteristics of it. Luminance is a joint sensation of rods and cones whereas Hue is a joint sensation of cones of three types. A further advantage is that there is

a 1:1 conversion defined between RGB and HLS, whereas this may not be the case with other colour spaces.

From the HLS data, two histograms are calculated over the whole image: one for Hue and one for Luminance. Next, the analysis starts in one of the two and identifies some regions in the image. The other histogram is then calculated for the identified regions and analysed and so on (see below).

The goal of the analysis process is to identify, from the two types of histograms, areas of similar colour in the image. The order in which histograms are considered depends on the colour content of the image and is determined by an initial analysis of the peaks in both histograms. The rule is that the histogram with the most separated peaks is analysed first. The exception is the case of single-hue images (where there is a distinct peak in the Hue histogram) in which case the analysis takes place only in the Luminance histogram. In practice, more often than not, the Hue histogram is analysed first. The Hue histogram for the image of Figure 1 can be seen in Figure 2.

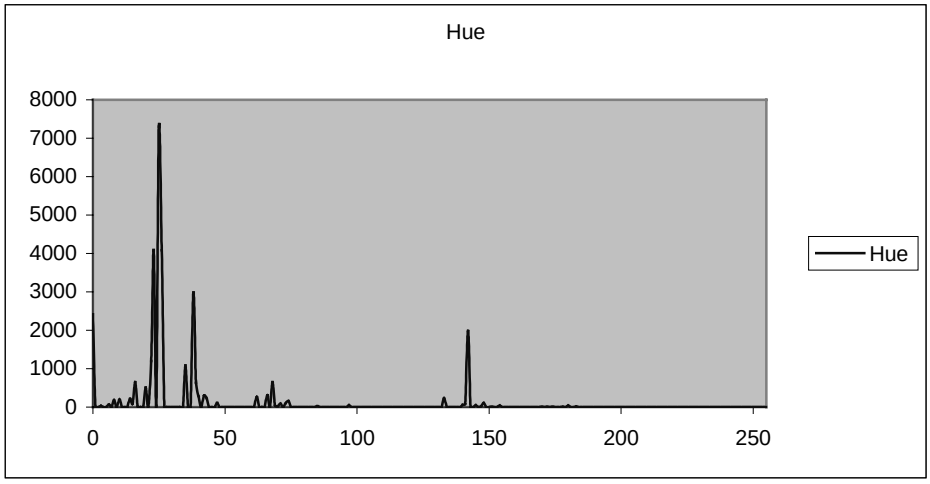

Figure 2. The Hue histogram of the image in Figure 1.

In the construction of the Hue histogram, pixels with luminance close to the ends of the range (very near white or very near black) are excluded as they do not contain useful chromaticity information (humans cannot differentiate colours in those cases).

It should also be noted that, due to the nature of hue representation, a distinction must be made between pixels with hue 0 (zero). A pixel may have hue 0 and be either some shade of grey (in which case its saturation S value will be zero) or some shade of red (saturation is non-zero). The pixels corresponding to grey values are not considered for the Hue histogram (but are of course for the Luminance histogram). These grey-valued pixels are extracted and placed on a separate layer (see below).

Next, in the analysis of the Hue histogram, peaks corresponding to regions of similar chromaticity are identified. A peak in the histogram is considered to correspond to a significant distinct region in the image if it satisfies certain conditions. Features that can be used to express attributes of peaks are:

- The area under the peak (sum of pixel counts) should be large enough to contain text. A fixed threshold is used as minimum. At the same time, the area under a peak should not be approaching very close to the total area of the image.
- A symmetrical peak shows good separation from others. For instance, a peak that rises high and descends a little will not be well separated from the next one. One way of expressing symmetry is as the ratio of the height of the left side to the height of the right side of the

peak. The ratio of the width of the left to the width of the right component can also be used, or both ratios can be combined.

- The width/height ratio indicates the broadness of range of colour within a peak. A narrow and high peak indicates that colours within it are well separated from the rest. On the other hand, a wide and relatively low peak indicates the presence of gradient colour.

Currently, the area under the peak and the height ratio are used to identify peaks. These gives satisfactory experimental results but can be improved upon by the use of additional features as above.

For each identified peak in the histogram the regions in the image corresponding to the colours contained in the peak are identified and separated as a distinct layer. The image is therefore decomposed into a number of layers. Each layer may be further decomposed by calculating and examining the Luminance histogram for the regions in the given layer and identifying peaks in that. The selection of peaks in the Luminance histogram is very similar to the identification of peaks in the Hue histogram but instead of colour separation the goal is to analyse lightness differences. The process continues alternating between the two histogram types, sub-dividing layers until peaks in a histogram cannot be further separated (according to the rules above).

The result is a number of layers in a tree structure, each layer containing regions of distinct (range) hue and/or luminance, with neighbouring (child or parent) layers containing regions that are closest in colour.

3.2 Segmentation

On each layer, each identified region is considered and its neighbourhood in the original image is analysed. This analysis concentrates on determining which pixels could be annexed to the area based on visual similarity. The criterion is that a human should not be able to perceive significant difference between the core region and the annexed pixels. For layers resulting from Hue the analysis refers to wavelength discrimination while for layers resulting from Luminance it refers to luminance discrimination.

The determined surrounding pixels of each core region are noted as a potential extension of that region. These potential extensions are subsequently analysed to determine overlaps in order to form larger connected components. This is particularly important when the layer analysis results in oversplitted regions. The connected component analysis takes into account the size, aspect ratio and density of individual components, in addition to overlaps of potential extensions of components as described above. Character-like connected components are then analysed in the context of other components in close proximity. Similarity and distance measures are computed, however not collinearity as characters are not assumed to be placed on straight lines.

A confidence value (calculated from the above features) is assigned to each individual connected component. Components whose confidence value falls within a middle range (regions with low values are rejected and regions with very high values accepted outright) are further analysed. This analysis entails attempting to merge them with components from adjacent layers and reassessing them with the above criteria. Characters composed of a number of colours (each colour expressed with different chromaticity and/or luminance) can be identified in this way.

4 Results and discussion

The resulting text extracted from the image of Figure 1 can be seen in figures 3–6. Each figure represents a separate layer and the information depicted (the constant blue colour indicates the background) shows the character-like components identified in that layer. Further filtering eliminates the solid rectangular and the very small components. The Hue histogram of the original

image has been calculated and for each of the candidate peaks selected the Luminance histogram of the corresponding pixels has been calculated. The extracted text in figures 3–5 is the result after these two stages (Hue then Luminance). The text of Figure 6 is the result of a further stage (Hue histogram peak selection of the second stage Luminance peak) as its colour resembles that of the background.

An example of an image of single hue with luminance variation can be seen in Figure 7. The extracted text can be seen in Figure 8. The text is extracted after examination of the Luminance histogram and merging of the different layers resulting from peak selection. This example would pose difficulties for other text extraction methods as the dark parts of the characters would be merged into the very similar dark colour of the background.

Overall, initial results on a representative test image set show that the method is feasible (extracts text in similar circumstances as other methods) and is capable of extracting text in more complex circumstances than other methods. One of the main aspects on which further work is concentrating is the refinement of the component merging process across adjacent layers. The current version of the method can extract relatively easily characters designed in gradient colour with the same hue (one resulting component) or with a limited number of different hues (more than one component segment in each layer). Complex colour changes where different hues are involved will result in a large number of component fragments across different layers. The extraction of such characters is a common problem across all approaches and a desirable goal.

The recognition of the text is the next open problem with a significant number of hurdles to overcome. The proposed text extraction method is arguably well placed to exploit the information contained in the layers for recognition, which can be incremental and collaborate with the extraction of characters.

In conclusion, a new method under development is described here for the extraction of text from images on WWW pages. It exploits the unique characteristics of the design of such images for human viewing on a screen by adopting a human perspective of colour perception. Preliminary results indicate that it is capable of extracting characters in more complex circumstances than existing methods.

Figure 3: Result from Hue -> Luminance

Figure 4: Result from Hue -> Luminance

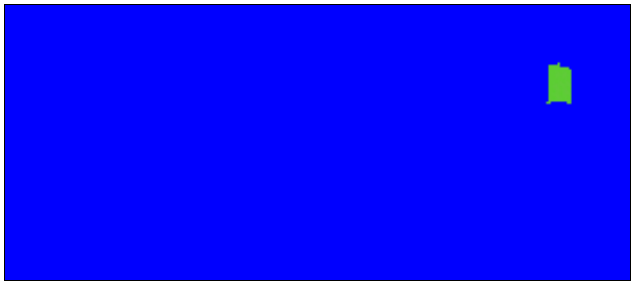

Figure 5: Result from Hue -> Luminance

Figure 6: Result from Hue -> Luminance -> Hue

Figure 7: Image with single hue.

New Adventures in Science Fiction

Figure 8: Text extracted from Figure 6.

References

[1] D. Amor, *The E-Business (R)evolution*, Prentice Hall, 1999.

[2] D. Lopresti and J. Zhou, "Document Analysis and the World Wide Web", Proceedings of the Workshop on Document Analysis Systems, Marven, Pennsylvania, October 1996, pp. 417–424.

[3] Search Engine Watch, http://searchenginewatch.com

[4] 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

[5] 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.

[6] A.K. Jain and B. Yu, "Automatic Text Location in Images and Video Frames", *Pattern Recognition*, vol. 31, no. 12, 1998, pp. 2055–2076.

[7] R.W.G. Hunt, *Measuring Colour*, John Wiley & Sons, 1987.

[8] G. Wyszecki and W. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2e, Wiley, New York, 1982.