Proceedings of the 1st International Workshop on Web Document Analysis (WDA2001), Seattle, USA, September 2001, pp. 35-38.

Text Extraction from Web Images Based on Human Perception and Fuzzy Inference

A. Antonacopoulos and D. Karatzas

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

Abstract

There is a significant need to extract and recognise the semantically-important text contained in images on Web pages. This paper proposes a new approach to text extraction from this special class of images. The method attempts to emulate closer than before the way humans perceive colour differences in order to differentiate between text and background regions. Pixels of similar colour (as humans see it) are merged into components and a fuzzy inference mechanism (using connectivity and colour distance features) is devised to group components into larger character-like regions.

1. Introduction

Semantically important entities on Web pages such as page titles, headers and menu items are routinely created in image form in order to add impact to the textual message they carry. The authors conducted a study [1] to assess the impact and consequences of text contained in images. The results agree with earlier findings [2] and clearly indicate an alarming trend. 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 textual description (ALT tags) of the images in question are incomplete, wrong or do not exist in 56% of the cases.

Although images contain some of the most important textual indexing terms of a Web Page, the plain text on the Web page remains the primary (the only one, in most cases) medium for indexing and searching. The fact that the search engines cannot access any text inside the images (see [3] for a list of indexing and ranking criteria for different search engines), introduces a significant problem since Web Pages may not be indexed and ranked correctly (since significant text—present only in image form—is missed).

The extraction and recognition of text in images on Web pages is more complex than traditional OCR on scanned documents. Images on the Web are optimised for viewing on a monitor screen and this fact introduces a number of problems including very small character sizes, low resolution, complex background, large number of colours, and quantization and compression artefacts [4]. Some of the above problems are shared with application such as text extraction from real scenes but in the latter domain there is relatively more control on a number of these problems (e.g., resolution).

Previous approaches for text extraction from colour images mainly assume that the characters are of uniform (or almost uniform) colour, work with a relatively small number of colours (reducing the original colours if necessary) and restrict all their operations in the RGB colour space [5][6][7]. A novel method that is based on information on the way humans perceive colour differences has been proposed by the authors [4]. That method works on full colour images and uses different colour spaces in order to approximate the way humans perceive colour. It comprises the splitting of the image into layers of similar colour by means of histogram analysis and the merging of the resulting components using criteria drawn from human colour discrimination observations.

This paper describes a new method for the extraction of text from Web images. In contrast to the authors' previous method [4], it is a bottom-up approach. This is an alternative method devised in an attempt to emulate even closer the way humans differentiate between text and background regions. Information on the ability of humans to discriminate between colours is used throughout the process. Pixels of similar colour (as humans see it) are merged into components and a fuzzy inference mechanism is devised to group components into larger character-like regions. The method is described in the next section and the paper concludes with a discussion of preliminary experimental results.

2. Text Extraction Method

The method starts by identifying colour connected components in the image, based on human perception merging criteria. The resulting components are merged, at the second step, into larger components with the aid of fuzzy techniques using a combined distance measure that takes into account topological and colour distance features between components.

2.1. Colour Connected Component Identification

Colour component labelling is performed in order to identify connected components of similar colour that will be used as the basis for the subsequent merging process. Although the merging process would still work with pixels rather than connected components as input, this first labelling step reduces significantly the computational load of the comparisons required at the merging step. The rationale is to avoid wrong groupings of pixels as – this is true for all bottom-up techniques – early errors have potentially a significant impact on the final results. To this effect, strict criteria are imposed, resulting in relatively smaller connected components.

The labelling algorithm used is a one-pass segmentation algorithm adapted from a previously proposed fast labelling algorithm used for binary images [8]. In the case of colour images as encountered here, a similarity comparison is performed instead of a straightforward match of black/white. The novel aspect of the comparison is that the criteria for similarity determination between pixels are based on human perception of colour differences. The measure of similarity follows the data from observations on the ability of humans to discriminate between different colours described in terms of wavelength and luminance [9]. According to this data, the distance between some colours is perceived as different than that between others having the same distance in the RGB space. In broad terms, neighbouring pixels are grouped together if their colour cannot be discriminated by a human. Otherwise, they belong to different components.

2.2. Combined Distance and Fuzzy Inference

At this step, colour connected components are merged to form larger components based on a fuzzy combined distance measure. The combined distance for any pair of components is based on two individual features: the *connections ratio*, which expresses topological properties, and the *colour distance* between the two components. The combined distance is defined using the fuzzy inference system developed.

The first feature used to express the degree of connectivity between two components is the *connections ratio*. A *connection* is defined as any one of the 8 neighbours of a single pixel. A connection of a pixel can be either internal (i.e. the neighbour is a pixel in the same component) or external (the neighbour is a pixel of another component), as shown in Fig. . 1.

Given any two components *a* and *b*, the connections ratio, $CR_{a,b}$ is defined as

$$CR_{a,b} = \frac{C_{a,b}}{\min(Ce_a, Ce_b)}$$

where $C_{a,b}$ is the number of external connections of component *a* to pixels of component *b* (it should be noted that $C_{a,b}=C_{b,a}$), and Ce_a and Ce_b refer to the total number of external connections of component *a* and *b*, respectively. The connections ratio is, therefore, the number of connections between the two components, divided by the total number of external connections of the component with the smaller boundary. The connections ratio ranges from 0-1.

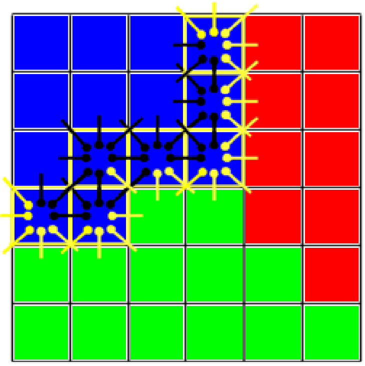

Figure 1. A connected component (blue) and its external (yellow) and internal (black) connections to its neighbouring components (red, green).

The second feature used is the *colour distance* between two components. In order for the colour distance to be related to the perceived colour difference, a colour system that is perceptually uniform must be used. Both the *RGB* and *XYZ* systems do not exhibit perceptual uniformity. The CIE has standardised two colour systems ($L^*a^*b^*$ and $L^*u^*v^*$) which significantly improve the perceptual non-uniformity of *XYZ*. The current implementation of the algorithm described here uses the $L^*a^*b^*$ space for the calculation of the colour distance. However, experiments are still being carried out by the authors using the $L^*u^*v^*$ and other colour systems.

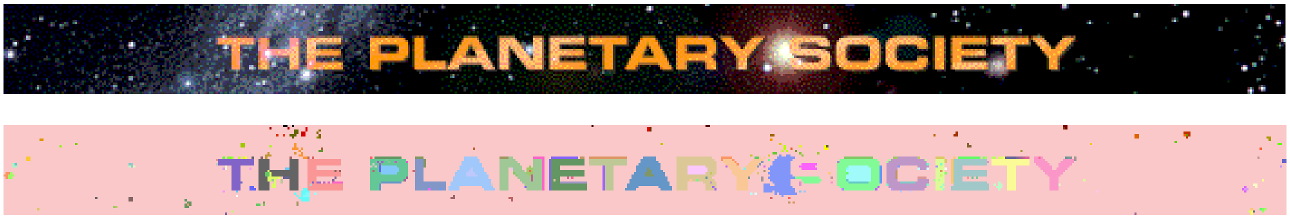

Figure 2. Example image and result

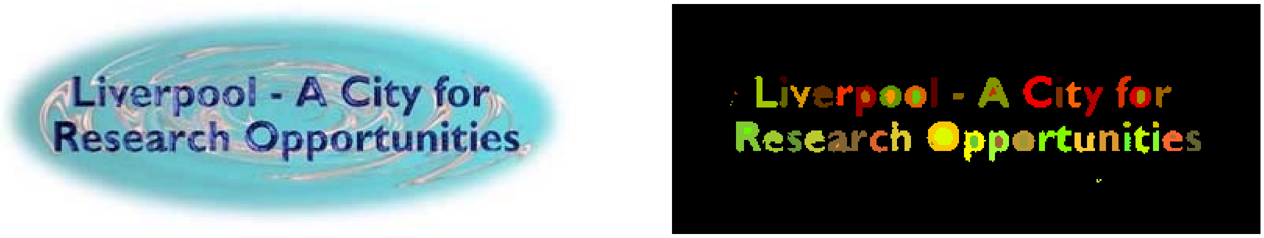
Liverpool - A City for Research Opportunities
Liverpool - A City for Research Opportunities

Figure 3. Example image and result

At first, a colour is assigned to each identified component, calculated as the average colour of the pixels of that component. This calculation is performed in the source colour system of the image i.e., the result is an *RGB* value. It should be noted at this point, that the average colour is chosen to represent the colour of a component as an attempt to cope with situations where characters of gradient colour fade into the background. The next step is to convert the colours of components to the *L*a*b** system. It is not feasible to convert from a device-dependant colour system to a device independent one without any extra knowledge about the hardware used, and the *RGB* colour system is not generally device independent. However between monitors that conform to the standard *Rec. 709* [10] within a specified tolerance, the *RGB* colours can be considered to be unvarying. The RGB_{709} colours are converted to the CIE *XYZ* system and subsequently converted to $L^*a^*b^*$ [11,12]. The colour distance used here is the Euclidean distance in the *L*a*b** space.

The *combined distance* between two components takes into account the above two features. It is defined within the fuzzy inference system developed with the aid of MatLab. It takes the two feature values as input and producing a value in the range of 0–1. Appropriate membership functions have been designed for both the *connections ratio* and the *colour distance*. The former takes into account that the components are usually parts of characters and, as characters are form by continuous strokes, the components should only share part of their border. The latter (colour distance) incorporates the results of ongoing experiments carried out by the authors to determine the minimum colour distance in the L*a*b* space required for a human to perceive a noticeable difference in colour. Appropriate membership functions are also defined for the combined distance giving output in five ranges (zero, small, medium, large, definite) allowing thus for flexibility in the definition of the rules for the inference system.

2.3. Merging Algorithm

The merging algorithm considers pairs of components and, based on the likelihood of two components belonging to the same character, combines them or not. An initial selection of components that could be parts of characters is made in terms of their size. In the current implementation all components whose size is up to 1/9th of the image size are initially considered.

For each component in the initial list, the combined distance is computed of that component and each neighbouring one and a sorted list of all the possible mergers is maintained. A merger between two components is possible if the combined distance between them is less than 0.5. The fuzzy inference system is designed in such a way that a tolerance of 0.5 for the combined distance is satisfactory for most of the cases. The algorithm proceeds to merge the components with the smallest distance, and recalculates only the combined distance between the new component that results from the merging and its neighbours, keeping the number of computations as low as possible. The algorithm continues merging the components with the smallest distance every time until there is no distance in the queue smaller than 0.5.

3. Results and Discussion

The algorithm described here is still in the development stage in terms of enhancement and refinement. Example images and corresponding results indicative of the performance of the method can be seen in Figures 2–3. Typical results for these types of images are between 80–95% accurate (character extraction). These figures are the result of manual evaluation. The development of an automated evaluation system (not a straightforward task due to the nature of the images) is currently being actively pursued. A critical aspect of the algorithm is the order of mergers, especially when the characters are of gradient colour fading into the background colour. In such cases, the background may be merged with part of the character (and the extraction rate becomes considerably lower). Current and further work is focussed on refining the merging decision process, identifying further features and fine-tuning the membership functions of the fuzzy inference system. Further topological and structural attributes of parts of characters are also being investigated and evaluated. Finally, possibilities for post-processing are currently being studied.

References

[1] A. Antonacopoulos, D. Karatzas and J. Ortiz Lopez, "Accessing Textual Information Embedded in Internet Images", *Proceedings of SPIE Internet Imaging II*, San Jose, USA, January 24-26, 2001, pp.198-205.

[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://www.searchenginewatch.com

[4] A. Antonacopoulos and D. Karatzas "An Anthropocentric Approach to Text Extraction from WWW Images", Proceedings of the 4th IAPR Workshop on Document Analysis Systems (DAS'2000), Rio de Janeiro, Brazil, December 2000, pp. 515–526.

[5] D. Lopresti and J. Zhou, "Locating and Recognizing Text in WWW Images", *Information Retrieval*, **2** (2/3), May 2000, pp. 177–206.

[6] 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.

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

[8] A. Antonacopoulos, "Page Segmentation Using the Description of the Background", Computer Vision and Image Understanding, Vol. 70, No 3, June, pp.350-369, 1998.

[9] R.E. Bedford and G.W. Wyszecki, "Wavelength Discrimination for Point Sources", Journal of the Optical Society of America, vol 48, no 2. February 1958.

[10] ITU-R Recommendation BT.709, Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange (1990), [formerly CCIR Rec. 709] (Geneva: ITU, 1990)

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

[12] K. McLaren, "The development of the CIE 1976 (L*a*b*) uniform colour-space and colour-difference formula", *Journal of the Society of Dyers and Colourists,* **92**, pp. 338-341