Fuzzy Segmentation of Characters in Web Images Based on Human Colour Perception

A. Antonacopoulos and D. Karatzas

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

Abstract. This paper describes a new approach for the segmentation of characters in images on Web pages. In common with the authors' previous work in this subject, this approach attempts to emulate the ability of humans to differentiate between colours. In this case, pixels of similar colour are first grouped using a colour distance defined in a perceptually uniform colour space (as opposed to the commonly used RGB). The resulting colour connected components are then grouped to form larger (character-like) regions with the aid of a fuzzy propinquity measure. This measure expresses the likelihood for merging two components based on two features. The first feature is the colour distance in the $L^*a^*b^*$ colour space. The second feature expresses the topological relationship of two components. The results of the method indicate a better performance than the previous method devised by the authors and comparable (possibly better) performance to other existing methods.

1 Introduction

Text is routinely created in image form (headers, banners etc.) on Web pages, as an attempt to overcome the stylistic limitations of HTML. This text, however, has a potentially high semantic value in terms of indexing and searching for the corresponding Web pages. As current search engine technology does not allow for text extraction and recognition in images (see [1] for a list of indexing and ranking criteria for different search engines), the text in image form is ignored. Moreover, it is desirable to obtain a uniform representation (e.g. UNICODE) of all *visible* text on a Web page. This uniform representation can be used by a number of applications such as voice browsing [2] and automated content analysis [3] for viewing on small screen devices such as PDAs.

There has been a provision for specifying the text included in images, in the form of ALT tags in HTML. However, a study conducted by the authors [4], assessing the impact and consequences of text contained in images indicates that the ALT tag strategy is not effective. It was found that the textual description (ALT tags) of 56% of images on Web pages was incomplete, wrong or did not exist at all. This can be a serious matter since, of the total number of words visible on a Web page, 17% are in image form (most often semantically important text). Worse still, 76% of these words in image form do not appear elsewhere in the encoded text. These results agree with earlier findings [5] and clearly indicate an alarming trend.

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 295–306, 2002.

It can be seen from the above that there is a significant need for methods to extract and recognise the text in images on Web pages. However, this is a challenging problem for the following reasons. First, these (sometimes complex) colour images tend to be of low resolution (usually just 72 dpi) and the font-size used for text is very small (about 5pt–7pt). Such conditions clearly pose a challenge to traditional OCR, which works with 300dpi images (mostly bilevel) and character sizes of usually 10pt or larger. Moreover, images on Web pages tend to have various artefacts due to colour quantization and lossy compression [6].

It should be mentioned that text in Web images is of quite different nature than text in video, for instance. In principle, although methods attempting to extract text from video (e.g., [7]) could be applied to a subset of Web images, they make restricting assumptions about the nature of embedded text (e.g., colour uniformity). As such assumptions are, more often than not, invalid for text in Web images, such methods are not directly discussed here.

Previous attempts to extract text from Web 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 [8][9][10]. A novel method that is based on information on the way humans perceive colour differences has been proposed by the authors [11]. 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 segmenting character regions in Web images. In contrast to the authors' previous method [11], 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 that uses a 'propinquity' measure is devised to group components into larger character-like regions.

The colour segmentation method and each of its constituent operations are examined in the next section and its subsections. Experimental results are presented and discussed, concluding the paper.

2 Colour Segmentation Method

The basic assumption of this paper is that, in contrast to other objects in general scenes, text in image form can always be easily separated (visually) from the background. It can be argued that this assumption holds true for all text, even more so for text intended to make an impact on the reader. The colour of the text in Web images and its visual separation from the background are chosen by the designer (consciously or subconsciously) according to how humans perceive it to 'stand out'.

To emulate human colour differentiation, a colour distance measure is defined in an alternative colour space. This distance measure is used first to identify colour connected components and then, combined with a new topological feature (using a fuzzy inference system), it is used to aggregate components into larger entities (characters).

Each of the processes of the system is described in a separate subsection below. First, the colour measure is described in the context of colour spaces and human colour perception. The connected components labelling process using this colour distance is described next. The two features (colour distance and a measure of spatial proximity) from which the new 'propinquity' measure is derived are presented in Section 2.3. Finally, the fuzzy inference system that computes the propinquity measure is the subject of Section 2.4 before the description of the last stage of colour connected component aggregation (Section 2.5).

2.1 Colour Distance

To model human colour perception in the form of a colour distance measure, requires an examination of the different colour spaces in terms of their perceptual uniformity. The *RGB* colour system, which is by far the most frequently used system in image analysis applications, lacks a straightforward measurement method for *perceived* colour difference. This is due to the fact that colours having equal distances in the *RGB* colour space may not necessarily be perceived by humans as having equal distances.[1] A more suitable colour system would be one that exhibits perceptual uniformity. The CIE (Commission Internationale de l'Eclairage) has standardised two colour systems ($L^*a^*b^*$ and $L^*u^*v^*$) based upon the CIE *XYZ* colour system [12][13]. These colour systems offer a significant improvement over the perceptual non-uniformity of *XYZ* [14] and are a more appropriate choice to use in that aspect than *RGB* (which is also perceptually non-uniform, as mentioned before).

The measure used to express the perceived colour distance in the current implementation of this method is the Euclidean distance in the $L^*a^*b^*$ colour space ($L^*u^*v^*$ has also been tried, and gives similar results). In order to convert from the *RGB* to the $L^*a^*b^*$ colour space, an intermediate conversion to *XYZ* is necessary. This is not a straightforward task, since the *RGB* colour system is by definition hardware-dependent, resulting in the same *RGB*-coded colour being reproduced on each system slightly differently (based on the specific hardware parameters). On the other hand, the *XYZ* colour system is based directly on characteristics of human vision (the spectral composition of the *XYZ* components corresponds to the colour matching characteristics of human vision) and therefore designed to be totally hardware-independent. In reality, the vast majority of monitors conform to certain specifications, set out by the standard *ITU-R recommendation BT.709* [15], so the conversion suggested by *Rec.709* can be safely used and is the one used for this method. The conversion from *XYZ* to $L^*a^*b^*$ is straightforward and well documented.

[1] For example, assume that two colours have RGB (Euclidean) distance δ. Humans find it more difficult to differentiate between the two colours if they both lie in the green band than if the two colours lie in the red-orange band (with the distance remaining δ in both cases). This is because humans are more sensitive to the red-orange wavelengths than they are to the green ones.

2.2 Colour Connected Component Identification

Colour connected component labelling is performed in order to identify components of similar colour. These components will form the basis for the subsequent aggregation process (see Section 2.5). It should be noted that although the aggregation process that follows would still work with pixels rather than connected components as input, using connected components significantly reduces the number of mergers and subsequently the computational load of the whole process.

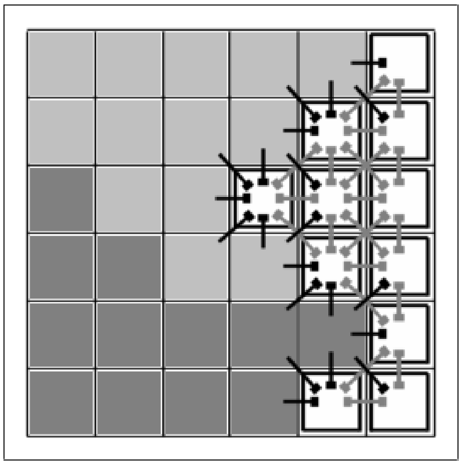

Fig. 1. A connected component (white) and its external and internal connections to its neighbouring components (shown in dark and light grey). Black lines indicate the external connections (to pixels belonging to different components) and light grey lines the internal connections (to pixels of the same component)

The idea behind this pre-processing step is to group pixels into components, if and only if a human being cannot discriminate between their colours. The rationale at this stage is to avoid wrong groupings of pixels as – this is true for all bottom-up techniques – early errors have potentially significant impact on the final results.

The identification of colour connected components is performed using a one-pass segmentation algorithm adapted from a previously proposed algorithm used for binary images [16]. For each pixel, the colour distance to its adjoining (if any) connected components is computed and the pixel is assigned to the component with which the colour distance has the smallest value. If the pixel in question has a distance greater than a threshold to all its neighbouring connected components, a new component is created from that pixel.

The threshold below which two colours are considered similar was experimentally determined and set to *20* in the current implementation. In fact, it was determined as the maximum threshold for which no character was merged with any part of the background. It should be noted, since the images in the training data set include cases containing text very similar to the surrounding background in terms of hue, luminance or saturation, this threshold is believed to be appropriate for the vast majority of text in Web images. Finally, the chosen threshold is small enough to conform to the opening statement that only colours that cannot be differentiated by humans should be grouped together.

2.3 Propinquity Features

The subsequent aggregation of the connected components produced by the initial labelling process into larger components is based on a fuzzy inference system (see next section) that outputs a *propinquity* measure. This measure expresses how close two components are in terms of colour and topology.

The propinquity measure defined here is based on two features: a colour similarity measure and a measure expressing the degree of 'connectivity' between two components. The colour distance measure described above (Section 2.1) is used to assess whether two components have perceptually different colours or not.

The degree of connectivity between two components is expressed by the *connections ratio* feature. A *connection* is defined here as a link between a pixel and any one of its 8-neighbours, each pixel thus having 8 connections. A connection can be either internal (i.e., both the pixel in question and the neighbour belong to the same component) or external (i.e. the neighbour is a pixel of another component). Figure 1 illustrates the external and internal connections of a given component to its neighbouring components.

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

$$CR_{a,b} = \frac{C_{a,b}}{\min(Ce_a, Ce_b)} \tag{1}$$

where $C_{a,b}$ is the number of (external) connections of component *a* to pixels of component *b*, and Ce_a and Ce_b refer to the total number of external connections (to all neighbouring components) of components *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*.

In terms of practical significance, the connections ratio is far more descriptive of the topological relationship between two components than other spatial distance measures (e.g., the Euclidean distance between their centroids). A small connections ratio indicates loosely linked components, a medium value indicates components connected only at one side, and a large connections ratio indicates that one component is almost included in the other. Moreover, the connections ratio provides a direct indication of whether two components are neighbouring or not in the first place, since it will equal zero if the components are disjoint.

2.4 Fuzzy Inference

A fuzzy inference system has been designed to combine the two features described above into a single value indicating the degree to which two components can be merged to form a larger one. The $L^*a^*b^*$ colour distance and the connections ratio described in the previous sections form the input to the fuzzy inference system. The output, called the *propinquity* between the two participating components, is a value ranging between zero and one, representing how close the two components are in terms of their colour and topology in the image. Each of the inputs and the output are coded in a number of membership functions described below, and the relationship between them is defined with a set of rules.

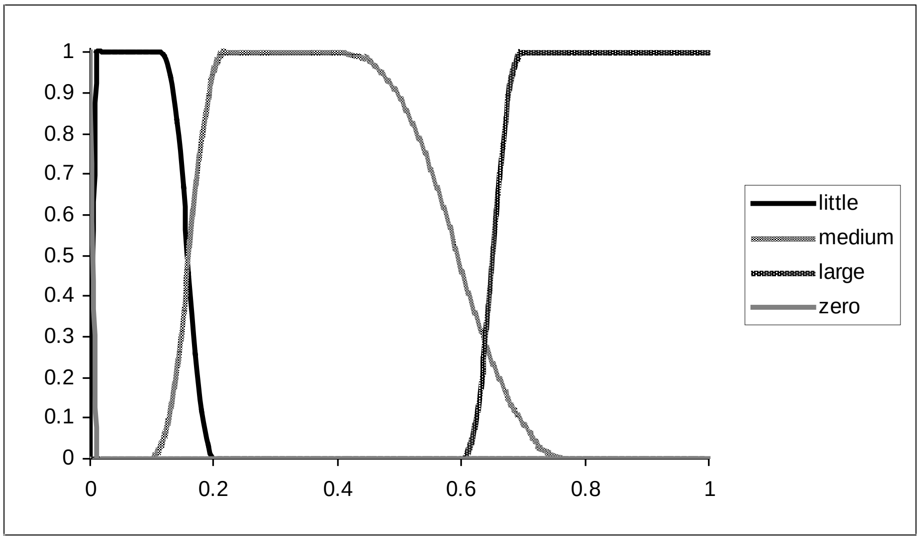

Fig. 2. Membership functions for connections ratio input

The membership functions defined for the connections ratio input can be seen in Figure 2. The components that should be combined are those that correspond to parts of characters. Due to the fact that characters consist of continuous strokes, the components in question should only partially touch each other. For this reason, a *medium* membership function is defined between *0.1* and *0.75*. It is considered advantageous for two components to have a connections ratio that falls in that range in order to combine them. This fact is reflected in the rules comprising the fuzzy inference system, which favour a connectivity ratio in the *medium* region, rather than one in the *small* or *large* regions. Furthermore, a membership function called *zero* is defined, in order to facilitate the different handling of components that do not touch at all, and should not be considered for merging.

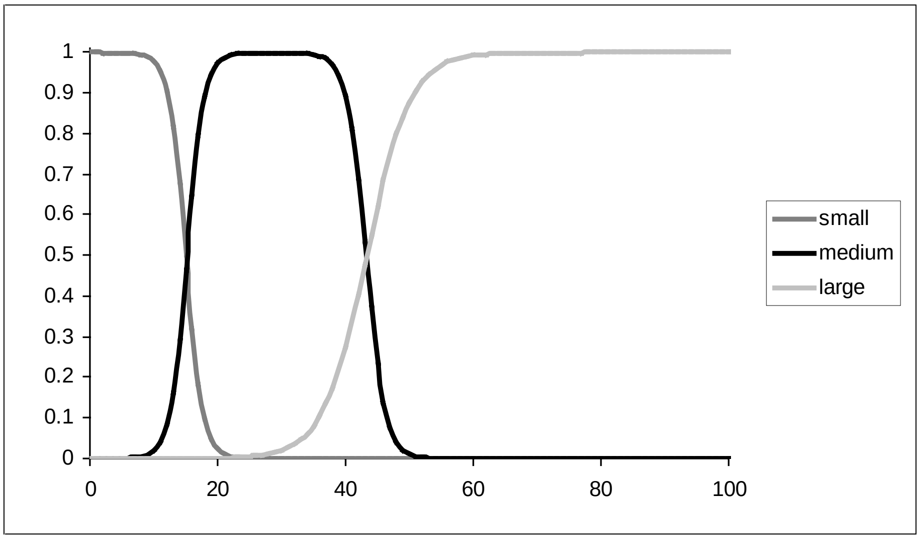

Fig. 3. The membership functions for the colour distance input

There are three membership functions defined for the $L^*a^*b^*$ colour distance input, namely *small*, *medium* and *large* (see Figure 3). The *small* membership function is defined between *0* and *15*. Colours having an $L^*a^*b^*$ distance less than *15* cannot be discriminated by a human being, therefore a colour distance falling in the small range is being favoured by the rules of the fuzzy inference system. In contrast, a large membership function has been defined for colour distances above *43*. Components having a colour distance in that range are considered as the most inappropriate candidates to be merged. The middle range, described by the medium membership function, is where there is no high confidence about whether two components should be merged or not. In that case, the rules of the system give more credence to the connections ratio feature. The thresholds of *15* and *43* were experimentally determined, as the ones that minimise the number of wrong mergers.

The single output of the fuzzy inference system, the propinquity, is defined with the help of five membership functions (see Figure 4). There are two membership functions at the edges of the possible output values range, namely *zero* and *definite*, and three middle range membership functions: *small*, *medium* and *large*. This set of membership functions allows for a high degree of flexibility in defining the rules of the system, while it encapsulates all the possible output cases.

The fuzzy inference surface, picturing the relationship defined by the rules of the system between the two inputs and the propinquity output can be seen in Figure 5. The fuzzy inference system is designed in such a way, that a propinquity of *0.5* can be used as a threshold in deciding whether two components should be considered for merging or not.

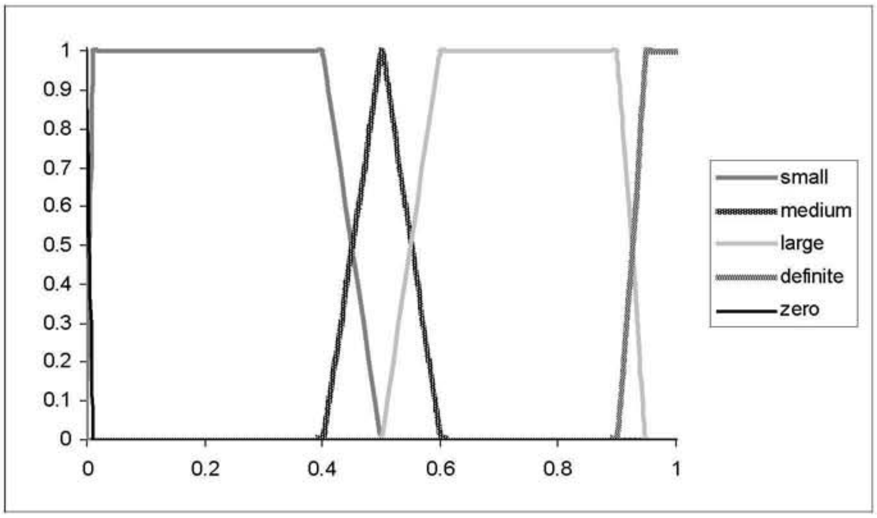

Fig. 4. The membership functions for the output (propinquity)

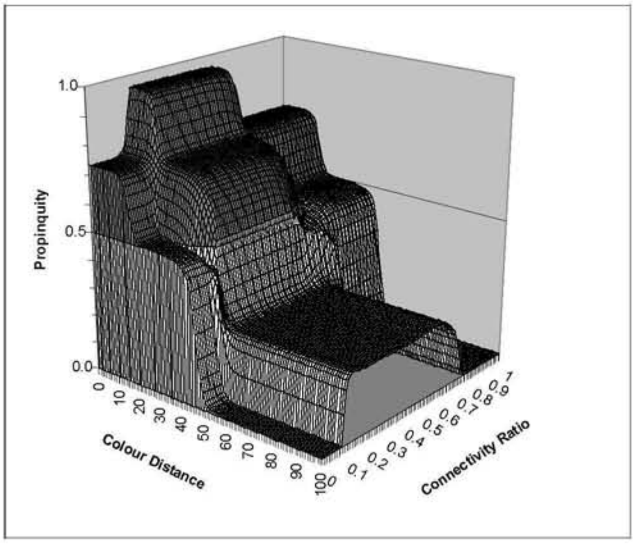

Fig. 5. The fuzzy inference system surface

2.5 Colour Component Aggregation

The merging algorithm considers pairs of connected components, and based on the propinquity output of the fuzzy inference system, combines them or not. All components produced by the initial colour connected components identification process are considered.

For each connected component, the propinquity to each of the neighbouring components is computed, and if it is greater than a set threshold, a possible merger is identified. A sorted list of all possible mergers is maintained, based on the computed propinquity value. The algorithm proceeds to merge the components with the largest propinquity value, and updates the list after each merger, including possible mergers between the newly created component and its neighbours. Only the necessary propinquity values are recalculated after each merger, keeping the number of computations to a minimum. The process continues in an iterative manner, as long as there are merger candidates in the sorted list having propinquity greater that the threshold. The threshold for propinquity is set (as a direct result of the design of the membership functions) to be *0.5*.

3 Results and Discussion

The colour segmentation method was evaluated using a variety of images collected from different Websites. The test set comprises 124 images, which are divided into four categories: (a) Multicoloured text over multicoloured background (24 images), (b) Multicoloured text over single-coloured background (15 images), (c) Single-coloured text over multicoloured background (30 images) and (d) Single-coloured text over single-coloured background (55 images). This distribution reflects the occurrence of images on Web documents. The number of colours in the images ranges from two to several thousand and the bits per pixel are in the range from 8 to 24. A width of four pixels was defined as the minimum for any character to be considered readable.

The evaluation of the segmentation method was performed by visual inspection. This assessment can be subjective for the following reasons. First, the borders of the characters are not precisely defined in most of the cases (due to anti-aliasing or other artefacts e.g. artefacts caused by compression). Second, no other information is available about which pixel belongs to a character and which to the background (no ground truth information is available for Web images). For this reason, in cases where it is not clear whether a character-like component contains any pixel of the background or not, the evaluator decides on the outcome based on whether by seeing the component on its own he/she can understand the character or not. The foundation for this is that even if a few pixels have been misclassified, as long as the overall shape can still be recognised, the character would be identifiable by OCR software.

The following rules apply regarding the characterisation of the results. Each character contained in the image is characterised as identified, partially identified or missed. Identified characters are those that are described by a single component. Partially identified ones are the characters described by more than one component, as long as each of those components contain only pixels of the character in question (not any background pixels). If two or more characters are described by only one component (thus merged together), yet no part of the background is merged in the same component, then they are also characterised as partially identified. Finally, missed are the characters for which no component or combination of components exists that describes them completely without containing pixels of the background as well.

The algorithm was tested with images of each of the four categories. In category (a) 223 out of 420 readable characters (53.10%) were correctly identified, 79 characters (18.57%) were partially identified and 119 characters (28.33%) were missed. In addition, out of the 487 non-readable characters of this category, the method was able to identify 245 and partially identify 129. In category (b) the method correctly identified 284 out of 419 characters (67.78%) while 88 (21.00%) were partially identified and 47 (11.22%) missed. There were no non-readable characters in this category. In category (c) 443 (72.74%) out of 609 readable characters were identified, 115 (18.88%) partially identified and 51 (8.37%) missed. In this category, the method was also able to identify 130 and partially identify 186 out of 388 non-readable characters. Finally, in category (d) 572 (73.71%) out of 776 readable characters were identified, 197 (25.39%) partially identified and 7 (0.9%) missed. In addition, 127 out of 227 non-readable characters were identified and 53 partially identified.

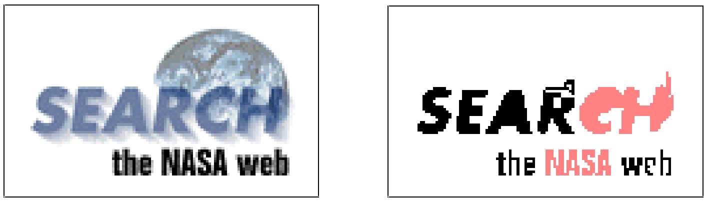

Fig. 6. An image containing gradient text blended with the background and the corresponding results

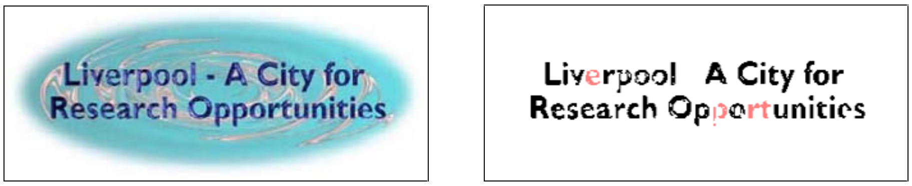

Fig. 7. An image containing multi-coloured characters over multi-coloured background and the corresponding results

The results mentioned above reflect the increasing difficulty in categories where the text and/or the background are multi-coloured. In figures 6 to 9, a number of images of the test set can be seen, along with the corresponding results. The black characters denote correctly identified ones, whereas the grey ones (red in the original) partially identified ones.

In conclusion, a new approach for the segmentation of characters in images on Web pages is described. The method is an attempt to emulate the ability of humans to differentiate between colours. A fuzzy propinquity measure is used to express the likelihood for merging two components, based on topological and colour similarity

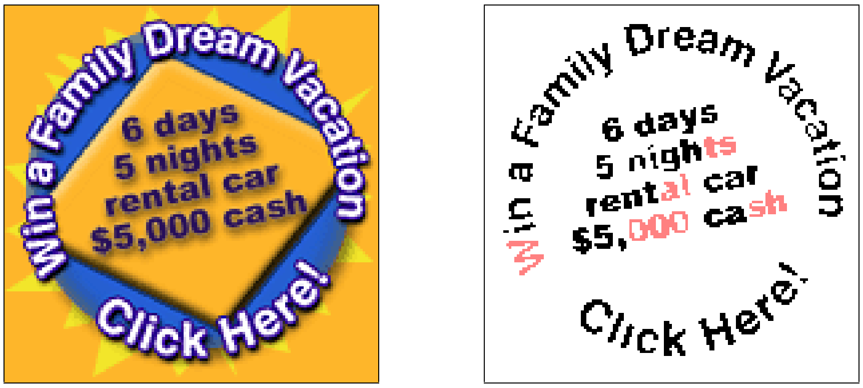

Fig. 8. An image containing shadowed and outlined characters and the corresponding results

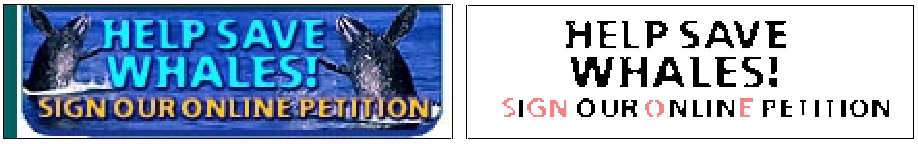

Fig. 9. An image containing single-colour characters over multi-coloured background and the corresponding results

features. The results of the method indicate a better performance than the previous method devised by the authors and comparable performance to other existing methods. Continuous work is concentrating on the possibilities to enhance the propinquity measure by adding more features and in the further optimisation of the fuzzy inference system. Results over a large test set indicate potential for better performance.

Acknowledgement. The authors would like to express their gratitude to Hewlett-Packard for their substantial equipment donation in support of this project.

References

1. Search Engine Watch, http://www.searchenginewatch.com
2. M.K. Brown, “Web Page Analysis for Voice Browsing”, *Proceedings of the 1st International Workshop on Web Document Analysis (WDA’2001)*, Seattle, USA, September 2001, pp. 59-61.
3. G. Penn, J. Hu, H. Luo and R. McDonald, “Flexible Web Document Analysis for Delivery to Narrow-Bandwidth Devices”, *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR’01)*, Seattle, USA, September 2001, pp. 1074–1078.
4. 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.

5. 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
6. D. Lopresti and J. Zhou, “Document Analysis and the World Wide Web”, *Proceedings of the 2nd IAPR Workshop on Document Analysis Systems (DAS’96)*, Marven, Pennsylvania, October 1996, pp. 417–424.
7. H. Li; D. Doermann and O. Kia, “Automatic text detection and tracking in digital video”, *IEEE Transactions on Image Processing*, vol. 9, issue 1, Jan. 2000, pp. 147-156.
8. D. Lopresti and J. Zhou, “Locating and Recognizing Text in WWW Images”, *Information Retrieval*, **2** (2/3), May 2000, pp. 177–206.
9. A.K. Jain and B. Yu, “Automatic Text Location in Images and Video Frames”, *Pattern Recognition*, vol 31, no. 12, 1998, pp.2055-2076.
10. 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.
11. 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.
12. R. C. Carter and E. C. Carter, "CIE L*u*v* Color-Difference Equations for Self-Luminous Displays," *Color Research and Applications*, vol. 8, 1983, pp. 252-253.
13. K. McLaren, "The development of CIE 1976 (L*a*b*) Uniform Colour Space and Colour-diference Formlua," *Journal of the Society of Dyers and Colourists*, vol. 92, 1976, pp. 338-341.
14. G. Wyszecki and W. S. Stiles, *Color Science - Concepts and Methods, Quantitative Data Formulas*. John Wiley, New York, 1967.
15. *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange*, ITU-R Recommendation BT.709 [formely CCIR Rec.709] Geneva, Switzerland: ITU 1990.
16. A. Antonacopoulos, “Page Segmentation Using the Description of the Background”, *Computer Vision and Image Understanding*, vol. 70, 1998, pp. 350-369.