Text Area Identification in Web Images

Stavros J. Perantonis[1], Basilios Gatos[1], Vassilios Maragos[1,3],
Vangelis Karkaletsis[2], and George Petasis[2]

[1] Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
{sper,bgat}@iit.demokritos.gr
http://www.iit.demokritos.gr/cil

[2] Software and Knowledge Engineering,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
{vangelis,petasis}@iit.demokritos.gr
http://www.iit.demokritos.gr/skel

[3] Department of Computer Science,
Technological Educational Institution of Athens,
122 10 Egaleo, Greece

Abstract. With the explosive growth of the World Wide Web, millions of documents are published and accessed on-line. Statistics show that a significant part of Web text information is encoded in Web images. Since Web images have special characteristics that sometimes distinguish them from other types of images, commercial OCR products often fail to recognize Web images due to their special characteristics. This paper proposes a novel Web image processing algorithm that aims to locate text areas and prepare them for OCR procedure with better results. Our methodology for text area identification has been fully integrated with an OCR engine and with an Information Extraction system. We present quantitative results for the performance of the OCR engine as well as qualitative results concerning its effects to the Information Extraction system. Experimental results obtained from a large corpus of Web images, demonstrate the efficiency of our methodology.

1 Introduction

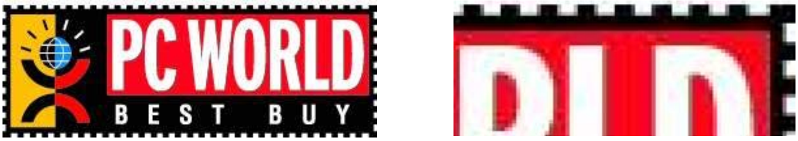
With the explosive growth of the World Wide Web, millions of documents are published and accessed on-line. The World Wide Web contains lots of information but even modern search engines just index a fraction of this information. This issue poses new challenges for Web Document Analysis and Web Content Extraction. While there has been active research on Web Content Extraction using text-based techniques, documents often include multimedia content. It has been reported [1][2] that of the

G.A. Vouros and T. Panayiotopoulos (Eds.): SETN 2004, LNAI 3025, pp. 82–92, 2004.

total number of words visible on a Web page, 17% are in image form and those words are usually the most semantically important.

Unfortunately, commercial OCR engines often fail to recognize Web images due to their special key characteristics. Web images are usually of low resolution, consist mainly of graphic objects, are usually noiseless and have the anti-aliasing property (see Fig. 1). Anti-aliasing smoothes out the discretization of an image by padding pixels with intermediate colors.

Several approaches in the literature deal with text locating in color images. In [3], characters are assumed of almost uniform colour. In [4], foreground and background segmentation is achieved by grouping colours into clusters. A resolution enhancement to facilitate text segmentation is proposed in [5]. In [6], texture information is combined with a neural classifier. Recent work in locating text in Web images is based on merging pixels of similar colour into components and selecting text components by using a fuzzy inference mechanism [7]. Another approach is based on information on the way humans perceive colour difference and uses different colour spaces in order to approximate the way human perceive colour [8]. Finally, approaches [9][10] restrict their operations in the RGB colour space and assume text areas of uniform colour.



(a) (b

Fig. 1. A Web image example (a) and a zoom in it (b) to demonstrate the web image key characteristics.

In this paper, we aim at two objectives: (a) Development of new technologies for extracting text from Web images for Information Extraction purposes and (b) Creation of an evaluation platform in order to measure the performance of all introduced new technologies.

Recently, some of the authors have proposed a novel method for text area identification in Web images [11]. The method has been developed in the framework of the EC-funded R&D project, CROSSMARC, which aims to develop technology for extracting information from domain-specific Web pages. Our approach is based on the transitions of brightness as perceived by the human eye. An image segment is classified as text by the human eye if characters are clearly distinguished from the background. This means that the brightness transition from the text body to the foreground exceeds a certain threshold. Additionally, the area of all characters observed by the human eye does not exceed a certain value since text bodies are of restricted thickness. These characteristics of human eye perception are embodied in our approach. According to it, the Web color image is converted to gray scale in order to record the transitions of brightness perceived by the human eye. Then, an edge extraction technique facilitates the extraction of all objects as well as of all inverted objects. A conditional dilation technique helps to choose text and inverted text objects among all objects. The criterion is the thickness of all objects that in the case of characters is of restricted value. Our approach is mainly based on the detected character edges and character thickness that are the main human eye perception characteristics.

The evaluation platform used in order to assess the performance of the proposed method for text area location was based on the Segmentation Evaluation Tool v.2 of the Computational Intelligence Laboratory (NCSR "DEMOKRITOS") [12]. We measured the performance of the proposed scheme for text area identification and recorded a significant facilitation in the recognition task of the OCR engine. Our methodology for text area identification has been fully integrated with an OCR engine and with an Information Extraction system (NERC module [13]). We present quantitative results for the performance of the OCR engine as well as qualitative results concerning its effects to the Information Extraction system. Experimental results obtained from a large corpus of Web images, demonstrate the efficiency of our methodology.

2 Text Area Location Algorithm

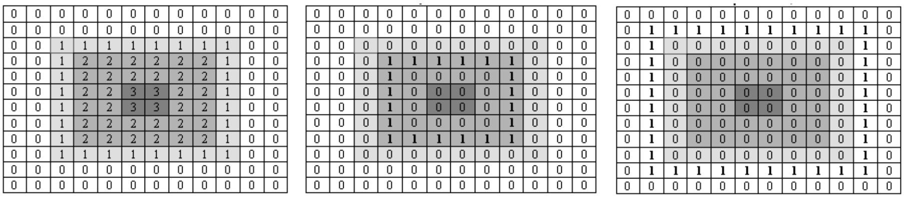
2.1 Edge Extraction

Consider a color Web image *I*. First, we covert it to the gray scale image *Ig*. Then, we define as *e* and e^{-1} the B/W edge and invert edge images that encapsulate the abrupt increase or decrease in image brightness:

$$e(x,y)=\begin{cases}1,\ \text{if}\ \exists(\mathrm{m,n}):\mathrm{Ig(m,n)}-\mathrm{Ig(x,y)}>\mathrm{D}\ \wedge\\ \qquad\qquad |\mathrm{m}-\mathrm{x}|<=\mathrm{d}\ \wedge\ |\mathrm{n}-\mathrm{y}|<=\mathrm{d}\\ 0,\ \text{otherwise}\end{cases} \tag{1}$$

$$e^{-1}(x,y)=\begin{cases}1,\ \text{if}\ \exists(\mathrm{m,n}):\mathrm{Ig(m,n)}-\mathrm{Ig(x,y)}<\mathrm{D}\ \wedge\\ \qquad\qquad |\mathrm{m}-\mathrm{x}|<=\mathrm{d}\ \wedge\ |\mathrm{n}-\mathrm{y}|<=\mathrm{d}\\ 0,\ \text{otherwise}\end{cases} \tag{2}$$

where *D* is the gray level contrast visible by the human eye and *d* defines the window at x,y in which we search for a gray level contrast. Fig. 2 shows an example for *e* and e^{-1} calculation.



0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	2	2	2	2	2	2	1	0	0
0	0	1	2	2	2	2	2	2	1	0	0
0	0	1	2	2	3	3	2	2	1	0	0
0	0	1	2	2	3	3	2	2	1	0	0
0	0	1	2	2	2	2	2	2	1	0	0
0	0	1	2	2	2	2	2	2	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(a)

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(b)

0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0	1	0
0	1	1	1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0

(c)

Fig. 2. (a) Gray scale image Ig, (b) edge image e and (c) invert edge image e-1 (parameters used: D=2, d=2).

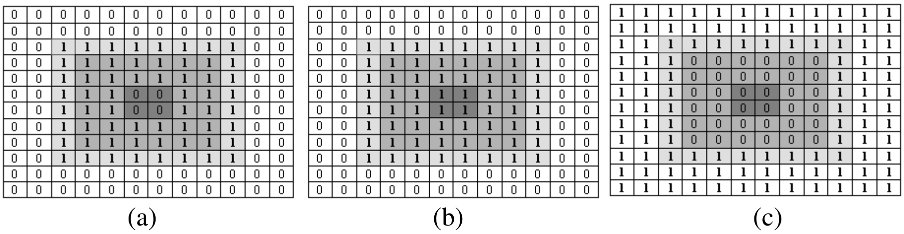
2.2 Object Identification

Objects are defined as groups of pixels that neighbor with edge pixels and have similar gray scale value. To calculate image objects, we proceed to a conditional dilation of edge images. A pixel is added only if it has a similar gray scale value in the original image *Ig*. The dimension of the structuring element defines the expected maximum thickness of all objects. Objects O_s and inverted objects O_s^{-1} are defined as follows:

$$O_s(x,y)=\begin{cases}1,\ \text{if}\ \exists(m,n): e(m,n)=1 \wedge |m-x|<=s\ \wedge\ |n-y|<=s \\ \qquad\qquad \wedge\,|Ig(x,y)-Ig(m,n)|<S \\ 0,\ \text{otherwise}\end{cases} \tag{3}$$

$$O_s^{-1}(x,y)=\begin{cases}1,\ \text{if}\ \exists(m,n): e^{-1}(m,n)=1 \wedge |m-x|<=s\ \wedge\ |n-y|<=s \\ \qquad\qquad \wedge\,|Ig(x,y)-Ig(m,n)|<S \\ 0,\ \text{otherwise}\end{cases} \tag{4}$$

where *s* the dimension of the structuring element and *S* is the expected maximum difference in gray scale values within the same object. Fig. 3 shows an example for O_s and O_s^{-1} calculation.



0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	0	0	1	1	1	0	0
0	0	1	1	1	0	0	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(a)

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(b)

1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1

(c)

Fig. 3. For the example of fig. 2 we calculate object O_1 (a), object O_n, n>1 (b) and object O_1^{-1} (c) (parameters used: S=1).

2.3 Text Identification

The above conditional dilation technique applied with several iterations (several values for the structuring elements) helps to choose text and inverted text objects among all objects. The criterion is the thickness of all objects that in the case of characters is of restricted value.

Let *P(f)*, the set of points of a b/w image *f*:

$$O(f) = \{(x,y):f(x,y)=1\} \tag{5}$$

$p_i(f)$, the set of points of all the connected components that comprise image *f*:

$$P(f) = \cup\ p_i(f) \tag{6}$$

S$(p_i(f))$, the number of pixels of the connected component, $E(p_i(f))$, the set of background points that have a 4-connected relation with the connected component, S$(E(p_i(f)))$, the number of pixels of $E(p_i(f))$, and $C(p_i(f))$, the category a connected component belongs to:

$$C(p_i(f)) = \text{TEXT or OTHER CATEGORY} \tag{7}$$

A connected component of image object O_n is classified as text region if while increasing n the set of background pixels that have a 4-connected relation with the connected component remains almost the same (see the example of Fig. 3b where object O_n remains the same for n>1):

$$\begin{aligned} C(p_i(O_n)) = \text{TEXT if } \exists j: (p_i(O_n) \subseteq p_j(O_{n+1}) \text{ AND} \\ S(E(p_i(O_n)) \cap E(p_j(O_{n+1}))) / S(E(p_i(O_n))< s) \text{ AND } n<N \end{aligned} \tag{8}$$

where N depends on the maximum expected letter thickness and s is the allowed tolerance in changes of the 4-connected background pixel set. The reason we trace the changes to the 4-connected background pixels and not to the foreground pixels is that due to dilation with a larger structuring element, the connected components may be joined together. In the same way, we define the condition for locating inverse text objects.

At Fig. 4 the flowchart of the proposed method is demonstrated.

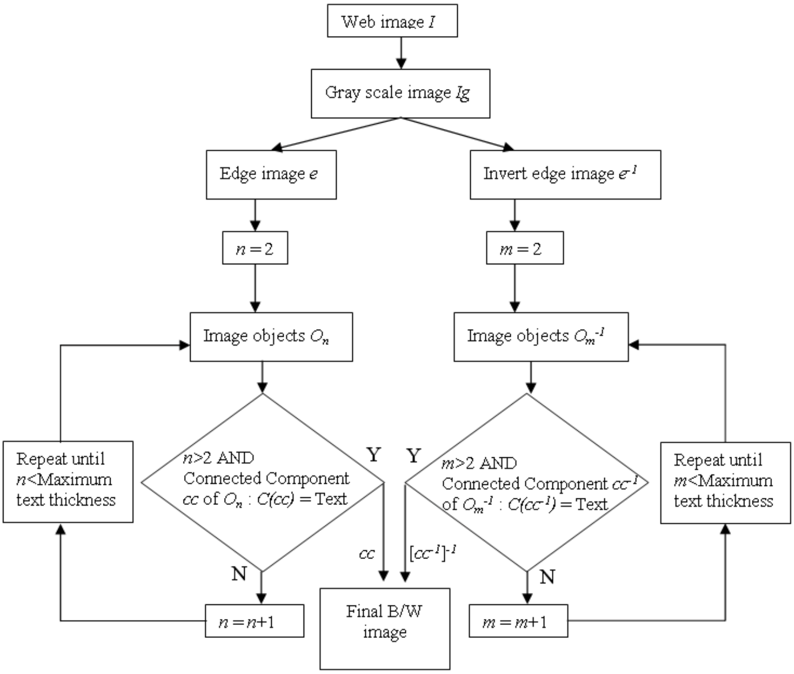
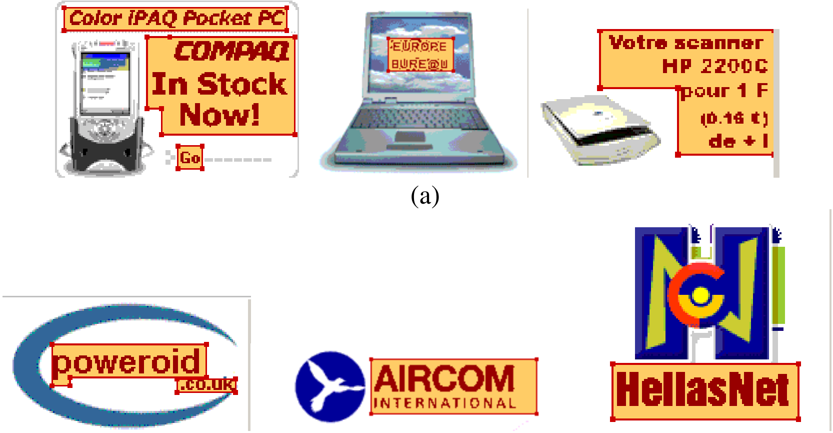

Web image *I*
Gray scale image *Ig*
Edge image *e*
Invert edge image *e*⁻¹
n = 2
m = 2
Image objects *O*ₙ
Image objects *O*ₘ⁻¹
Repeat until *n*<Maximum text thickness
n>2 AND Connected Component *cc* of *O*ₙ : *C(cc)* = Text
Y
Y
m>2 AND Connected Component *cc*⁻¹ of *O*ₘ⁻¹ : *C(cc*⁻¹*)* = Text
Repeat until *m*<Maximum text thickness
N
cc
[*cc*⁻¹]⁻¹
N
n = *n*+1
Final B/W image
m = *m*+1

Fig. 4. Text area identification algorithm flowchart.

3 System Evaluation

3.1 Corpus Preparation

The corpus for the evaluation of the proposed technique was prepared by selecting more than 1100 images from English, French, Greek and Italian Web pages. These images contain text, inverse text and graphics and concern laptop offers and job offers. In order to record the performance of the proposed method for text area location we annotated the text areas for all images (see Fig.5) using the Ground Truth Maker v.1 of the Computational Intelligence Laboratory (NCSR “DEMOKRITOS”) [11].



(a)

(b)

Fig. 5. Example of the ground truth text annotations: (a) From the laptop offers domain (b) from the job offers domain.

3.2 Evaluation Methodology

The proposed technique for text area identification in Web images has been implemented and tested with the large Web image corpus. We compared the results obtained by the well-known OCR engine FineReader 5 (FineReader) with and without applying our text area location technique. FineReader which has come out on top in major OCR comparative tests, can recognize the structure of a document including columns, graphic inserts and table formatting and can readily retain the page layout. It is also very effective in recognizing characters in different languages.

In order to record the performance of the proposed method for text area location we used the Segmentation Evaluation Tool v.2 of the Computational Intelligence Laboratory. We created a ground truth set with the annotations of the text areas. The performance evaluation method used is based on counting the number of matches between the text areas detected by the algorithm and the text areas in the ground truth. We calculated the intersection of the ON pixel sets of the result and the ground truth images.

Let I be the set of all image points, G the set of all points inside the ground truth text regions, R the set of all points of the result text regions and T(s) a function that counts the elements of set s. For every ground truth region we exclude all points that have approximate the same color with the surrounding of the annotation area. Detection rate and recognition accuracy are defined as follows:

$$\text{DetectionRate} = \frac{T(G \cap R \cap I)}{T(G)} \tag{9}$$

$$\text{RecognitionAccuracy} = \frac{T(G \cap R \cap I)}{T(R)} \tag{10}$$

A performance metric for text location can be extracted if we combine the values of detection rate and recognition accuracy. We used the following Text Detection Metric (TDM):

$$\text{TDM} = \frac{2\text{DetectionRate}\ \text{RecognitionAccuracy}}{\text{DetectionRate} + \text{RecognitionAccuracy}} \tag{11}$$

The evaluation strategy we followed concerns three main tasks: (a) Evaluation of the text locating module (b) evaluation of the OCR result after applying our text locating module, and (c) evaluation of the performance of an information extraction system using the OCR results.

3.3 Evaluation of the Text Locating Module

The evaluation results concerning the performance of the text location module for the laptop offers and the job offers domains are shown in tables 1 and 2.

3.4 Evaluation of the OCR Result after Applying Our Text Locating Module

In almost all cases, the recognition results were improved after applying our text area identification technique. A list of OCR results with and without the text extraction tool are presented in Table 3.

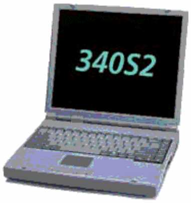
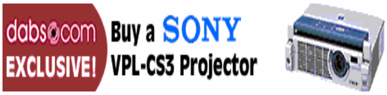
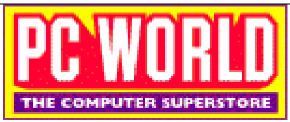
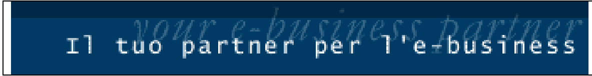
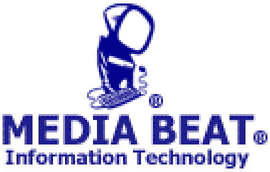
Table 1. Text location evaluation results for the laptop offers domain.

	Detection Rate	**Recognition Accuracy**	**Text Detection Metric**
English web image corpus	85,08	61,53	71,41
French web image corpus	84,32	61,61	71,20
Greek web image corpus	80,93	61,73	70,03
Italian web image corpus	78,41	61,50	68,93
TOTAL	**83,58**	**61,58**	**70,91**

Table 2. Text location evaluation results for the job offers domain.

	Detection Rate	Recognition Accuracy	Text Detection Metric
English web image corpus	72,67	61,84	66,81
French web image corpus	78,50	74,16	76,27
Greek web image corpus	79,58	64,83	71,46
Italian web image corpus	78,33	66,95	72,19
TOTAL	**77,31**	**66,57**	**71,54**

Table 3. OCR results with and without the text extraction tool.

	FineReader	Text extraction + FineReader
	-	340S2
	SONY 1VPL-CS3 Projector	da*sOc*m exclusive! Buy a SONY VPL-CS3 Projec-tor
	• ι π «*-):^·	PC WORLD THE COMPUTER SUPER-STORE
	jmrfc flrηsm &wtm	il tuo partner per l'e-business
Consultation des offres d'emploi	-	.Consultation des offres d'emploi
	MEDIA BEAT» I riformatici n Ttch nalóg y	MEDIA BEAT* Information Technology

A quantitative evaluation of the performance of the text extraction and preprocessing tool in combination with the OCR engine in terms of detection rate and recognition accuracy is shown in table 4.

Table 4. Evaluation of the performance of the text extraction and preprocessing tool in combination with the OCR engine in terms of detection rate and recognition accuracy.

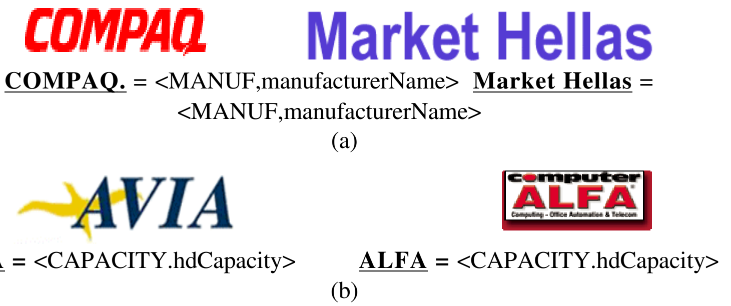
Laptop offers domain			
Detection Rate		Recognition Accuracy	
FineReader	Text extraction + FineReader	FineReader	Text extraction + FineReader
19.11%	22.09%	75.03%	74.13%
Job offers domain			
Detection Rate		Recognition Accuracy	
FineReader	Text extraction + FineReader	FineReader	Text extraction + FineReader
27.06%	33.63%	71.13%	70.28%

3.5 Evaluation of the Performance of the Information Extraction System Using the OCR Results

The evaluation results concerning the performance of the information extraction system (NERC module [13]) after adding to the web text information the OCR results show that:

- For the words added by the OCR procedure, 30% are correctly classified by the NERC module while the 70% of it are misclassified.
- If we had the perfect OCR engine with 100% recognition rate, then we would have a 45% correct classification by the NERC module while the 55% of it would be misclassified.
- From the above two remarks, we can state that the proposed text extraction and preprocessing module working with an OCR engine adds textual information to the NERC module and produces 66% of the correct results we would have if we used an 100% correct OCR scheme.

Some examples of correct classification results and misclassifications of the information extraction system (NERC module) are shown in figure 5.



COMPAQ. = <MANUF,manufacturerName> **Market Hellas** = <MANUF,manufacturerName>

(a)

AVIA = <CAPACITY.hdCapacity> **ALFA** = <CAPACITY.hdCapacity>

(b)

Fig. 6. Results from the information extraction system (NERC module). (a) Correct classification results (b) Misclassifications.

4 Concluding Remarks

The evaluation results show that many cases, where text is present as part of an image, are recovered by our text location algorithm. Moreover, it must be stressed that our method not only locates text areas, but it also preprocesses the characters present in them, so that the OCR engines are significantly facilitated in their recognition task.

The quantitative evaluation of the performance of the text extraction and preprocessing tool in combination with the OCR engine in terms of detection rate and recognition accuracy shows an approximate 20% increase in Recognition Rates. On the other hand, the evaluation results concerning the performance of text locating after applying our extraction and preprocessing tool module show that we have satisfactory results with more that 70% success. The main reason we did not achieve higher recognition rates is that we used the well-known OCR engine FineReader that is not oriented to work with low resolution images. Our future work concerning the improvement of our text extraction tool involves integration with a low resolution oriented OCR engine.

References

1. Antonacopoulos, A., Karatzas, D., Ortiz Lopez, J.: Accessing Textual Information Embedded in Internet Images. SPIE Internet Imaging II, San Jose, USA (2001) 198-205
2. Lopresti, D., Zhou, J.: Document Analysis and the World Wide Web. Workshop on Document Analysis Systems, Marven, Pennsylvania (1996) 417-424
3. Jain, A. K., Yu, B.: Automatic Text Location in Images and Video Frames. Pattern Recognition, Vol. 31, No. 12 (1998) 2055-2076
4. Huang, Q., Dom, B., Steele, D., Ashley, J., Niblack, W.: Foreground/background segmentation of color images by integration of multiple cues. Computer Vision and Pattern Recognition (1995) 246-249
5. Li, H., Kia, O., Doermann, D.: Text enhancement in digital video. Doc. Recognition & Retrieval VI (IS&SPIE Electronic Imaging'99), San Jose, Vol. 3651 (1999) 2-9
6. Strouthopoulos, C., Papamarkos, N.: Text identification for document image analysis using a neural network, Image and Vision Computing, Vol. 16 (1998) 879-896
7. Antonacopoulos, A., Karatzas, D.: Text Extraction from Web Images Based on Human Perception and Fuzzy Inference. 1st Int'l Workshop on Web Document Analysis (WDA 2001), Seattle, USA (2001) 35-38
8. Antonacopoulos, A., Karatzas, D.: An Anthropocentric Approach to Text Extraction from WWW Images. 4th IAPR Workshop on Document Analysis Systems (DAS2000), Rio de Janeiro (2000) 515-526
9. Antonacopoulos, A., Delporte, F.: Automated Interpretation of Visual representations: Extracting textual Information from WWW Images. Visual Representations and Interpretations, R. Paton and I. Neilson (eds.), Springer, London (1999)
10. Lopresti, D.,Zhou, J.: Locating and Recognizing Text in WWW Images. Information Retrieval, Vol. 2 (2/3) (2000) 177-206

11. Perantonis, S. J., Gatos, B., Maragos, V.: A Novel Web Image Processing Algorithm for Text Area Identification that Helps Commercial OCR Engines to Improve Their Web Image Recognition Efficiency. Second International Workshop on Web Document Analysis (WDA2003), Edinburgh, Scotland (2003).
12. Antonacopoulos, A., Gatos, B., Karatzas, D.: ICDAR 2003 Page Segmentation Competition. 7th International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, Scotland (2003)
13. Petasis G., Karkaletsis V and Spyropoulos C. D: Cross-lingual Information Extraction from Web pages: the use of a general-purpose Text Engineering Platform. 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003), Borovets, Bulgaria (2003)