TOWARDS IDENTIFICATION OF VERY LOW RESOLUTION, ANTI-ALAISED CHARACTERS

Farshideh Einsele, Jean Hennebert and Rolf Ingold
Department Of Informatics, University Of Fribourg, Switzerland
{farshideh.einsele, jean.hennebert, rolf.ingold}@unifr.ch

ABSTRACT

Current Web indexing technologies suffer from a severe drawback due to the fact that web documents often present textual information that is encapsulated in digital images and therefore not available as actual coded text. Moreover such images are not suited to be processed by existing OCR software, since they are generally designed for recognizing binary document images produced by scanners with resolutions between 200-600 dpi, whereas text embedded in web images is often anti-aliased and has generally a resolution between 72 and 90 dpi. The presented paper describes two preliminary studies about character identification at very low resolution (72 dpi) and small font sizes (3-12 pts). The proposed character identification system delivers identification rates up to 99.93% for 12'600 isolated character samples and up to 99.89% for 300'000 character samples in context.

1. INTRODUCTION

The explosive growth of the World Wide Web has resulted in a colossal data collection with billions of electronic documents. These documents contain images with textual information providing very high semantic value which could be used for indexing. Of the total number of words visible on a WWW page, 17% are in image form and 75% of these words do not appear elsewhere in the encoded text [1]. Furthermore, the ALT tag, which is recommended for describing an image in HTML language, is not or wrongly used [2]. However, current search engines do not use this valuable information. The reason is that current OCR technologies are not capable of extracting and recognizing such embedded text accurately. OCRs perform very well on images with 200 dpi (dot per inch) or more. But text contained in WWW-images has a resolution between 72 and100 dpi and small point sizes (i.e. less than 12 points) cannot be recognized in an accurate manner. Therefore, a significant part of electronic information is not considered for indexing. A so called Web-Image-OCR could immensely overcome this drawback.

Several studies have been done by D. Lopresti and J. Zhou [3-6]. Their work is concentrated mostly on pre-processing of the embedded textual information and passing them through the existing commercial OCRs. They report a character recognition rate between 80-92% and a text extraction rate of 68.3%.

Another approach has been done by A. Antonacopolous and al. [7-9]. They introduce methods for extracting characters based on the way humans perceive color differences and achieved extraction rates between 80 and 95%.

Notwithstanding, these studies dealt with bi-level (black and white) images and did not address the variability of text embedded in web images.

The ultimate goal of our research is to develop a Web-Image OCR specially designed for very low resolution rendered text images with small point sizes.

The novelty of the presented paper is to present results on an identification experiment that considers the various sources of variability of very low resolution characters. The presented study has two major parts.

The first part is about a study on isolated characters that are anti-aliased with small font sizes (6-9 pts) and aims at determining the identification performance of such individual characters. Results of several experiments on 168 fonts are presented.

The second part reports of a study about identification of the same kind of characters, which were in context of words. In spite of recognition of text by resolutions more than 200 dpi, there is no space available between characters Therefore, none of the known segmentation methods, for instance, connected components analysis, aspect ratio analysis, profile analysis, etc. can be used to isolate each character for the recognition process[10]. Thus segmentation and recognition tasks cannot be separated. For an accurate performance evaluation of our character identification system, we assume in this study that segmentation is known. Under this assumption we have conducted two systematical classification experiments. The results of these experiments on a database including 300'000 characters are presented with our interpretation, conclusion. Both studies apply the Bayesian classification method using multivariate density functions.

The remainder of this study is organized as follows:

In section 2, we introduce the sources of variability encountered in text embedded in web-images. In section 3, we present the results of a study about identification of isolated characters and section 4 presents the results of a study about identification of individual characters in context. This paper ends up with our conclusions and a short sketch of our future work.

2. DIFFICULTIES FOR WEB-IMAGE-OCR

2.1 Sources of variability

In this chapter we introduce sources of variability in characters at very low resolution with small font sizes:

2.1.1. Anti-aliasing

Fig. 1 illustrates the images of the bi-level "A" and the anti-aliased "A". Anti-aliasing is a rendering method to smooth edges and diagonals by very low resolution characters, i.e. when few pixels are available for rendering the image. Anti-aliasing uses 256 gray levels hoping to profit from the way our eyes tend to average adjacent pixels.

1-4244-0779-6/07/$20.00 ©2007 IEEE
IEEE International Symposium on Signal Processing and its Applications (ISSPA'07), Sharjah, United Arab Emirates. 2007. pp. 1-4

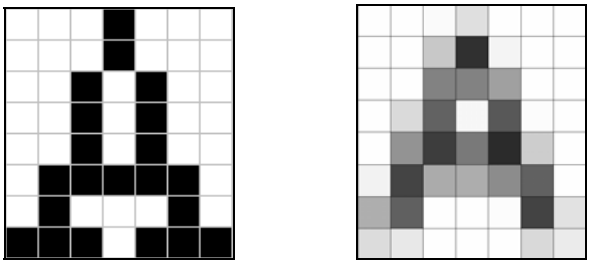

Fig.1: The images of the bi-.level "A" & the anti-aliased "A"

2.1.2 Grid-alignment

As is illustrated in Fig.2, the anti-aliased image of the same character in web images varies according to the sampling grid, that we call the grid alignment.

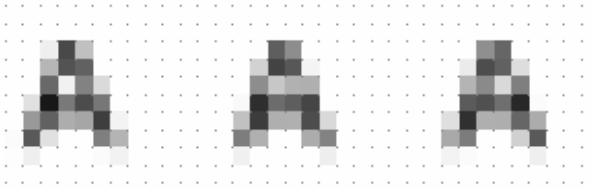

Fig. 2: Variability of "A" by different horizontal shifts

2.1.3. Influence of adjacent characters

The images of characters in context are influenced by their adjacent characters at their left and right borders; this is another source of variability, since the images of characters in context vary from the images of their isolated counterparts as illustrated in Fig.3:

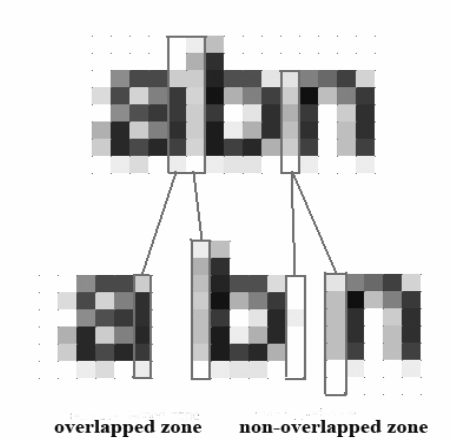

Fig. 3: the noisy zones by characters in context

2.2 Segmentation problem

As the Fig. 4 illustrates, there are no character interspaces available to segment characters within the word "School". This is a big challenge as segmentation and recognition can't be separated. To develop a Web-Image-OCR, segmentation and recognition have to be combined in the same process. However, in this paper we do not address this problem.

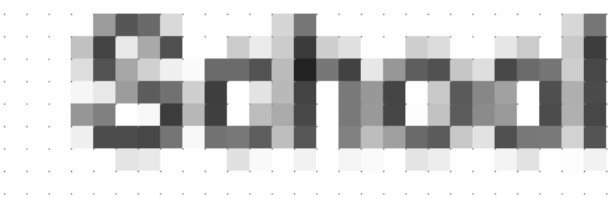

Fig. 4: A typical word extracted from a web image

3. DESCRIPTION OF THE IDENTIFICATION SYSTEM

As stated earlier, we are focusing in this work on single isolated character identification. The more complex task of full word recognition is not addressed here. More specifically, our main objective in this work is to investigate the impact of the different rendering variabilities on the identification rate of single characters. We also would like to determine if we can build optimized training sets that reach fair levels of identification performances on the most common types of font families and font styles.

We have set up a character identification system as illustrated in Fig.5. With this system, one test corresponds to an identification task that aims to determine which character is the most probable given a character image as input. In our tests, character images are supposed to be gray-level and to contain lower-case letters.

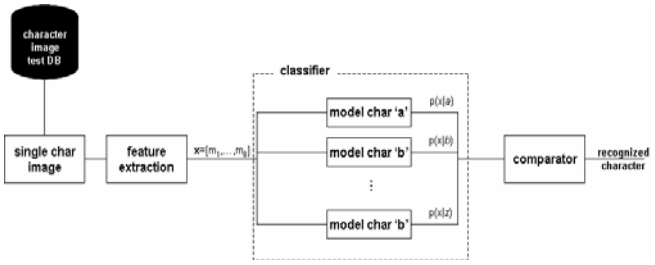

Fig. 5: Scheme of the single character identification system

The first component of the system is a feature extraction front-end that computes 7 features on the grey-level image. We decided to use geometrical central moments μ_{00}, μ_{11}, μ_{20}, μ_{02}, μ_{21}, μ_{12} *and* μ_{22}, which play an important role in shape analysis due to their translation invariant nature. Central moments are defined as follows:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q \, p(x,y) \quad (1)$$

with

$$\bar{x} = \frac{m_{10}}{m_{00}}, \quad \bar{y} = \frac{m_{01}}{m_{00}} \quad (2)$$

and

$$m_{pq} = \sum_x \sum_y x^p y^q \, p(x,y) \quad (3)$$

where x, y represent the coordinates and $p(x,y)$ the pixel value at (x,y) The feature vector x =$[m_1, \ldots . m_7]$ is then fed into M parallel models with M equal to the number of different characters to recognize. In the first study we consider upper &lower case letters without any accentuation (M=52), whereas in the second study we just consider lower case letters (M=26). Each model is based on a single multivariate Gaussian that estimates the probability density function of **x** given the character category c_i

$P(x|c_i) = N(x,\mu_i,\Sigma_i)$

in which $i=1 \ldots M$ is the index of the character category, and in which the Gaussian density N is parameterized by a mean D x 1 vector μ_i and a D x D covariance matrix Σ_i (in 1st study D=7, in 2nd study D=8). With this modelling, we make the implicit assumption that the moments are distributed according to a normal density, which is potentially not the case for all characters and fonts. Using Gaussian mixture models (GMMs)

instead of single Gaussian would allow us to relax this assumption but they are not investigated here.
Classification can then be performed in the comparator block of Fig. 5 with

$$I^* = argmax_{i=1...M}\, p(x|c_i)$$

Where I^* is the index of the winner character category. In this last equation, we make the assumption that character priors are all equal which is of course not the case in practice.
For the different experiments described in the next two sections, we generate training and testing databases of isolated characters using a flexible rendering procedure that allows us to explore the impacts of the different sources of variability:

- Two different anti-aliasing algorithms, the one of Photoshop from Adobe and the one included in the Java graphical library(java.awt) [11]
- Different grid alignments
- Contextual noise that can be simulated by extracting single characters from words (nota: this extraction is performed by exploiting the a priori knowledge of font metric information

4. STUDY OF ISOLATED CHARACTERS

The objective of this preliminary study is to evaluate the performance of our character identification system for isolated characters, which are without contextual noise as described in section 2.

4.1 Evaluation experiments

Three different evaluation tests have been done for classification of isolated characters. This study has been published in [12].

4.1.1. General study:

We have conducted a systematical study on 168 fonts, which consisted of 3 serifs, 3 non-serif fonts, with point sizes between 3-9 points and 4 different font styles. Table I presents the global recognition rates according to the font size. The experiment has shown very high accuracy for font sizes higher or equal to 5 points. The recognition is drastically reduced for font sizes lower than 5 points.

Table I: Recgnition results for 6 font types

Font Size	9	8	7	6	5	4	3
	99.93	99.92	99.99	99.82	99.84	99.66	92.59

4.1.2. Influence of the rendering method

We performed a classification experiment to determine the impact of the rendering method on the character identification results. We have fed our character identification system with both features gained from our two rendering methods (section3), and gained character identification rates of 99.95% for Java patterns and 99.93% for Photoshop patterns. Theses results have lead to the conclusion that using the mixed training sets enables to build a classifier, which delivers reliable results independently of their rendering methods.

4.1.3. Multi-font experiment

As the above experiments were done in a mono- font context, we decided to measure the character identification rate by considering multi-font classifiers. We separated serif and sans serif fonts with the same rule as the above test and observed a character identification rate of 98.48% for serif fonts and 98.5% for sans serif fonts. This result shows that we can build reliable multi-font classifiers by grouping serif and sans-serif fonts.

4.2 Conclusion

The presented study shows that if the characters are isolated, they can be recognized very accurately. The next step is studying the influence of adjacent characters in words.

5. STUDY OF CHARACTERS IN CONTEXT

In this study we aimed firstly at analyzing the mutual influence of adjacent characters in words and secondly at measuring the performance of our identification system for those characters.

5.1 Evaluation experiments

Two different evaluation tests have been done for identification of characters in context: a mono-font experiment and a multi-font experiment. In both studies 3 serif fonts, 3 non-serif fonts with point sizes between 8-12 points and 4 different font styles have been considered.

5.1.1. Mono-font experiment

The objective of this experiment was to measure the character identification rates in a mono-font context; i.e. when the font supposed to be known. By using 6 font types with point sizes between 8-12 and 4 different font styles, we constructed training sets and data sets containing each 300'000 character samples. Table II shows the overall recognition rates according to the font size.

Table II: Recognition Results for 6 font Types

Font Size	12	11	10	9	8
	99.17	98.79	98.73	98.25	97.26

A systematical analysis of misclassification errors has shown two major sources of failure:
1) Misclassification between the letters "i" and "l".
2) Misclassification by having the letter "f" as the precedence letter:

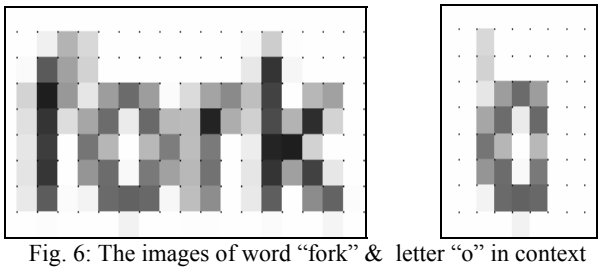

Fig. 6: The images of word "fork" & letter "o" in context

Fig. 6 at the left shows the letter "o" as the precedence letter of "f". One can assume due to this Fig. 6 at the right that "o" could be misclassified as "b". The reason here fore is that the letter "f" has a negative right side bearing and this leads to an unwished noise in the left border of the bounding box of the adjacent character.
By omitting these two sources of errors, we obtain the following overall identification table by font size and font style (each row represents the overall results of 6 font types):

Table III: Recognition Results for 6 font Types

	12	11	10	9	8
plain	99.62	99.84	97.97	96.46	95.25
bold	99.61	99.75	99.28	98.22	98.88
italics	94.83	97.53	97.12	97.25	93.28
bold+italics	98.89	99.00	98.62	98.34	94.83

The results for italic fonts are worse than the roman fonts. The reason is that we used rectangular segmentation windows,, which are not appropriate. Improvement should be obtaine with slanted windows.

5.1.2. Multi-font experiment

In this experiment we have clustered the fonts into two groups: serif and non serif fonts and passed them through either a serif or a non serif identification system.

The results for serif and sans serif fonts by font size and font style are illustrated in tables IV & V:

Table IV: Overall Recognition Results for 2 serif font Types

	12	11	10	9	8
plain	99.60	99.32	98.70	97.64	92.85
bold	99.34	99.34	99.20	99.36	98.61
italics	96.01	96.06	94.88	92.30	92.19
bold+italics	97.71	97.00	94.62	95.37	94.39

Table V: Recognition Results for 3 sans serif font Types

	12	11	10	9	8
plain	97.61	99.26	97.15	92.63	91.85
bold	98.44	97.72	98.47	98.43	96.44
italics	97.25	98.08	95.29	95.97	86.04
bold+italics	98.52	97.86	96.03	93.25	93.93

The results for sans serif fonts are worse than the ones for serif, fonts. The reason here fore is that similar letters like “i” and “l in serif fonts” are better recognizable, since the serif fonts use small decorative marks to embellish characters.

5.2 Conclusion

The study of characters in context shows that the performance of our proposed character identification system is not as accurate as its performance on isolated characters. We expect an improvement by extending character models with contextual information.

6. DISCUSSSION & FUTURE WORK

In this paper we have introduced two studies about very low resolution character identification. The first experiment about identification of isolated characters delivered very accurate results, whereas the second experiment about characters in context delivered results, which were less accurate. Therefore, we plan to design a more reliable character identification system by using Hidden Markov Models (HMMs), which is able to combine segmentation and recognition within the same process. Furthermore, HHMs are also well suited to include higher level linguistic knowledge such as probabilities of character and word sequences, which will allow increasing the future recognition rate.

6. REFERENCES

[1] A. Antonacopoulos, D. Karatzas, J.O.Lopetz “Accessing Textual Information Embedded in Internet Images”, *Proceedings of Electronic Imaging*, Internet Imaging II, San Jose, California, Jan. 2001

[2] E.V. Munson, Y. Tsymbalenko, “Using HTML Metadata to Find Relevant Images on the Web”, *Proceedings of Internet Computing* 2001, Volume II, Las Vegas, pages 842-848, CSREA Press, June 2001

[3] D. Lopresti, J. Zhou, “Document Analysis and the World Wide Web*”*, *International Association for Pattern Recognition*, Workshop on Document Analysis Systems, pp 651-671, 1996

[4] J. Zhou, D. Lopresti, “Extracting Text from WWW Images*”*, *Proceedings of the 4th ICDAR*, pp 248-252, 1997

[5] J. Zhou, D. Lopresti, “OCR for World Wide Web Images*”*, *Proceedings of SPIE on Document Recognition IV*, pp 58-66, 1997

[6] D. Lopresti, J. Zhou, “Locating and Recognizing Text in WWW Images*”*, *Information Retrieval 2,* pp 177-206, , 2000

[7] A. Antonacopoulos, D. Karatzas, “An Anthropocentric Approach to Text Extraction from WWW Images*”*, *IAPR* Rio de Janiero, 2000

[8] A. Antonacopoulos, D. Kartazas, “Text Extraction from Web Images Based on Human Perception and Fuzzy Inference*”*, *Document Analysis Systems V*: 5th International Workshop, DAS 2002, Princeton, NY, USA, August 19-21, 2002

[9] A. Antonacopoulos, D. Karatzas, “Text Extraction from Web Images Based on a Split-and-Merge Segmentation Method Using Color Perception”, *Proceedings of the 17th International Conference on Pattern Recognition* (ICPR2004), Cambridge, UK, August 23-26, 2004

[10] T. Hong, J.J. Hull, “Character Segmentation Using Visual Inter-Word Constraints in Text Page”, *SPIE Conf. on Document Recognition* II, pp. 76-83, San Jose, Feb 6-7, 1995,

[11] http://java.sun.com/j2se/1.4.2/docs/api/java, 18.9.2006

[12] F.Einsele, R.Ingold, "A Study of the Variability of Very Low Resolution Characters and the Feasibility of their Discrimination Using Geometrical Features." **,** *proc. of 4th World Enformatica Congress*, International Conference on Pattern Recognition and Computer Vision, pp. 213-217, Istanbul (Turkey), June 24 - 26 2005