International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:6, 2007

A Study of the Variability of Very Low Resolution Characters and the Feasibility of Their Discrimination Using Geometrical Features

Farshideh Einsele, and Rolf Ingold

***Abstract*—**Current OCR technology does not allow to accurately recognizing small text images, such as those found in web images. Our goal is to investigate new approaches to recognize *very low resolution* text images containing anti-aliased character shapes.

This paper presents a preliminary study on the variability of such characters and the feasibility to discriminate them by using geometrical features. In a first stage we analyze the distribution of these features. In a second stage we present a study on the discriminative power for recognizing isolated characters, using various rendering methods and font properties. Finally we present interesting results of our evaluation tests leading to our conclusion and future focus.

***Keywords*—**World Wide Web, document analysis, pattern recognition, Optical Character Recognition.

I. Introduction

THE World Wide Web has grown up to a vast electronic data collection with billions of documents. Search engine crawlers have been developed for indexing this information. These search engines index the existing mainly plain text and meta information contained in HTML documents. However, the significance of web has persuaded the web designer to adorn the originally text with eye catching and colorful web images. Some of these images contain important textual information and play often a key role for indexing the information [1]. Unfortunately the current search engines are not able to use these embedded texts in web images for indexing [2]. The recommended ALT tag in HTML language, which has been designed for that purpose, is either not used or incorrectly used [3]. Thus, momentarily archiving and indexing important embedded text in web images has been almost ignored. This has a negative impact on searching the existing web information.

The text embedded in web images differs in many ways from the text contained in scanned document images. Firstly, these images have to be of very low resolution (72 dpi = monitor resolution) and most of them have very small point sizes (less than 10 points). Thus, these images are very small and only a few pixels are available for their presentation. A bilevel presentation produces aliasing leading to character shapes with sharp edges, curves and diagonals. In order to overcome this drawback most font rendering engines use a method called anti-aliasing to smooth edges, curves and diagonals by using several color levels hoping to profit from the way our eyes tend to average two adjacent pixels and see one in the middle. Figure 1 demonstrates this fact:

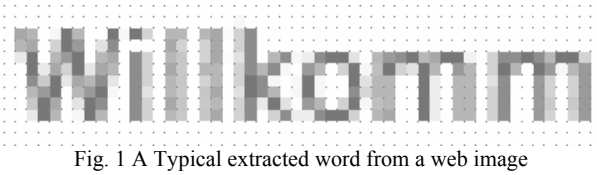

Fig. 1 A Typical extracted word from a web image

Furthermore, despite of the bilevel image of text in printed documents, which is independent of the grid alignment, the anti-aliased image of the same character in web images varies by different shift phases. This causes an additional difficulty for recognizing the text in web images. This fact is shown in figure 2:

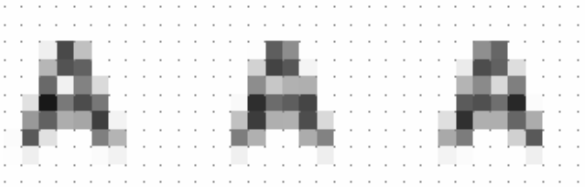

Fig. 2 Variability of character image "A" by different horizontal shift phases

Current existing commercial Optical Character Recognition Software [4] are not designed to accurately recognize the text in figure 1, since they are developed to binarize the images into bilevel shapes and as a consequent the anti-aliased information shown in figure 2 will be lost. However these

Manuscript received Mai 17, 2005 This work was supported by Swiss National Foundation and University of Fribourg, Switzerland.

F. Einsele is with the University of Fribourg, Department of Informatics, Chemin du Musée 3, 1700 Fribourg, (phone: +41 26 300 84 73; fax: +41 26 300 97 31; e-mail: farshideh.einsele@unifr.ch).

R. Ingold is with the University of Fribourg, Department of Informatics, Chemin du Musée 3, 1700 Fribourg, (phone: +41 26 300 97 31; fax: +41 26 300 97 31; e-mail: rolf.ingold@unifr.ch).

commercial OCRs have been developed to recognize the text in printed document images with a resolution of 200 dpi and above and character sizes of 10 point or larger and represented as bilevel images.

Recently, new several studies [5]-[8] have been done by D. Lopresti and J. Zhou concentrating mostly on pre-processing the web images containing textual information and then passing them through existing commercial OCRs. Another approach [9]-[11] has been done by A. Antonacopoulos and D. Karatzas introducing basically two methods for extracting text from web images based on the way humans perceive color differences.

Notwithstanding, the researchers mentioned paid no attention on the anti-aliased rendering nature of the textual information in web images and their variability by different shift phases, which we consider as highly important for developing an effective and reliable tool for our purpose.

In this paper we present a preliminary study, which aims at understanding the behavior of the anti-aliased isolated characters with very low resolution and small point sizes, considering their variability by different shift phases as those found in web images. In our approach we consider the geometrical characteristics of these characters first in a mono-font context. Such a restricted approach is justified when it can be combined with font recognition tools, such as ApOFIS (A priori Optical Font Identification System), a tool that has been developed in our research group [12]. Since the evaluation tests for the described isolated character images has been very encouraging in a mono-font context, the possibility of clustering fonts into serif and non-serif groups has also been evaluated and has still delivered reliable results.

The remainder of this paper is organized as follows: In section 2 we introduce our method to simulate isolated anti-aliased characters based on their variability by different grid alignments such as those found in WWW-images. In section 3 we discuss the features that we use and study their distributions. In section 4 we study the discriminative power of the chosen features using a Bayesian classifier. This paper ends up with conclusions and a short sketch of our future work.

II. Simulation of Low Resolution Character Rendering

The goal of this section was to simulate low resolution anti-aliased rendered characters by different shift phases that we call also grid alignments. For that purpose, we produced images with large font sizes (e.g. 40 and 80 points) and sampled them down by a factor of K (e.g. K=5 or K=10) using a 1/K pixel shift (e.g. 0.2 or 0.1) to simulate different shift phases. By doing this, we gained a set of K*K different shapes for each character. This procedure has been repeated for 52 classes containing upper- and lowercase letters.

"Downsampling" was performed using two different rendering tools: 1) the library provided by Adobe Photoshop and 2) the method used by the Java library as shown in figure 3.

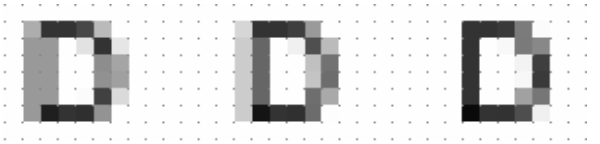

Fig. 3 (a) "Downsampling" with Java

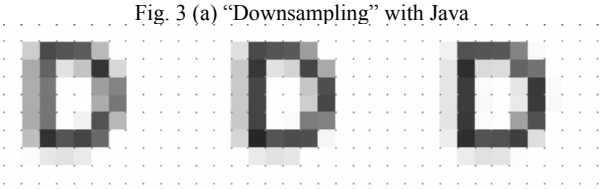

Fig. 3 (b) "Downsampling" with Photoshop

III. Proposed Features

In this task we followed two major goals: first we worked on choosing discriminative and reliable features and second we studied their distribution.

A. Definition of Central Moments

The aim of this part was choosing powerful features to design our classifiers. Thus we decided to use geometrical central moments μ_{00}, μ_{11}, μ_{20}, μ_{02}, μ_{21}, μ_{12} *and* μ_{22} due to their translation invariant nature. Central moments play an important role in the analysis of shapes for the generation of useful features in image analysis and 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)

B. Feature Distribution

The objective of this part was to analyze the statistical distribution of the chosen features. A systematically statistical evaluation was performed on features of 52 classes (uppercase- and lowercase characters) for one font: Arial roman, 8 point. Following the procedure described in section 2 we produced 2x100 different patterns for each class using both rendering methods. We extracted the features introduced in section 3.1 and decided to analyze if they were normally distributed.

Under the null hypothesis that these features follow a normal distribution, we performed the $\chi 2$-test with a degree of freedom of 6 and with a significance level of P = 95% on each individual feature. The features' normality was mostly not rejected (accepted) for patterns simulated with Photoshop rendering method (85% of all cases), whereas for the Java rendering method the proportion for the acceptance of the hypothesis dropped to (35%). A possible explanation of this result is that the Java patterns have a very small variability. Thus by comparing the same features for each pattern, just a narrow gap among them is evident. This could be the reason

for this high rejection rate for a normal distribution by features gained from Java patterns.

IV. Classification Experiment Based on Bayesian Classifier

This section aims at demonstrating the power of the chosen geometrical features to discriminate isolated, anti-aliased characters at very low resolution, with various shift phases as described in the previous chapters. The evaluation was divided into the following experiments:

- General study
- Influence of the rendering method
- Multi-font experiment

A. General Study

The objective of this experiment was gaining global recognition rates by a systematic evaluation using different font families, font styles and font sizes.

As we could verify the "quasi-normality" of the geometrical features obtained from the simulated patterns (at least when using the Photoshop rendering method), we decided to evaluate the discriminative power of these features by the means of a *Bayesian classifier*, using parameter estimation and assuming multivariate normal density functions.

For this purpose we produced 25 different patterns by "downsampling" them with a factor of k=5 (K=5, number of patterns=5*5=25) for each of the 52 classes (upper and lower case letters) under the conditions described in section III.

This evaluation has been done for:

- 7 font sizes: 3-9 points
- 7 font families: Arial, Times New Roman, Courier New, Comic sans MS, Verdana, Tahoma, Century
- 4 font styles: roman, italics, bold, bold & italics

The density function parameters were estimated using "downsampled" Photoshop patterns.

Table 1 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
DIFFERENT POINT SIZES

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

Table 2 shows the recognition rates obtained for different font styles obtained with font sizes in the range of 4-9 points. We can observe that the overall recognition rates of italic fonts are much better than those obtained for roman fonts; identically, the recognition rates of bold fonts were better than for normal font.

TABLE II
DIFFERENT FONT STYLES

	Roman	Italic	Mean
Normal	99.51	99.71	99.61
Bold	99.72	99.93	99.83
Mean	99.62	99.82	

B. Influence of the Rendering Method

In this experiment, we aimed at determining the impact of the chosen rendering method. This evaluation has been performed under the same conditions as 4.1 with font sizes between 4 and 9 points. Four experiments have been made as described below:

a. First, we evaluated our classifiers, which were trained with Photoshop patterns, to recognize Java Patterns. The achieved global recognition rate has dropped to 89.51%.

b. In a second step, we used the Java patterns for training, assuming again normally distributed features, despite the fact that the experiment in section 3.1 showed these features were not normally distributed. Surprisingly, we obtained a global recognition rate of 100% (perfect recognition) of all Java patterns. We interpret the reason of this astonishing result as a consequence of the fact that, due to hinting rules the variability of Java patterns is much lower than that of Photoshop patterns.

c. We used then the same training sets (Java patterns) for recognizing the Photoshop patterns. This experiment did confirm the interpretation above since the achieved global recognition rate was reduced to 37.41%.

d. Finally, we mixed both sets of patterns (Photoshop and Java patterns) and used them as training sets. We observed excellent recognition rates for both test sets: 99.95% for Java patterns and 99.93% for Photoshop patterns.

The findings of the above experiments has lead to the conclusion that the classifier can be trained to make it capable of recognizing characters independently to the rendering method

C. Multi-Font Experiment

Due to the astonishing results of the previous experiments, especially for small font sizes, we decided to evaluate the robustness of the method in a multi-font context using a *Bayesian classifier*.

This evaluation has been done for:

- 3 serif font families: Century, Georgia, Garamond
- 3 non-serif font families: Arial, Verdana, Tahoma
- 2 point sizes : 8 and 7 points
- 4 font styles : roman, italics, bold, bold + italics
- the training set driven from section 4.2.d

Our experiments were organized as follows:

a. Firstly we mixed the patterns of all 6 fonts (serif, non-serif) obtained from both rendering methods (Photoshop and Java) and then randomly generated 1/3 of them as training sets and 2/3 of them for recognition. The purpose of this step was generating different sets for training and test, though for a Bayesian decision using the same patterns for training and recognition has a negligible influence. The achieved recognition rates are shown in table 3.

b. Secondly as the results of previous step as shown in table 3 were reduced comparing to those of previous chapters, we repeated the same experiment with a small modification in the number of patterns used for test sets, in which we separated the serif and non-serif fonts and obtained better recognition rates than former experiment. Table 3 demonstrates the achieved results for both experiments a and b:

TABLE III
MULTI-FONT RESULTS

	Roman	Bold	Italic	Bold+Italic
6 fonts	91.85	91.94	88.61	90.05
Sans-serif	98.23	98.10	98.77	98.91
Serif	98.87	99.37	97.25	98.43

V. CONCLUSION

In this paper we have presented a preliminary study about the variability and discrimination capabilities of very low resolution (72 dpi = monitor resolution), anti-aliased, isolated rendered characters at small point sizes (less than 10 points) as those found in WWW-images. We have presented a method to simulate these characters considering the fact that for the same character the resulted anti-aliased image was different due to its grid-alignment. By using geometrical central moments as features we evaluated their ability to be recognized by a Bayesian classifier**.**

It has been shown that geometrical features based on central moments are able to discriminate characters very accurately. In addition to this it could be shown that character discrimination performs very well for font sizes down to 5 points. Further it has been shown that the classifier can be trained to avoid the dependency of the applied rendering method. Finally the possibility of clustering the serif and sans-serif fonts as classifier still delivered reliable recognition rates.

The presented study shows that if the characters are isolated, they can be recognized very accurately. We are currently studying the influence of adjacent characters in words. The next step will consist in a novel segmentation algorithm, based on hypothesis checking, which will use the results of this study. Our final goal is to come up with a novel OCR approach for very low resolution character recognition, which combines segmentation and character classification.

REFERENCES

[1] A. Antonacopoulos, D. Karatzas nad J.O.Lopetz "Accessing Textual Information Embedded in Internet Images", Proceedings of Electronic Imaging, Jan. 2001, Internet Imaging II, San Jose, California.
[2] D. Amor, The E-Business (R)evolution, Prentice Hall, 1999.
[3] 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.
[4] G. Nagy, "Twenty Years of Document Image Analysis in PAMI", IEEE Transactions on Pattern Analysis and Machine Intelligence", 1999125.
[5] D. Lopresti, J. Zhou, "*Document Analysis and the World Wide Web"*, International Association for Pattern Recognition, Workshop on Document Analysis Systems, 1996, pp 651-671.
[6] J. Zhou, D. Lopresti, "*Extracting Text from WWW Images"*, Proceedings of the 4th ICDAR, 1997, pp 248-252.
[7] J. Zhou, D. Lopresti, "*OCR for World Wide Web Images"*, Proceedings of SPIE on Document Recognition IV, 1997, pp 58-66.
[8] D. Lopresti, J. Zhou, "*Locating and Recognizing Text in WWW Images"*, Information Retrieval 2,, 2000, pp 177-206.
[9] A. Antonacopoulos, D. Karatzas, "*An Anthropocentric Approach to Text Extraction from WWW Images"*, IAPR Rio de Janiero, 2000.
[10] 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.
[11] 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, IEEE-CS Press.
[12] A. Zramdini and R. Ingold, "*Optical Font Recognition Using Typographical Features"*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), August 1998.