

What Fraction of Images on the Web Contain Text?

Tapas Kanungo
IBM Almaden Research Center
650 Harry Road, 52BC
San Jose, CA 95120, USA
kanungo@almaden.ibm.com

Chang Ha Lee
Department of Computer Science
University of Maryland
College Park, MD 20740, USA
chlee@cs.umd.edu

Roger Bradford
Science Applications International Corp.
1953 Gallows, Road Vienna, VA 22182, USA
bradfordr@saic.com

Abstract

Web search engines index text represented in symbolic form. However, it is well known that a fraction of the text on the web is present in the form of images, and the textual content of these images is not indexed by the search engines. This fact immediately raises a few questions: i) What fraction of the images on the web contain text? ii) What fraction of the text content of these images does not appear in the web page in symbolic form? Answers to these questions will give the web users an idea about the amount of information being missed by the search engines, and, justify whether or not Optical Character Recognition should be a standard part of search engine indexing. To answer these questions we statistically sample the images referenced in the web pages retrieved by a search engine for specific queries and then find the fraction of sampled images that contain text.

1. Introduction

Researchers [8, 11] have reported that “considerable portion of the text is on the World Wide Web (WWW) is embedded in images.” However, we are not aware of quantitative estimates of “considerable portion” in the literature. Furthermore, it is not clear whether the text in images is completely independent of the text in the corresponding HTML file. That is, if a high proportion of the text in images also appears in the corresponding HTML files, not much information is lost by the internet search engines.

In this paper we try to estimate the fraction of images on the web that contain text, and also the percentage of image text that does not appear in the corresponding HTML file.

Lopresti and Zhou [8, 11, 9] have proposed an approach for extracting text strings from complex images on the web that first quantizes the color and then detects connected components. Wu, Manmatha and Riseman [10] propose a text detection and extraction algorithm that is based on analyzing the image texture. Lienhart and Stuber [7] and Li, Kia and Doermann [6] present algorithms for detecting text in video image sequences. However, none of above papers have addressed the issue of the “size” of the problem. Lawrence and Giles [5] have estimated the number of web pages on the internet. However this work does not address the issue of images on the web. Antonacopoulos and Karatzas [1, 2] recently proposed an anthropocentric approach to extract textual information from web images, and studied 200 web sites to investigate the proportion of the significant text in image form. However, they did not provide the details of the sampling strategy used in their experiment, and it is not clear if they considered things like stopwords which are not significant as keyword.

2. The Problem

Let H be a web page, and $W(H)$ be the set of words represented in it. Let $I(H)$ be the set of images referenced by the web page H and let $W(I(H))$ and $W(T(H))$ be the set of words in the image and text part of the web page, respectively. Thus $W(H) = W(T(H)) \cup W(I(H))$. Search engines index the words $T(H)$. Now let's raise the questions:

1. On average, what fraction of $I(H)$ contains text?
2. On average, what fraction of words in the images, $W(I(H))$, do not appear in the symbolic part $W(T(H))$. That is, what is the expected value of $\#(W(I(H)) \cap W(T(H))^c) / \#W(I(H))$.

3. Methodology

The presense or absence of text in a web image can be determined by three methods. First, is to use a commercial OCR system and test whether the output of the OCR system is non-empty. The second method is to use an algorithm that classifies an image as text or non-text based on color or texture features. The third approach is to ask a human to view each image and record whether or not it contains text.

Each approach has its problems. The problem of detecting and recognizing text in web images via OCR is quite difficult. The images on the web typically happen to be textured color images with various types of stylized fonts, overlays, occlusion etc. Most commercial OCR products like OmniPage and TextBridge, however, are trained on bitonal document images that have very ‘regular’ text — like text in simple memos, journal and magazine articles, etc. Thus commercial products typically can not detect the existence of text in the web images, let alone recognize it. Thus if we use OCR as a means of detecting text in web images, we will underestimate the fraction of web images containing text.

Feature-based classification of web images into text and non-text also has misdetection and false alarm classification errors. This can make our estimate biased high or low depending upon whether the false alarms are more or less than the misdetections.

Finally, manual detection of text is time consuming and laborious. However, manual detection can be more accurate than either of the other two methods.

In our approach we select a small representative sample of images by randomly selecting images from the collection of web pages returned for a query. We then manually transcribe the text within each image. We then compute the fraction of images that contain text in the sample. This fraction is used as an estimate of the probability of finding text in a web image.

Next each word in the image text is searched in the corresponding HTML file. We then count the number of image words that do not appear in the corresponding HTML file and compute the fraction of image words that do not appear in the corresponding HTML files.

4. Experimental Protocol

We used the Google internet search engine for our experiments. The batch search jobs were run using Perl 5.6 scripts that invoked `get` function to conduct the search. The individual web pages referenced in the search result were retrieved using the GNU `wget` version 1.5.3 package. The HTML files were parsed using the `HTML::Parser` in the Perl package.

In our current experiments we used one query – “news-papers” – and requested a maximum of one thousand search

results. Google returned 934 web pages, of which 72 either did not exist or were having network problems.

The 862 functional web pages were retrieved and each reference to an image within each HTML file was recorded. There were a total of 18161 images referenced in the 862 HTML files. We randomly selected 300 images from the 18161 images. We were able to download only 265 of the 300 selected images. Each of these 265 images were viewed using an image viewer and the existence of text was recorded and the text string in the image was entered into a corresponding text file. The fraction of images that contained textual information was recorded.

Next, each word in the human-entered text file was searched in the corresponding HTML file. Care was taken to omit the comments sections while searching. In addition, we used a stopword list with 320 words to exclude stopwords. The fraction of words in image files not found in the HTML file was computed.

Query selection is an issue. The queries should be able to retrieve representative web pages. Our proposal is to use the categories of Yahoo as queries. The query used in our experiments – “newspapers” – is a Yahoo category. There are 378 such categories. Manually typing text corresponding to 300 images in each of these category searches can be time consuming, which took about 3 hours for each category in our case. One way is to select a small random sample from the 378 categories and then manually specify whether or not there is text in each image. The results reported in this article are for one query only. Experiments are currently being run for 20 queries/categories.

5. Results and Discussion

Each HTML file contains references to various images. In Figure 5 we show the distribution of the number of images contained in the HTML files. We see that the distribution looks exponential with the highest number of HTML files not referring to any image at all and some of the HTML files referring to more than 200 images. In Figure 1(a) we show examples of images that contain text which do not appear in the corresponding HTML page. In Figure 1(b) we show examples of images that contain text, all of which appear in the corresponding HTML file. In Figure 2 we show examples of images that do not contain any text.

Our experimental findings reveal the following:

1. 42% of the images in the sample contain text.
2. Of the images with text, 59% of the images contain at least one word that does not appear in the corresponding HTML file.
3. Of the images with text, 36% of the images are such that $W(I(H)) \subset W(T(H))$. That is, for 36% of the

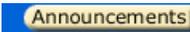
 Club Freep	
Home Delivery FAQ	
	
Volume 3 No. 10	
	
(a) Images containing text that do not appear in the HTML file.	(b) Images containing text that appear in the HTML file.

Figure 1. Sample images with text.

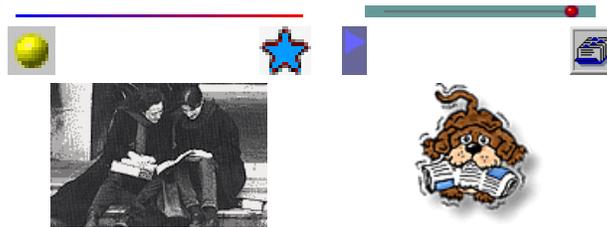
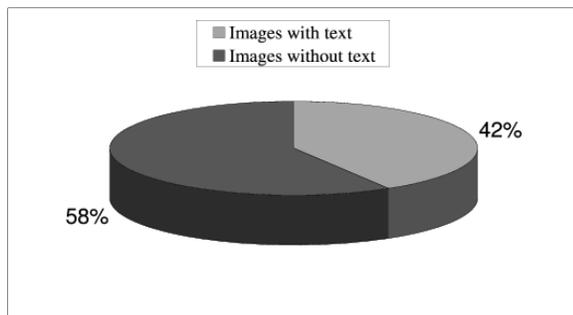
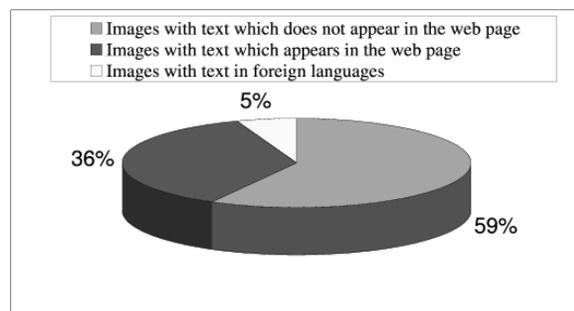


Figure 2. Sample images without text.



(a) The proportion of text images in randomly selected 265 images.



(b) The proportion of images containing text.

Figure 3. The proportion of images.

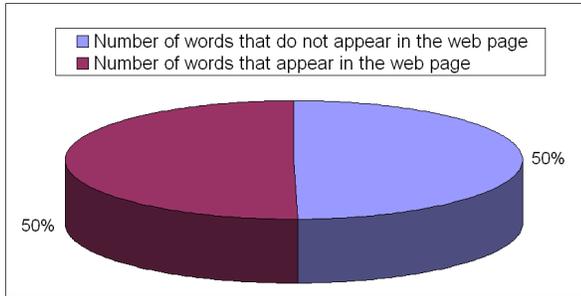


Figure 4. The proportion of words appearing in the text images.

images with text, all the words in the image are contained in the corresponding HTML file. Thus the text in images are not completely independent of the text in the HTML files.

4. 50% of all the non-stopwords in text images are not contained in the corresponding HTML file. Before excluding stopwords, 42% of all the words in the images are not contained in the corresponding HTML file. 78% of all the words in text images are non-stop words, and 93% of the words which are not contained in the corresponding HTML file are non-stopwords.
5. 5% of the images with text contain non-English script.

Thus of the $N = 18161$ images the expected number of images that contain text is $Np = 7627$ and the standard deviation is $\sqrt{Np(1-p)} = 66$ where p is the estimated fraction. The large sample 95% confidence interval [4, 3] for the p is $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})} = 0.42 \pm 0.059$.

Our quantitative estimates concur with the subjective hypothesis of researchers that the amount of text images is increasing on the web.

Preliminary experiments with other queries indicate similar text image proportions. However, in some subcategories, such as Arabic newspapers, we have noticed that many newspapers tend to have images of text instead symbolic text. Korean newspapers tend to be symbolic, however.

Experiments in this article raise new questions. Is the image content of web pages increasing with time? Is the fraction of images containing text increasing every year? Can we get better estimates of the proportions using image-feature based text detection algorithms?

Distribution of Web Pages Containing Images

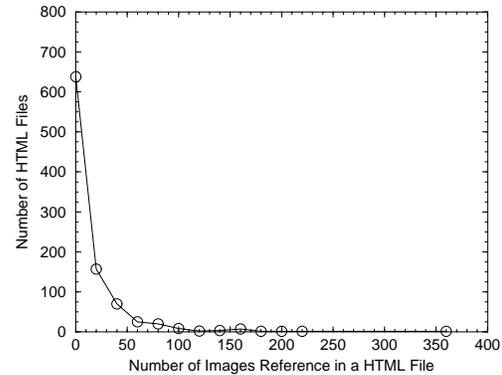


Figure 5. The distribution of web pages containing images.

References

- [1] A. Antonacopoulos and D. Karatzas. An anthropocentric approach to text extraction from WWW images. In *Proceedings of IAPR Workshop on Document Analysis Systems*, pages 515–525, Rio de Janeiro, Brazil, December 2000.
- [2] A. Antonacopoulos, D. Karatzas, and J. O. Lopez. Accessing textual information embedded in internet images. In *Proceedings of SPIE Conference on Internet Imaging*, pages 198–205, San Jose, CA, January 2001.
- [3] G. Casella and R. L. Berger. *Statistical Inference*. Wadsworth and Brooks/Cole, Pacific Grove, California, 1990.
- [4] J. L. Devore. *Probability and Statistics*. Brooks/Cole Publishing Company, Pacific Grove, CA, 1995.
- [5] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280:98–100, 1998.
- [6] H. Li, O. Kia, and D. Doermann. Text enhancement in digital video. In *Proceedings of SPIE Conference on Document Recognition IV*, pages 1–8, 1999.
- [7] R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In *Proceedings of ACM Multimedia*, pages 11–20, 1996.
- [8] D. Lopresti and J. Zhou. Document analysis and the World Wide Web. In *Proceedings of IAPR Workshop on Document Analysis Systems*, Marven, PA, 1996.
- [9] D. Lopresti and J. Zhou. Locating and recognizing text in WWW images. *Information Retrieval*, 2:177–206, 2000.
- [10] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, 1997.
- [11] J. Zhou and D. Lopresti. Extracting text from WWW images. In *Proceedings of the IAPR International Conference on Document Analysis Recognition*, Ulm, Germany, 1997.