Classifying Objectionable Websites Based on Image Content *

James Ze Wang Jia Li Gio Wiederhold Oscar Firschein

Stanford University, Stanford, CA 94305, USA

Abstract. This paper describes IBCOW (Image-based Classification of Objectionable Websites), a system capable of classifying a website as objectionable or benign based on image content. The system uses WIPE$_{TM}$ (Wavelet Image Pornography Elimination) and statistics to provide robust classification of on-line objectionable World Wide Web sites. Semantically-meaningful feature vector matching is carried out so that comparisons between a given on-line image and images marked as "objectionable" and "benign" in a training set can be performed efficiently and effectively in the WIPE module. If more than a certain number of images sampled from a site is found to be objectionable, then the site is considered to be objectionable. The statistical analysis for determining the size of the image sample and the threshold number of objectionable images is given in this paper. The system is practical for real-world applications, classifying a Web site at a speed of less than 2 minutes each, including the time to compute the feature vector for the images downloaded from the site, on a Pentium Pro PC. Besides its exceptional speed, it has demonstrated 97% sensitivity and 97% specificity in classifying a Web site based solely on images. Both the sensitivity and the specificity in real-world applications is expected to be higher because our performance evaluation is relatively conservative and surrounding text can be used to assist the classification process.

1 Introduction

With the rapid expansion of the Internet, every day large numbers of adults and children use the Internet for searching and browsing through different multimedia documents and databases. Convenience in accessing a wide range of information is making the Internet and the World-Wide Web part of the everyday life of ordinary people. Because there is freedom of speech, people are allowed to publish various types of material or conduct different types of business on the Internet. However, due to this policy, there is currently a large amount of domestic and foreign objectionable images and video sequences available for free download on the World-Wide Web and usenet newsgroups. Access of objectionable graphic images by under-aged "netters" is a problem that many parents are becoming concerned about.

* We thank the Stanford University Libraries and Academic Information Resources for providing computer equipment during the development and testing process.
Correspondence to: wangz@cs.stanford.edu

1.1 Related Work in Industry

There are many attempts to solve the problem of objectionable images in the software industry. Pornography-free web sites such as the *Yahoo! Web Guides for Kids* have been set up to protect those children too young to know how to use the web browser to get to other sites. However, it is difficult to control access to other Internet sites.

Software programs such as *NetNanny*, *Cyber Patrol*, or *CyberSitter* are available for parents to prevent their children from accessing objectionable documents. However, the algorithms used in this software do not check the image contents. Some software stores more than 10,000 IP addresses and blocks access to objectionable sites by matching the site addresses, some focus on blocking websites based on text, and some software blocks all unsupervised image access. There are problems with all of the approaches. The Internet is so dynamic that more and more new sites and pages are added to it every day. Maintaining lists of sites manually is not sufficiently responsive. Textual matching has problems as well. Sites that most of us would find benign, such as the sites about breast cancer, are blocked by text-based algorithms, while many objectionable sites with text incorporated in elaborate images are not blocked. Eliminating all images is not a solution since the Internet will not be useful to children if we do not allow them to view images.

1.2 Related Work in Academia

Academic researchers are actively investigating alternative algorithms to screen and block objectionable media. Many recent developments in shape detection, object representation and recognition, people recognition, face recognition, and content-based image and video database retrieval are being considered by researchers for use in this problem.

To make such algorithms practical for our purposes, extremely high sensitivity (or recall of objectionable websites) with reasonably high speed and high specificity is necessary. In this application, *sensitivity* is defined as the ratio of the number of objectionable websites identified to the total number of objectionable websites accessed; *specificity* is defined as the ratio of the number of benign websites passed to the total number of benign websites accessed. A perfect system would identify all objectionable websites and not mislabel any benign websites, and would therefore have a sensitivity and specificity of 1. The "gold standard" definition of objectionable and benign images or websites is a complicated social problem and there is no objective answer. In our experiments, we use human judgment to serve as a gold standard.

For real-world application needs, a high sensitivity is desirable, i.e., the correct identification of almost every objectionable website even though this may result in some benign websites being mislabeled. Parents might be upset if their children are exposed to even a few objectionable websites.

The following properties of objectionable images found on the Internet make the problem extremely difficult:

– mostly contain non-uniform image background;
– foreground may contain textual noise such as phone numbers, URLs, etc;
– content may range from grey-scale to 24-bit color;
– some images may be of very low quality (sharpness);
– views are taken from a variety of camera positions;
– may be an indexing image containing many small icons;
– may contain more than one person;
– persons in the picture may have different skin colors;
– may contain both people and animals;
– may contain only some parts of a person;
– persons in the picture may be partially dressed.

Forsyth's research group [2, 3] has designed and implemented an algorithm to screen images of naked people. Their algorithms involve a skin filter and a human figure grouper. As indicated in [2], 52.2% sensitivity and 96.6% specificity have been obtained for a test set of 138 images with naked people and 1401 assorted benign images. However, it takes about 6 minutes on a workstation for the figure grouper in their algorithm to process a suspect image passed by the skin filter.

1.3 Overview of Our Work

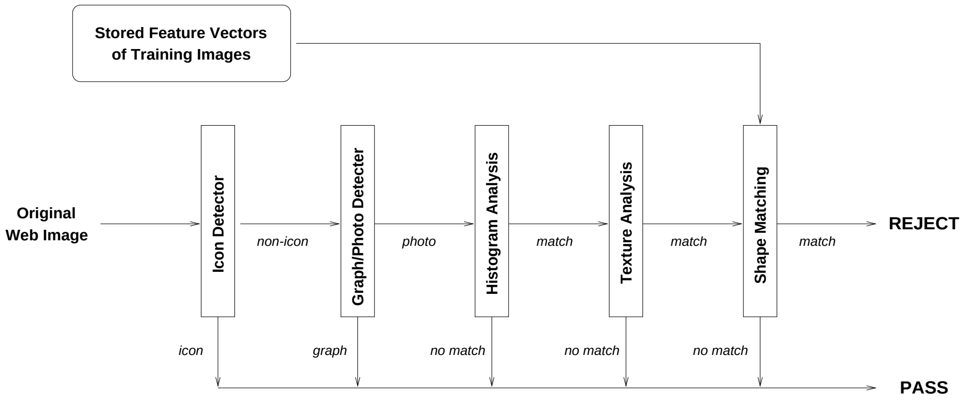

Fig. 1. Basic structure of the algorithm in WIPE.

WIPE Our group has built the WIPE [9, 11] system that is capable of classifying an image as objectionable or benign in a much efficient way, processing an image within 2 seconds on a Pentium Pro PC. Instead of carrying out a detailed analysis of an image, we match it against a small number of feature vectors obtained from a training database of 500 objectionable images and 8,000 benign images, after passing the images through a series of fast filters. If the image is close in content to a threshold number of pornographic images, e.g., matching two or more of the marked objectionable images in the training database within the closest 15 matches, it is considered objectionable. To accomplish this, we attempt to effectively code images based on image content and match the query with statistical information on the feature indexes of the training database. The foundation of this approach is the content-based feature vector indexing and matching developed in our multimedia database research. Image feature vector indexing has been developed and implemented in several multimedia database systems such as the IBM QBIC System [1] developed at the IBM Almaden Research Center. Readers are referred to [5–8, 10] for details on this subject. However, for WIPE we use quite specialized features. Using Daubechies' wavelets, moment analysis, texture analysis, histogram analysis and statistics, the algorithm in the WIPE system is able to produce a 96% sensitivity and a higher than 91% specificity, tested on more than 1,000 objectionable photograph images and more than 10,000 benign photograph images.

IBCOW Based on the WIPE system, we have developed IBCOW, a system that can classify a website as objectionable or benign. The system downloads and classifies the first N images from a new website. If at least a subset of the N images are classified as objectionable by the existing WIPE system, the website is classified as objectionable by the IBCOW system; otherwise, the website is classified as a benign website.

In the real world, IBCOW can be incorporated in a World Wide Web client software program so that a website is first screened by the system before the client starts to download the contents of the website. Once the website is screened, it is memorized on the local storage so that it is considered safe for some period of time. IBCOW can also be used as a tool for screening software companies to generate lists of potentially objectionable World Wide Web sites.

2 Screening Algorithms in IBCOW and Statistical Analysis

In this section, we derive the optimal algorithm for classifying a World Wide Web site as objectionable or benign based on an image classification system like the WIPE system developed by us.

2.1 Screening Algorithms

As discussed in [9, 11], the screening algorithm in the WIPE module uses several major steps, as shown in Figure 1. The layout of these filters is a result of a cost-effectiveness analysis. Faster filters are placed earlier in the pipeline so that benign images can be quickly passed. The algorithm in WIPE has resulted a 96% sensitivity and higher than 91% specificity for classifying a photographic image as objectionable or benign. Figure 2 and 3 show typical images being mistakenly marked by the WIPE system.

(a) (b) (c) (d) (e)

Fig. 2. Typical benign images being marked mistakenly as objectionable images by WIPE. (a) areas similar to human body (b) fine-art (c) some dog images are difficult to classify (d) partially undressed (e) hard to tell (bathing) .

Figure 4 shows the basic structure of the IBCOW system. For a given suspect website, the IBCOW system first downloads as many pages as possible from the website by following the links from the front page. The process is terminated after a pre-set timeout period. Once the pages are downloaded, we use a parser to extract image URLs, i.e., URLs with suffixes such as '.jpg' or '.gif'. N randomly selected images from this list are downloaded from this website. Then we apply WIPE to classify the images. If at least a subset, say $r \times N$ images ($r < 1$), of the N images are classified as objectionable by the WIPE system, the website is classified as objectionable by the IBCOW system; otherwise, the website is classified as a benign website. The following subsection will address the ideal combination of N and r given the performance of WIPE using statistical analysis.

(a) (b) (c) (d) (e)

Fig. 3. Typical objectionable images being marked mistakenly as benign images by WIPE. (a) undressed part too small (b) frame (c) hard to tell (d) image too dark and of extremely low contrast (e) dressed but objectionable . Some areas of objectionable images are blackened and blurred.

2.2 Statistical Analysis

In order to proceed with the statistical analysis, we must make some basic assumptions. A World Wide Web site is considered a benign website if none of the images provided by this website is objectionable; otherwise, it is considered an objectionable website. This definition can be refined if we allow a benign website to have a small amount, say, less than 2%, of objectionable images. In our experience, some websites that most of us would find benign, such as some university websites, may still contain some personal homepages with small number of half-naked movie stars' images.

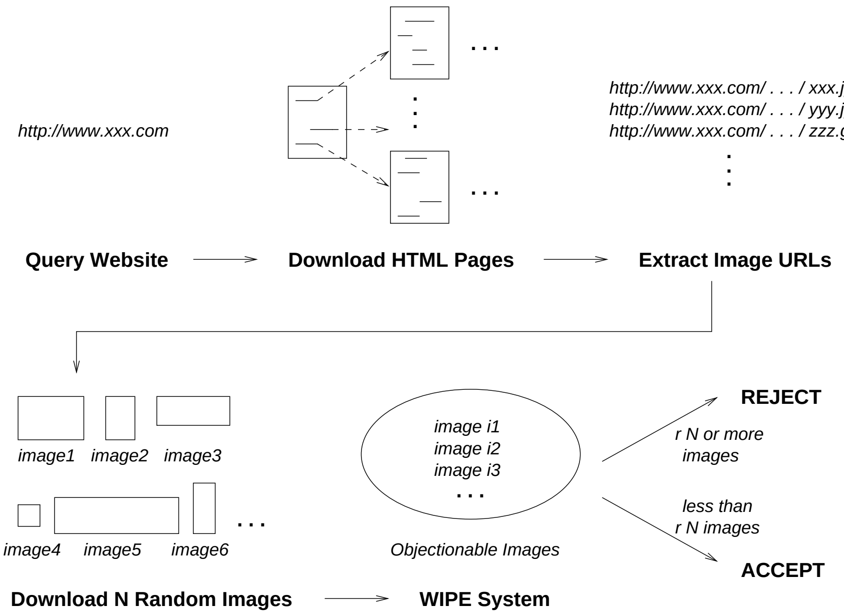

Fig. 4. Basic structure of the algorithm in IBCOW.

For a given objectionable website, we denote p as the chance of an image on the website to be an objectionable image. The probability p varies between 0.02 and 1 for various objectionable websites. Given a website, this probability equals to the percentage of objectionable images over all images provided by the website. The distribution of p over all websites in the world physically exists, although we would not be able to know what the distribution is. Therefore, we assume that p obeys some hypothetical distributions, which are as shown in Figure 5.

The performance of our WIPE system was evaluated by two parameters: sensitivity, denoted as q_1, is the accuracy of detecting an objectionable image as objectionable, and specificity, denoted as q_2, the accuracy of detecting a benign image as benign. The false positive rate, i.e., the failure rate of blocking an objectionable image, is thus $1 - q_1$, and the false negative rate, i.e., the false alarm rate for benign images, is $1 - q_2$.

For IBCOW, we must find out the minimum number of images, denoted as N, from a suspect website to be tested by the WIPE module in order to classify a website as objectionable or benign at a confidence level α, i.e., with a probability $1-\alpha$ of being correct. The confidence level requirements on objectionable websites and benign websites may differ. For objectionable websites, we denote the desired confidence level to be α_1, while for benign websites, we denote the desired confidence level to be α_2. Furthermore, we must decide the threshold, denoted as r, for the percentage of detected objectionable images at which the IBCOW system will classify the website as objectionable. Therefore, the system tests N images from a website and classifies the website as objectionable if more than $r \times N$ images are detected as objectionable by WIPE. Our objective is that when a website is rated as objectionable with probability higher than $1-\alpha_1$, it will be classified as objectionable, and when a website is rated benign, with probability higher than $1-\alpha_2$, it will be classified as benign.

According to the above assumptions, we can calculate the probabilities of misclassifying objectionable websites and benign websites. We start with the simpler case of benign websites.

$$P\{\ classified\ as\ benign \mid a\ website\ is\ benign\ \} = P(I_2 \leq rN) \quad ,$$

where I_2 is the number of images detected as objectionable by WIPE. Since I_2 is a binomial variable [4] with probability mass function

$$p_i = \binom{n}{i}(1-q_2)^i q_2^{n-i}\ ,\ i = 0,\ 1,\ \ldots,\ n \quad ,$$

we have

$$P(I_2 \leq rN) = \sum_{i=1}^{[rN]} \binom{N}{i}(1-q_2)^i q_2^{n-i} \quad .$$

Similarly, for objectionable websites, we get

$$P\{\ classified\ as\ objectionable \mid a\ website\ is\ objectionable\ \} = P(I_1 > rN) \quad .$$

For an objectionable website, suppose that any image in this website has probability p of being objectionable and it is independent of the other images, then the probability for this image to be classified as objectionable image is evaluated as follows:

$$\begin{aligned}
& P\{\ classified\ as\ objectionable\ \} \\
= & P(A)P(\ classified\ as\ objectionable \mid A) + \\
& P(\tilde{A})P(\ classified\ as\ objectionable \mid \tilde{A}) \\
= & pq_1 + (1-p)(1-q_2)
\end{aligned}$$

where

$$\begin{aligned}
A &= \{\ the\ image\ is\ objectionable\ \}, \\
\tilde{A} &= \{\ the\ image\ is\ benign\ \} \quad .
\end{aligned}$$

For simplicity, we denote

$$\lambda(p) = pq_1 + (1-p)(1-q_2) \quad .$$

Similarly, I_1 follows a binomial distribution with a probability mass function

$$p_i = \binom{n}{i} (\lambda(p))^i (1-\lambda(p))^{n-i} \ , \ i = 0,\ 1,\ \ldots\ ,\ n \quad .$$

For this specific website,

$$P(I_2 > rN) = \sum_{[rN]+1}^{N} \binom{N}{i} (\lambda(p))^i (1-\lambda(p))^{n-i} \quad .$$

If p follows a truncated Gaussian distribution, i.e., the first hypothetical distribution, we denote the probability density function of p as $f(p)$. Thus,

$$\begin{aligned} &P\{\ classified\ as\ objectionable \mid a\ website\ is\ objectionable\ \} \\ &= \int_0^1 \left[\sum_{[rN]+1}^{N} \binom{N}{i} (\lambda(p))^i (1-\lambda(p))^{n-i} \right] f(p)dp \quad . \end{aligned}$$

As N is usually large, the binomial distribution can be approximated by a Gaussian distribution [4]. We thus get the following approximations.

$$\begin{aligned} &P\{\ classified\ as\ benign \mid a\ website\ is\ benign\ \} \\ &= \sum_{i=1}^{[rN]} \binom{N}{i} (1-q_2)^i q_2^{n-i} \\ &\approx \Phi\left(\frac{(r-(1-q_2))\sqrt{N}}{\sqrt{q_2(1-q_2)}} \right) \quad , \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of normal distribution [4]. Supposing $r > (1-q_2)$, the above formula converges to 1 when $N \to \infty$.

$$\begin{aligned} &P\{\ classified\ as\ objectionable \mid a\ website\ is\ objectionable\ \} \\ &\approx \int_0^1 \left(1 - \Phi\left(\frac{(r-\lambda(p))\sqrt{N}}{\sqrt{\lambda(p)(1-\lambda(p))}} \right) \right) f(p)dp \quad , \end{aligned}$$

where $\lambda(p) = pq_1 + (1-p)(1-q_2)$ as defined before.
When $r < \lambda(p)$,

$$\lim_{N\to\infty} \Phi\left(\frac{(r-\lambda(p))\sqrt{N}}{\sqrt{\lambda(p)(1-\lambda(p))}} \right) \to 0 \quad .$$

Obviously, for any reasonable objectionable image screening system, $q_1 > 1 - q_2$, i.e., the truth positive (TP) rate is higher than the false positive (FP) rate. Hence, we can choose r so that $r \in (1-q_2,\ \epsilon q_1 + (1-\epsilon)(1-q_2))$ for $\epsilon > 0$. The inequality $r > 1 - q_2$ will guarantee that the probability of misclassifying benign websites approaches zero when N becomes large, which we concluded in a previous analysis. On the other hand, the inequality $r < \epsilon q_1 + (1-\epsilon)(1-q_2)$ will enable the probability of misclassifying objectionable websites to become arbitrarily small when N becomes large.

To simplify notation, we let

$$\Delta_{r,N}(p) = 1 - \Phi\left(\frac{(r - \lambda(p))\sqrt{N}}{\sqrt{\lambda(p)(1-\lambda(p))}}\right) \quad .$$

Note that

$$\int_0^1 \Delta_{r,N}(p) f(p) dp \geq \int_\epsilon^1 \Delta_{r,N}(p) f(p) dp \quad .$$

By increasing N, we can choose arbitrarily small ϵ so that $\Delta_{r,N}(p)$ is as close to 1 as we need, for all $p > \epsilon$. Hence, $\int_\epsilon^1 \Delta_{r,N}(p) f(p) dp$ can be arbitrarily close to $\int_\epsilon^1 f(p) dp$. Since we can choose ϵ arbitrarily small, this integration approaches to 1. In conclusion, by choosing r slightly higher than $1-q_2$ and N large, our system can perform close to 100% correctness for classification of both objectionable websites and benign websites.

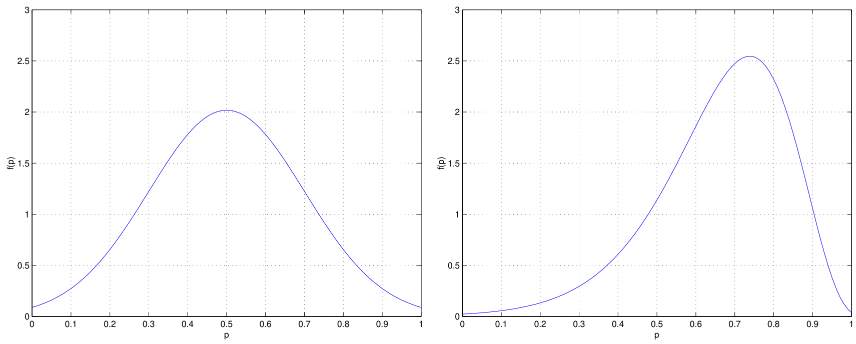

Fig. 5. Distributions assumed for the percentage (p) of objectionable images on objectionable websites.

As we only require a confidence level α, i.e., $1-\alpha$ correctness, we have much more freedom in choosing r and N. Our WIPE system can provide a performance with $q_1 = 96\%$ and $q_2 = 91\%$. The actual q_1 and q_2 in real world can be higher because icons and graphs on the Web can be easily classified by WIPE with close to 100% sensitivity. When we assume $f(p)$ being a truncated Gaussian with mean

$\bar{p} = 0.5$ and standard deviation $\sigma = 0.2$, which is plotted in Figure 5, we may test $N = 35$ images from each website and mark the website as objectionable once 7 or more images are identified as objectionable by the WIPE system. Under this configuration, we can achieve approximately 97% correctness for classifying both objectionable websites and benign websites.

If we fix the decision rule of our system, i.e., test a maximum of 35 images from each website and mark the website as objectionable once 7 or more images are identified as objectionable, the percentages of correctness for classification of both types of websites depend on the sensitivity parameter q_1 and the specificity parameter q_2. By fixing q_2 to 91% and changing q_1 between 80% to 100%, the percentages of correctness for both types of websites are shown in the left panel of Figure 6. Similarly, the results are shown in the right panel of Figure 6 for the case of fixing q_1 to 96% and changing q_2 between 80% to 100%. As shown in the graph on the left side, when q_2 is fixed, the correct classification rate for benign websites is a constant. On the other hand, the correct classification rate for objectionable websites degrades with the decrease of q_1. However, the decrease of the correct classification rate is not sharp. Even when $q_1 = 0.8$, the correct classification rate is approximately 95%. On the other hand, when $q_1 = 96\%$, no matter what q_2 is, the correct classification rate for objectionable websites is always above 92%. The rate of correctness for benign websites monotonically increases with q_2. Since benign images in an objectionable website are less likely to be classified as objectionable when q_2 increases, the number of objectionable images found in the set of test images is less likely to pass the threshold 7. As a result, the correct classification rate for objectionable websites decreases slightly with the increase of q_2. However, the correct classification rate will not drop below 92%.

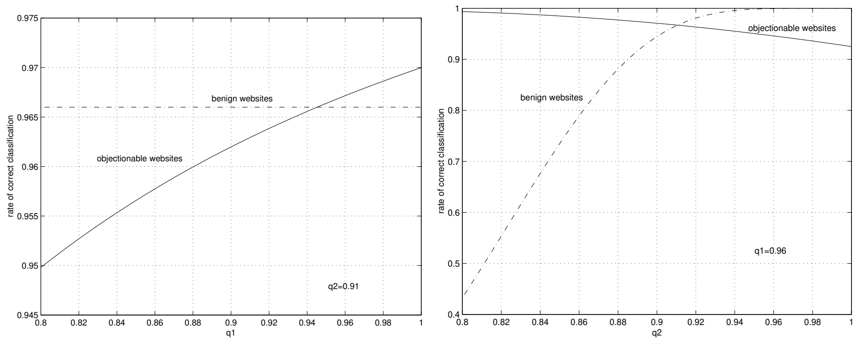

Fig. 6. Dependence of correct classification rates on sensitivity and specificity of WIPE (for the Gaussian-like distribution of p). Left: $q_2 = 91\%$, q_1 varies between 80% to 100%. Right: $q_1 = 96\%$, q_2 varies between 80% to 100%. Solid line: correct classification rate for objectionable websites. Dash dot line: correct classification rate for benign websites.

In the above statistical analysis, we assumed that the probability of an image being objectionable in an objectionable website has distribution $f(p)$ with mean 0.5. In real life, this mean value is usually higher than 0.5. With a less conservative hypothetical distribution of p, as shown in the right panel of Figure 5, we can achieve approximately 97% correctness by testing only 20 images from each website and marking the website as objectionable if 5 or more images are identified as objectionable by the WIPE system.

2.3 Limitations of the Algorithm

The screening algorithm in our IBCOW system assumes a minimum of N images downloadable from a given query website. For the current system set up, N can be as low as 20 for the less conservative assumption. However, it is not always possible to download 20 images from each website. We have noticed that some objectionable websites put only a few images on its front page for non-member netters to view without a password. For these websites, surround text will be more useful than images in the classification process. Also, we are considering to assign probabilities of objectionable to such sites based on accessible images.

In the statistical analysis, we assume each image in a given website is equally likely to be an objectionable image. This assumption may be false for some websites. For example, some objectionable websites put objectionable images in deep links and benign images in front pages.

3 Experimental Results

This algorithm has been implemented on a Pentium Pro 200MHz workstation. We selected 20 objectionable websites and 40 benign websites from various categories. It takes in general less than 2 minutes for the system to process each website. Besides the fast speed, the algorithm has achieved remarkable accuracy. It correctly identified all the 20 objectionable websites and did not mismark any one of the 40 benign websites. We expect the speed to be much faster once image and textual information is combined in the classification process.

4 Conclusions and Future Work

In this paper, we have presented the statistical analysis that provides us with the size of the sampling set and the number of objectionable images in that set needed to classify a website as objectionable. Using the WIPE system we developed, we have obtained a performance which already appears satisfactory for practical applications. Using statistical analysis, we expect the system to perform approximately 97% sensitivity and 97% specificity in classifying websites. Both the sensitivity and the specificity in real-world applications is expected to be much higher because our performance evaluation is relatively conservative and surrounding text can be used to assist the classification process.

We will further test the assumptions used in the statistical analysis part of this paper using real-world data. We are also working on refining both the WIPE and the IBCOW algorithms and the codes so that the system is more useful to real-world applications. Surround text will be used in the screening process. The algorithm can also be modified to execute in parallel on multi-processor systems. Experiments with our algorithm on video websites could be another interesting study.

References

1. C. Faloutsos et al, Efficient and Effective Querying by Image Content, *J. of Intelligent Information Systems*, 3:231-262, 1994.
2. Margaret Fleck, David A. Forsyth, Chris Bregler, Finding Naked People, *Proc. 4'th European Conf on Computer Vision*, UK, Vol 2, pp. 593-602, 1996.
3. David A. Forsyth et al, Finding Pictures of Objects in Large Collections of Images, *Proc. Int'l Workshop on Object Recognition*, Cambridge, 1996.
4. Alberto Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, Addison-Wesley Publishing Company, pp.99-110, 280-287, 1994.
5. Amarnath Gupta and Ramesh Jain, *Visual Information Retrieval*, Comm. of the ACM, vol.40 no.5, pp 69-79, 1997.
6. C. E. Jacobs et al., Fast Multiresolution Image Querying, *Proc. of SIGGAPH 95 Computer Graphics*, pp.277-286, August 1995.
7. J. R. Smith and S.-F. Chang, VisualSEEk: A Fully Automated Content-Based Image Query System, *ACM Multimedia Conference*, Boston, Nov 1996.
8. James Ze Wang et al., Wavelet-Based Image Indexing Techniques with Partial Sketch Retrieval Capability, *Proc. 4th ADL Forum (ADL'97)*, Washington D.C., May 1997.
9. James Ze Wang et al., System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms, *Proc. IDMS'97*, Springer-Verlag LNCS 1309, Sept. 1997.
10. James Ze Wang et al., Content-based Image Indexing and Searching Using Daubechies' Wavelets, *International Journal of Digital Libraries(IJODL)*, 1(4):311-328, Springer-Verlag, 1998.
11. James Ze Wang et al., System for Screening Objectionable Images, to appear in *Computer Communications Journal*, Elsevier, Amsterdam, 1998.