

# Practical capacity of digital watermark as constrained by reliability

Ryo Sugihara

Tokyo Research Laboratory, IBM Japan, Ltd.,  
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242-8502, Japan  
sugiryo@jp.ibm.com

## Abstract

*This paper presents a theoretic analysis of watermark capacity. First, a simplified watermark scheme is postulated. In the scheme, detection yields a multidimensional vector, in which each dimension is assumed to be i.i.d. (independent and identically distributed) and follow the Gaussian distribution. The major constraint on the capacity is detection reliability, which is one of the most important measures of the utility of watermarks. The problem is to figure out the maximum amount of information payload with the reliability requirement still satisfied. The reliability is represented by three kinds of error rates: the false positive error rate, the false negative error rate, and the bit error rate. These error rates are formulated under certain assumptions, and the theoretical capacity can be determined by setting the bounds on all of the error rates. Further, experiments were performed to verify the theoretic analysis, and it was shown that this approach yields a good estimate of the capacity of a watermark.*

## 1 Introduction

Most of the previous works on watermark capacity[1, 5] are based on information-theoretic considerations. They regard watermarking procedures as a communication over a noisy channel, and the capacity of the watermark corresponds to the communication capacity of the “watermark-channel”.

But the information payload of a watermark is usually smaller than these results indicate. One of the reasons for overestimation in information-theoretic studies is that they aim solely at maximizing mutual information, and do not care much about the reliability of detection. However in practical frameworks for watermarking, there are usually some constraints on reliability. Most of the previous studies ignore the case of erroneous detections, which includes both the case where a watermark is detected from non-watermarked content, and where a different watermark

is extracted from watermarked content. For these reasons, there are difficulties in applying their theoretical results to actual watermarking schemes. In other words, they are not focused on the “practical” capacity.

The capacity of a watermark is usually constrained by fidelity, robustness, and reliability[3, 4]. Of these constraints, fidelity will not be discussed here, since it has little relation to how many bits are in an image. Robustness can be considered as equivalent to reliability, as it is considered to be a measure related to the probability of correct detection after some degradation.

In this paper, we discuss the capacity of watermarks where there are some constraints on reliability. First we define the measurements of reliability and the constraints on those measurements. Second we consider a simple watermarking scheme and make some assumptions about its attributes. Based on these assumptions, we derive the theoretical probability of detection errors. Also we built a prototype watermarking system with those attributes, and compared the theoretical and experimental results. We conclude our reliability-driven approach for estimating the capacity of watermarks is more practical than the information-theoretic approaches, and is therefore more directly useful in real applications.

## 2 Problem statement

In short, the problem handled in this paper is to maximize the number of bits embedded in the image, when there are some constraints on the reliability of detection. Reliability can be measured by three metrics: the false positive error rate, the false negative error rate, and the bit error rate. The false positive error rate is defined as the probability that a watermark is detected from non-watermarked content. The false negative error rate is for the case that the watermark is not detected in watermarked content. The bit error rate is usually defined as the percentage of erroneous bits in an extracted message out of all bits. However in this paper, we define the bit error rate as the probability that a different watermark is detected in watermarked content.

Most of the applications of watermarks, such as copyright protection, require very high reliability, and it is useless in practice if there are any errors in the extracted messages. In other words, the number of error bits does not matter much, but whether or not there are any error bits does matter, so we use this restricted definition for bit error rate.

The constraints can be written as

$$P_{fp} \leq P_{fpmax} \quad (1)$$

$$P_{be} \leq P_{bemax} \quad (2)$$

$$P_{cd} \geq P_{cdmin} \quad (3)$$

where  $P_{fp}$  is false positive error rate,  $P_{be}$  is bit error rate, and  $P_{cd}$  is rate of correct detection, which is defined as  $P_{cd} \triangleq 1 - P_{fn}$ .  $P_{fn}$  is false negative error rate.

Here we must think about how much degradation is to be considered. If there is no limit, these constraints cannot be satisfied even for one bit embedding, because  $P_{cd}$  gets closer to  $P_{fp}$  as the image is more degraded. This holds true for almost any watermarking scheme.

## 2.1 Simple watermarking scheme

In order to solve the problem, we model a simple watermarking scheme. We postulate that the scheme has the following attributes:

- statistical watermarking, i.e. detection yields  $n$ -dimensional values  $x_i (1 \leq i \leq n; n$  is the number of bits embedded in an image) and they are statistically tested as follows:

$$\begin{cases} \text{if } |x_i| \geq T \ (\forall i) & \text{detected} \\ \text{else} & \text{not detected} \end{cases}$$

where  $T (> 0)$  is a predetermined threshold. In the first case, the message is extracted as follows<sup>1</sup>:

$$\begin{cases} x_i \geq T & \text{"1" for the } i\text{-th bit} \\ x_i \leq -T & \text{"0" for the } i\text{-th bit} \end{cases}$$

- multiple bit embedding
- $x_i$  is i.i.d.(independent and identically distributed) and follows a Gaussian distribution
- for the cover-image,  $x_i$  is distributed according to  $N(0, 1^2)$ , where  $N(\mu, \sigma^2)$  stands for the Gaussian distribution with mean  $\mu$  and standard deviation  $\sigma$
- for a stego-image,  $x_i$  is distributed according to  $N(\mu_{max}^{(n)}, 1^2)$  ( $\mu_{max}^{(n)} \geq 0$ )

<sup>1</sup>For simplification, the message is always "11...1" hereafter

- When the stego-image is degraded,  $x_i$  is distributed according to  $N(\mu^{(n)}, 1^2)$  ( $0 \leq \mu^{(n)} \leq \mu_{max}^{(n)}$ ). Note that the standard deviation of  $x_i$  is assumed to be held constant at 1.0 by an ideal normalization.

- $\mu^{(n)}$  is proportional to  $\sqrt{1/n}$

Actually, the "patchwork algorithm"[2] is one of the best and simplest examples with these attributes. For multiple bit embedding, we divide the pixels equally to represent each bit.

## 3 Theoretical analysis

First, we explain how we treat detection error rates theoretically.

### 3.1 Formulation of error rates

#### 3.1.1 Basic formulae

Let  $f_{\mu^{(n)}, \sigma^{(n)}}(x)$  denote the probability density function of the detection statistic from each bit. It is assumed to follow a Gaussian distribution written as

$$f_{\mu^{(n)}, \sigma^{(n)}}(x) = \frac{1}{\sqrt{2\pi\sigma^{(n)2}}} \exp\left\{-\frac{(x - \mu^{(n)})^2}{2\sigma^{(n)2}}\right\} \quad (4)$$

The probability of correct detection ( $p_c^{(n)}$ ) and wrong detection ( $p_w^{(n)}$ ) are given by

$$p_c^{(n)} = \int_{T^{(n)}}^{\infty} f_{\mu^{(n)}, \sigma^{(n)}}(x) dx \quad (5)$$

$$p_w^{(n)} = \int_{-\infty}^{-T^{(n)}} f_{\mu^{(n)}, \sigma^{(n)}}(x) dx \quad (6)$$

where  $T^{(n)}$  is the threshold when the number of bits is  $n$ .

The probability of  $i$ -bit error detection out of  $n$  bits is as follows:

$$p_e^{(n)}(i) = \binom{n}{i} \{p_w^{(n)}\}^i \{p_c^{(n)}\}^{n-i} \quad (7)$$

where  $\binom{n}{i}$  is the number of combinations for choosing  $i$  items out of  $n$ .

#### 3.1.2 For non-watermarked images

When the image is not watermarked, each detection statistic follows the standard Gaussian distribution  $N(0, 1^2)$ . The false positive detection rate for one bit is

$$\begin{aligned}
p_0^{(n)} &= \int_{T^{(n)}}^{\infty} f_{0,1}(x)dx + \int_{-\infty}^{-T^{(n)}} f_{0,1}(x)dx \\
&= \operatorname{erfc}\left(\frac{T^{(n)}}{\sqrt{2}}\right)
\end{aligned} \tag{8}$$

where  $\operatorname{erfc}(x)$  is the complementary error function defined as follows:

$$\operatorname{erfc}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2) dt \tag{9}$$

Using  $p_0^{(n)}$ , the false positive detection rate ( $P_{fp}^{(n)}$ ), and the no detection rate ( $P_{no}^{(n)}$ ), are given as follows:

$$P_{fp}^{(n)} = \left\{ p_0^{(n)} \right\}^n \tag{10}$$

$$P_{no}^{(n)} = 1 - P_{fp}^{(n)} \tag{11}$$

### 3.1.3 For watermarked images

When the image is watermarked, the probabilities of correct detection ( $P_{cd}^{(n)}$ ), bit error ( $P_{be}^{(n)}$ ), and no detection ( $P_{no}^{(n)}$ ) can be written:

$$P_{cd}^{(n)} = p_e^{(n)}(0) \tag{12}$$

$$P_{be}^{(n)} = \sum_{i=1}^n p_e^{(n)}(i) \tag{13}$$

$$P_{no}^{(n)} = 1 - P_{cd}^{(n)} - P_{be}^{(n)} \tag{14}$$

## 3.2 Theoretical capacity

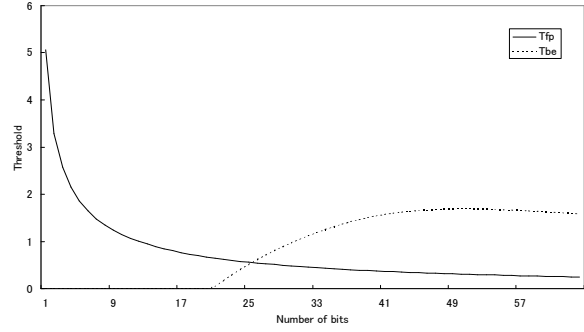
Under the current assumptions, the probability of a detection error for a particular value of  $n$  is totally controlled by the threshold value. In other words, we have to figure out the threshold value to satisfy the constraints on error rates.

As for the false positive error, the threshold value is derived from equations (8) and (10) as follows

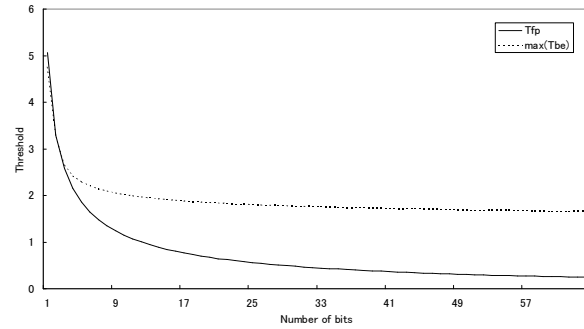
$$\begin{aligned}
T_{fp}^{(n)} &= \sqrt{2} \operatorname{erfc}^{-1}(p_0^{(n)}) \\
&= \sqrt{2} \operatorname{erfc}^{-1}\left\{ P_{fp}^{(n)\frac{1}{n}} \right\}
\end{aligned} \tag{15}$$

where  $T_{fp}^{(n)}$  is the threshold value constrained by the false positive error rate, and  $\operatorname{erfc}^{-1}(x)$  is the inverse function of the complementary error function. Note that  $T_{fp}^{(n)}$  only satisfies the constraint on the false positive error rate, and does not necessarily satisfy the bit error rate constraint.

For the bit error rate, the problem is much more difficult to solve equation (13) analytically. However, we can find an



**Figure 1.** Threshold ( $T_{fp}^{(n)}, T_{be}^{(n)}$ ). Here  $P_{fpmax} = P_{bemax} = 10^{-6}, \mu^{(12)} = 7.0, \sigma^{(n)} = 1.0$ .



**Figure 2.** Threshold ( $T_{fp}^{(n)}, \max_{\mu^{(n)}} \{ T_{be}^{(n)} \}$ ) when  $P_{fpmax} = P_{bemax} = 10^{-6}, 0 \leq \mu^{(12)} \leq 10.5, \sigma^{(n)} = 1.0$ .

approximate threshold value for each distribution, by using an iterative algorithm such as the Newton method.

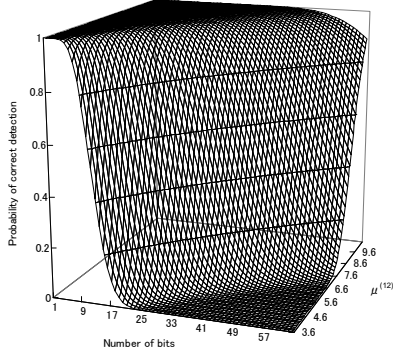
For example, if  $P_{fpmax} = P_{bemax} = 10^{-6}, \mu^{(12)} = 7.0^2$ , and  $\sigma^{(n)} = 1.0$ , Fig. 1 shows the threshold values ( $T_{fp}^{(n)}, T_{be}^{(n)}$ ).

The probability of the bit error rate must not exceed  $P_{bemax}$  in any of the distributions, that are defined by  $\mu^{(n)}$ . So the threshold is given as

$$T^{(n)} = \max \left\{ T_{fp}^{(n)}, \max_{0 \leq \mu^{(n)} \leq \mu_{max}^{(n)}} \{ T_{be}^{(n)} \} \right\} \tag{16}$$

where  $\mu_{max}^{(n)}$  is the maximum value of  $\mu^{(n)}$  and corresponds to the value when the embedded image suffers no degradation.

<sup>2</sup>Here,  $\mu^{(12)}$  represents all  $\mu^{(n)}$ . The values for other  $n$  is calculated according to  $\mu^{(n)} \propto \sqrt{1/n}$ , e.g.  $\mu^{(48)} = 3.5$



**Figure 3. Theoretically derived probability of correct detection ( $P_{cd}^{(n)}$ ), when  $P_{f_{pmax}} = P_{b_{max}} = 10^{-6}$ ,  $3.6 \leq \mu^{(12)} \leq 10.5$ ,  $\sigma^{(n)} = 1.0$ .**

Fig. 2 shows an example of two threshold values when  $\mu^{(12)}$  takes any positive value below 10.5, and  $\sigma^{(n)}$  is a constant 1.0.

Fig. 3 shows the relationship between the number of bits,  $\mu^{(12)}$ , and  $P_{cd}^{(n)}$ . The conditions are almost the same as in Fig. 2, but the minimum value of  $\mu^{(12)}$  ( $\equiv \mu_{min}^{(12)}$ ) is now set to 3.6, which is an example of limit of degradation. It can be seen that  $\mu_{min}^{(n)}$  is the primary constraint on the capacity. Actually, it is almost sufficient to consider only  $\mu^{(n)} = \mu_{min}^{(n)}$  to estimate the capacity.

From Fig. 3, the capacity of the watermark can be determined. For example, if  $P_{cdmin} = 0.5$ , the capacity is 12 bits ( $P_{cd}^{(12)}|_{\mu^{(12)}=3.6} = 0.533$ ).

## 4 Experiments

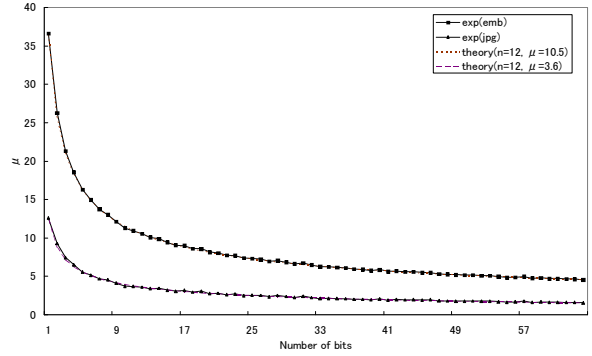
As we discussed in the previous section, we can estimate the capacity of watermark theoretically, if the requirements on error rates ( $P_{f_{pmax}}, P_{b_{max}}, P_{cdmin}$ ) and the limit of degradation ( $\mu_{min}^{(n)}$ ) are given. Note that  $\mu_{min}^{(n)}$  has to be given for any particular  $n$ , as we assume  $\mu^{(n)} \propto \sqrt{1/n}$ .

For the verification of the analytical results, we implemented a prototype watermarking system and performed experiments. As mentioned previously, we used the patchwork algorithm, and modified it to realize multiple bit embedding.

The constraints were as follows:

- reliability :  $P_{f_{pmax}}, P_{b_{max}} = 10^{-6}$ ,  $P_{cdmin} = 0.5$
- survive through JPEG compression (Quality: 50)

In the experiment, 1000 images (resolution:  $640 \times 426$ ) were used. We tested for  $n$  between 1 and 64 bits. On em-



**Figure 4. The number of bits and  $\mu^{(n)}$ , the average detection statistic. Experiments were performed on 1000 images, and the results for uncompressed images and JPEG-compressed images are shown.**

bedding, the change for each pixel was constant  $\pm 5$ , except for flat regions that were not changed<sup>3</sup>.

Before the consideration of capacity, we performed a preliminary experiment to measure how much the watermarks were degraded by the JPEG compression. In the preliminary experiment,  $n$  was set to 12. The resulting averages for the detection statistics over 1000 images were 10.548 for the uncompressed images, and 3.598 for the JPEG-compressed images, respectively. Based on these results, we used  $\mu^{(12)} = 10.5$  for uncompressed images, and 3.6 for JPEG-compressed images for the theoretical analysis. As we consider JPEG compression as causing the heaviest degradation here,  $\mu_{min}^{(12)}$  is assumed to be 3.6.

Fig. 4 shows the relationship between the number of bits and the average detection statistic ( $\mu^{(n)}$ ). The experimental data are plotted over the theoretical curves, and it shows that  $\mu^{(n)} \propto \sqrt{1/n}$  is a valid assumption.

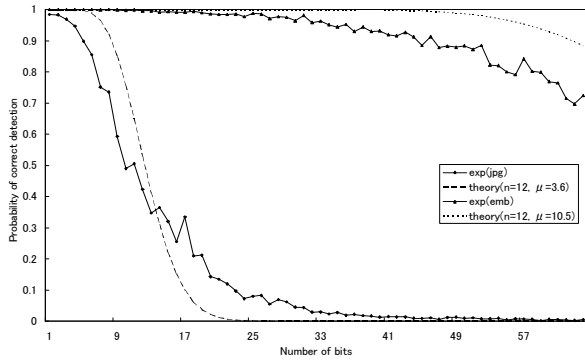
Fig. 5 shows the relationship between the number of bits and the probability of correct detection. Here we also included the experimental data on the theoretical curves.

From this figure, and remembering that  $\mu_{min}^{(12)}$  is assumed to be 3.6, the capacity is approximately 11 bits. This is nearly equal to the theoretical result of the previous section.

## 5 Discussion

We have shown that the experimental results nearly follow the theoretical calculations, which shows that our underlying model is sound. The discrepancy in cases where

<sup>3</sup>The resulting SNR(Signal-to-Noise Ratio) was 28.56dB on average of 1000 images.



**Figure 5. Probability of correct detection ( $P_{cd}^{(n)}$ ) and the number of bits. The details are the same as for Fig. 4.**

$P_{cd}$  is very high or very low are presumably caused by the variation between each image. In other words, the amount of changes from embedding varies between images, and the detection statistic  $x_i$  should not be considered as following the same distribution over all images. Moreover, there is also a large variation in how much the image degrades from the same degree of compression. But it is still remarkable that the theoretical results are so close to the experiments, even though we used such a rough assumption for the attributes of detection statistics.

It might seem to be strange that the resulting capacity is much smaller than not only the information-theoretic capacity, but also the payload in the commercial products of watermark. It is because the prototype watermarking system used in the experiment was not designed to be robust against compressions. The watermark pattern was concentrated on high spatial frequency, which is susceptible to most of the image compression algorithms. If it were designed to be robust, for example by using larger “patch” for the patchwork algorithm, the value of  $\mu^{(n)}$  would not fall down so much. The resulting capacity becomes large when the value of  $\mu^{(n)}$  is large, as seen from Fig. 3.

## 6 Conclusion

We have described one approach for the theoretical analysis of watermark capacity. We considered a simple statistical watermarking scheme based on the patchwork algorithm, and formulated the probability of detection errors theoretically with some assumptions. The parameters concerning the distribution of detection statistics were determined by a preliminary experiment. Using these parameters, the theoretical capacity of the watermark was calculated for the given constraints. We successfully verified

the results by further experiments, and showed that our approach is much more practical than the usual information-theoretic approach.

However, there are several problems and questions. The major problem is that our assumptions for the detection statistic are too strict and idealistic. Also the experimental watermarking algorithm is too crude for actual use, as it does not protect the fidelity of the image well enough. Moreover, there are some questions about how the error correcting codes (ECCs) would affect these results, how we could apply these results to video and/or audio watermarking schemes, and so on. Such topics remain to be investigated in our future work.

## References

- [1] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, “Capacity of the Watermark-Channel: How Many Bits Can Be Hidden Within a Digital Image?”, *Proceedings of SPIE*, Vol. 3657, pp. 437-448, 1999
- [2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for data hiding”, *IBM Systems Journal*, Vol. 35, No. 3/4, pp. 313-336, 1996
- [3] J. Hernandez and F. Perez-Gonzalez, “Statistical analysis of watermarking schemes for copyright protection on images”, *Proceedings of IEEE*, Vol. 87, No. 7, pp. 1142-1166, 1999
- [4] S. Katzenbeisser and F.A.P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, 2000
- [5] M. Ramkumar and A.N. Akansu, “Theoretical Capacity Measures for Data Hiding in Compressed Images”, *Proceedings of SPIE*, Vol. 3528, pp. 482-492, 1998