6

Data recovery

The definition of a reliable procedure to retrieve the information hidden within the host signal is of fundamental importance for the proper development of any data hiding system. This is not an easy task, because of the many modifications the host asset may undergo after embedding. As a matter of fact, modelling the watermark channel in the presence of attacks is a very complicated problem, due to the wide variety of possible attacks to be taken into account. Another factor which somewhat complicates the recovery of the hidden information is the unavailability of the original asset at the detector/decoder. As we will discuss in chapter 9, in line of principle this is not a problem, however, in practice, blind systems are much more complicated than non-blind ones.

Due to the wide variety of attacks and to the difficulties of developing an accurate statistical model of host features, the structure of the detector/decoder is, usually, derived by considering a simplified channel model. The performance of the system in the presence of more complicated channels, then, is evaluated either theoretically (by assuming that the detector/decoder structure is known) or experimentally. In this chapter we follow a similar approach: we derive the detector/decoder structure in some simple cases, dealing with over-simplified channel models, where attacks are either absent or modelled as noise addition. Then we evaluate the error probability of the system for the simplified channel, being aware that a more accurate, experimental, analysis is needed to assess the performance of the systems in more realistic situations.

We first consider the detection problem, in which the detector is only asked to decide whether the asset at hand contains a given watermark or not. Then we pass to the decoding problem, in which the decoder has to take a decision on which message, among those possible, is actually em-

bedded within the asset under analysis. In this second case, it is usually assumed that the host asset is surely marked, nevertheless this is not necessarily the case. For this reason, we also touch the problem of watermark presence assessment in readable watermarking systems.

6.1 Watermark detection

We start by considering the detection problem, i.e. given a digital asset A and a watermark code $\mathbf{b}$, decide whether A contains $\mathbf{b}$ or not. Of course, the problem depends on the particular embedding rule adopted by the system. When embedding passes through the injection of a watermarking signal $\mathbf{w}$ into the host feature set, the problem can be easily formulated as one of signal detection in a noisy environment, where noise accounts for both the unknown host signal and the possible presence of attacks. In informed embedding systems, the situation is rather different. In line of principle, the detector structure, i.e. the detection regions associated to each watermark message, could be defined without making any reference to the embedding process, for example, by adopting random coding arguments. In this case embedding would reduce to mapping the host asset into the detection region subject to the invisibility constraint. Moreover, the performance of the system would largely depend on the adopted embedding strategy and the asset at hand, since it would depend on the position of the marked asset within the embedding region.

In most cases, though a more tortuous path is followed. Detection regions are computed (often optimally) by assuming that a blind embedding strategy is adopted (e.g. by applying additive or multiplicative spread spectrum watermarking), then the informed embedding paradigm is applied to actually watermark the host asset[1]. Additionally, (informed coding) the same watermark may be associated to several detection regions. If the above approach is followed, informed embedding does not have any impact on the structure of the detector, however it contributes to diminish the error probability (or improve imperceptibility, or increase the payload). The actual computation of the error probability in the presence of informed embedding is usually a cumbersome task, since it intimately depends on the host asset. The only noticeable exception is when a constant robustness strategy is adopted (see "MD" approach in section 4.3.1), since in this case the same error probability is obtained regardless of the asset at hand.

With the above considerations in mind, while deriving the structure of the various detectors proposed so far, we will not take explicitly into account whether blind or informed embedding was used, since we will always assume that detection amounts to the extraction of a signal $\mathbf{w}$ immersed within noise.

In the attempt to be as general as possible, we first reformulate the detection problem as a classical hypothesis testing problem. Then we consider more specific situations in which simple statistical models are used to characterize the host feature set and attack noise. We develop most of our analysis by considering single channel systems (e.g. the watermarking of grey level images), where features assume scalar values. Only at the end of the section, we will give some hints on how the analysis can be extended to multichannel cases.

[1] This way of designing the embedding and detection part of a watermarking system is not necessarily optimum, since, in general, the joint optimization of the detection and embedding processes would be preferable, however no such system has been developed yet for detectable watermarking.

6.1.1 A hypothesis testing problem

In order to formalize the detection problem, let us assume we want to verify whether an asset A' contains the watermark code $\mathbf{b}^*$ or not. The host asset is indicated by A' instead of A or $A_{\mathbf{w}}$ to make it explicit that A' may coincide neither with the original asset nor with the marked asset $A_{\mathbf{w}}$. Dealing with blind embedding systems, we can assume that looking for the presence of $\mathbf{b}^*$ within A' amounts to looking for the presence of a certain watermark signal $\mathbf{w}^*$ [2]. The decision must be taken on the basis of a set of observed variables coinciding with the set of features $\mathbf{f}'$ extracted from A'. In the framework of statistical detection theory, the above problem corresponds to deciding in which of a finite number of states the observed system, namely the possibly marked asset, resides. More specifically, let us consider the following alternative hypothesis:

H_0: A' does not contain $\mathbf{w}^*$;

H_1: A' contains $\mathbf{w}^*$.

where H_0 is a composite hypothesis accounting for the following two situations:

Case a_0: A' is not watermarked;

Case b_0: A' contains a watermark other than $\mathbf{w}^*$.

Watermark detection amounts to defining a test of the simple hypothesis H_1 versus the composite alternative H_0 that is optimum with respect to a certain criterion.

[2] In some cases, e.g. with certain informed embedded systems, the detector has to look for the presence of one out many signals associated to the same message $\mathbf{b}^*$.

Likelihood ratio

In Bayes theory of hypothesis testing, the criterion is minimization of risk. Bayes risk is defined as the average of a loss function L_{ij}, where L_{01} is the loss sustained when hypothesis H_0 is in force but H_1 is chosen, and L_{10} is the loss sustained when hypothesis H_1 is in force and H_0 is chosen. By remembering that in our case observation variables correspond to the vector $\mathbf{f}'$, the decision criterion can be given the form of a decision rule Φ mapping each $\mathbf{f}'$ into 1 or 0, corresponding to H_1 and H_0:

$$\Phi(\mathbf{f}') = \begin{cases} 1, & \mathbf{f}' \in R_1 \quad (H_1 \text{ is in force}) \\ 0, & \mathbf{f}' \in R_0 \quad (H_0 \text{ is in force}) \end{cases} \tag{6.1}$$

where R_1 and R_0 are acceptance and rejection regions for hypothesis H_1. Minimization of Bayes risk leads to a decision criterion which is based on the, so called, likelihood ratio $\ell(\mathbf{f}')$:

$$\ell(\mathbf{f}') = \frac{p(\mathbf{f}'|H_1)}{p(\mathbf{f}'|H_0)}, \tag{6.2}$$

where $p(\mathbf{f}'|H_i)$ is the pdf of vector $\mathbf{f}'$ conditioned to hypothesis H_i. More specifically, minimum Bayes risk is achieved by letting:

$$R_1 = \{\mathbf{f}' : \ell(\mathbf{f}') > p_0 L_{01}/p_1 L_{10}\}, \tag{6.3}$$

or, equivalently:

$$\Phi(\mathbf{f}') = \begin{cases} 1, & \ell(\mathbf{f}') > p_0 L_{01}/p_1 L_{10} \\ 0, & \text{otherwise} \end{cases} \tag{6.4}$$

where p_0 and p_1 are the a priori probabilities of H_0 and H_1.

The exact specification of Φ requires that the watermark embedding rule is specified and that both the host features and the attack noise are characterized statistically, which will be the goal of next sections.

Threshold selection

By analyzing the decision rule defined by equation (6.4), we can see that the detector operates by comparing the likelihood ratio $\ell(\mathbf{f}')$ against a detection threshold λ, where:

$$\lambda = \frac{p_0 L_{01}}{p_1 L_{10}}. \tag{6.5}$$

A common approach to set λ consists in trying to minimize the overall error probability P_e. By letting P_f be the probability of revealing the presence of $\mathbf{w}^*$ when $\mathbf{w}^*$ is not actually present (false alarm probability),

and $P_m = 1 - P_d$ the probability of missing the watermark presence (P_d denotes the probability of correctly revealing the watermark), we have:

$$P_e = p_0 P_f + p_1(1 - P_d). \tag{6.6}$$

From decision theory it is known that to minimize P_e we must set $\lambda = 1$, which corresponds to the common situation in which $L_{01} = L_{10}$ and $p_0 = p_1$. Note also that in this case we have $P_f = P_m$, that is the minimum error probability is obtained by letting the probability of missing the watermark and that of falsely revealing its presence equal.

A problem with the above, minimum error, detector is that usually the model used to derive the detection rule only accounts for very simple attacks, e.g. addition of white Gaussian noise. When facing different kinds of attacks, however, it leads to a probability of missing the watermark which is considerably higher than the probability of falsely revealing the watermark presence, which is not a desirable behavior in many cases (the reason for such a behavior will be clear after the analysis in the next sections). In addition, in many applications, false detection probability can not fall below a certain level, regardless of the probability of missing the watermark. In these cases, it is preferable to minimize the probability of missing the watermark subject to a constraint on the maximum false detection probability. This is the aim of the Neyman-Pearson detection criterion,False alarm probability!Neyman Pearson criterion according to which the probability of correctly detecting the watermark is maximized subject to a prescribed limit on P_f.

As for the Bayes criterion, detection relies on the comparison of the likelihood ratio against a threshold λ:

$$\Phi(\mathbf{f}') = \begin{cases} 1, & \ell(\mathbf{f}') > \lambda \\ 0, & \text{otherwise} \end{cases} \tag{6.7}$$

what changes here is how the threshold is computed. More specifically, λ is calculated so that the desired false detection probability is achieved, i.e. we must have:

$$P\{\ell(\mathbf{f}') > \lambda | H_0\} = \bar{P}_f; \tag{6.8}$$

where $\bar{P}_f$ is the target false detection probability. By letting $p(\ell|H_0)$ be the pdf of ℓ under hypothesis H_0, we can rewrite the condition in (6.8) as:

$$\int_{\lambda}^{+\infty} p(\ell|H_0)d\ell = \bar{P}_f. \tag{6.9}$$

By solving the above equation for λ, we obtain the threshold to be adopted in the Neyman-Pearson criterion. Once λ has been fixed (thus determining

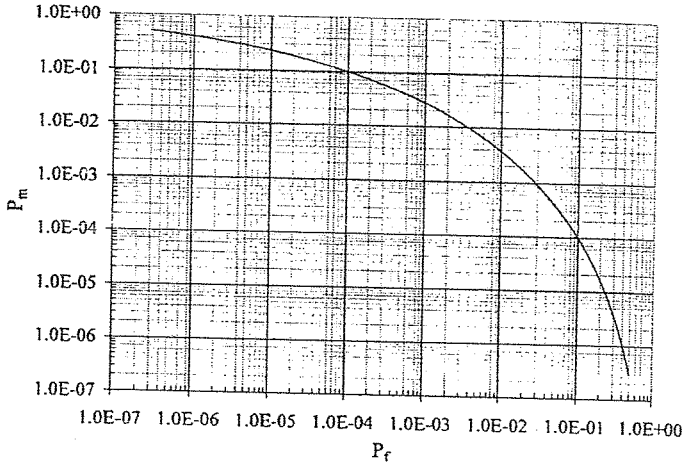

Figure 6.1: Example of a ROC curve defining the characteristic of a watermark detector. Each point of the curve corresponds to a different detection threshold, for which the false alarm and missed detection probabilities given in the figure are obtained.

P_f), the probability of missing the watermark can be calculated as:

$$P_m = \int_{-\infty}^{\lambda} p(\ell|H_1)d\ell. \tag{6.10}$$

The performance of a detector based on the Neyman-Pearson criterion are usually expressed by ROC (Receiver Operating Characteristic) curves, in which P_m is plotted against P_f, as exemplified in figure 6.1[3].

It is often convenient to replace the likelihood ratio with a log-likelihood ratio, defined as:

$$\mathcal{L}(\mathbf{f}') = \ln \ell(\mathbf{f}'), \tag{6.11}$$

leading to a decision rule having the form:

$$\mathcal{L}(\mathbf{f}') \lesseqgtr \ln \lambda = \Lambda, \tag{6.12}$$

where Λ can be derived by exploiting equation (6.9), or calculated directly by letting:

$$P\{\mathcal{L}(\mathbf{f}') > \Lambda|H_0\} = \bar{P}_f; \tag{6.13}$$

[3]Instead of P_m, sometimes P_d is plot against P_f, however the meaning of ROC curves does not change

yielding:

$$\int_{\Lambda}^{+\infty} p(\mathcal{L}|H_0)d\mathcal{L} = \bar{P}_f. \tag{6.14}$$

The actual implementation of a watermark detector based on the Neyman-Pearson criterion, requires that the exact relationship between the watermark signal, the host features and attack noise is defined. This in turn, requires that the watermark embedding rule is specified and that proper statistical models are available to describe the watermark, the host features and the noise introduced by attacks. With reference to noise modelling, it has to be noted that in some cases, the detector structure is derived in the absence of attacks, the only source of uncertainty at the detector being the host features. In such a case, the optimality of the detector is clearly lost, when attacks are taken into account, and the robustness of the watermark in the presence of attacks must be verified experimentally.

In the sequel we apply the above statistical framework to some practical cases, thus deriving the structure of some of the most popular watermark detectors. We start by considering a very simple case following the basic AWGN (Additive White Guassian Noise) assumptions, to finish with more complicated situations dealing with multiplicative watermarks hosted by non-Gaussian features.

6.1.2 AWGN channel

The simplest channel model to deal with is the Additive White Gaussian channel model, in which both host features and attacks are modelled as uncorrelated, Gaussian noise added to the watermark signal. In this case we can write:

$$f_{w,i} = f_i + \gamma w_i + n_i, \tag{6.15}$$

where n_i is a white Gaussian noise accounting for attacks. Actually, the degradation introduced by attacks is much more complicated than pure white noise addition, however by modelling attacks as additive white noise, the problem is considerably simplified, thus allowing the derivation of the detector structure in closed form. Additionally, under certain assumptions, AWGN channel may be considered as a worst-case attack with the corresponding analysis giving an upper bound on the achievable performance.

As to host features, we are making two important assumptions. The first one is that f_i's follow a Gaussian distribution. This is only rarely the case, hence the results we derive below are clearly suboptimal, thus calling for an experimental validation of detector performance. The second assumption is that host features, as well as noise, form an uncorrelated