GRAPHICAL MODELS FOR DESYNCHRONIZATION-RESILIENT WATERMARK DECODING

Shankar Sadasivam, Pierre Moulin

Beckman Inst., Coord. Sci. Lab and ECE Dept.,
Univ. of Illinois at Urbana-Champaign, USA.
{ssadasi2,moulin}@ifp.uiuc.edu

Ralf Koetter

Inst. for Communications Engg.,
Technical Univ. of Munich, Germany.
ralf.koetter@tum.de

ABSTRACT

The performance of current blind watermark decoders against desynchronization attacks is rather poor. Such attacks include filtering, amplitude modulation, gamma correction, time-varying delays, and spatial warping. We propose a new family of watermark decoders based on modern methods for iterative decoding using graphical models. This approach addresses the "curse of dimensionality" problem that seemingly results when the desynchronization parameter space has high dimensionality.

Index Terms: Watermarking, data hiding, desynchronization, coding, graphical models

1. INTRODUCTION

One of the main technical obstacles to the deployment of watermarking systems has been the limited resilience of commonly employed coders and decoders to *desynchronization attacks*. Such attacks include filtering, amplitude modulation, gamma correction, time-varying delays, and spatial warping. The problem can be addressed by embedding watermarks in a domain that is invariant to such operations [1, 2], by embedding pilots (synchronization sequences) [3, 4, 5, 6, 7] or by jointly decoding the watermark and estimate the desynchronization attack parameters [8, 9, 10]. Our recent research has established optimality properties of the last approach [11, 12] in an asymptotic setup, but the practicality of this approach is a concern due to the need for a search over a possibly large parameter space.

This paper introduces a practical computational framework for decoding in the presence of desynchronization attacks, using graphical models for the host signal, the watermarking code, and the attack channel. The reader is referred to the books by Lauritzen [13], Pearl [14] and Frey [15] as well as the articles [16, 17, 18] for a general introduction to graphical models. These models appear to be particularly appropriate for watermarking of media signals because the underlying probabilistic models are local, and inference problems such as watermark decoding can be solved using iterative belief propagation algorithms. In our study of this problem, we have chosen to focus on the problem of *blind decoding*, in which the host signal is not known to the decoder. The reason for this choice is that it is known be be an extremely challenging problem [19]. Only limited success has been achieved against desynchronization attacks on blind watermarking systems, and this limited success was obtained using simple attacks – such as a pure delay, or a pure amplitude scaling. In contrast, the methodology presented here is applicable to fairly general desynchronization attack models involving large parameter spaces. We illustrate our approach with numerical results for an *amplitude modulation* attack.

This research was supported in part by NSF grant CCR 03-25924.

2. DESYNCHRONIZATION-RESILIENT DECODING

A fairly general communication model for watermark decoding is depicted in Fig. 1. A boldface notation is used for sequences and vectors. A message (digital signature) m is embedded in a length-N host sequence $\mathbf{s} = \{s(1), \cdots, s(N)\}$, aided by side information $\mathbf{k}$ shared with the receiver. The signal after embedding is denoted by $\mathbf{x} = f(\mathbf{s}, m, \mathbf{k})$. In some cases, no embedding takes place (say when $m = \emptyset$), and the encoding function f simply reproduces $\mathbf{s}$. The receiver does not observe $\mathbf{x}$ directly, but at the output of an *insecure channel* modeled by a conditional distribution $p(\mathbf{y}|\mathbf{x})$. For instance $p(\mathbf{y}|\mathbf{x})$ could be a simple memoryless channel, such as an additive white Gaussian noise channel. But the insecure channel need not be memoryless or even causal. The watermarking problem is therefore intimately related to the problem of communication over an uncertain channel. There exist well established information-theoretic methods to analyze such problems [20, 21].

The receiver has access to $\mathbf{y}$ and $\mathbf{k}$ and produces an estimate of m. The side information $\mathbf{k}$ may be a cryptographic key, but may also be used to convey information about $\mathbf{s}$ to the receiver. A *blind receiver* is not given access to $\mathbf{s}$; this problem is the most challenging one, calling for the use of special codes (binning codes) to deliver high performance

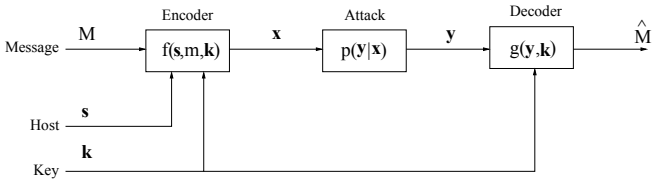

Fig. 1. Communication model for watermarking.

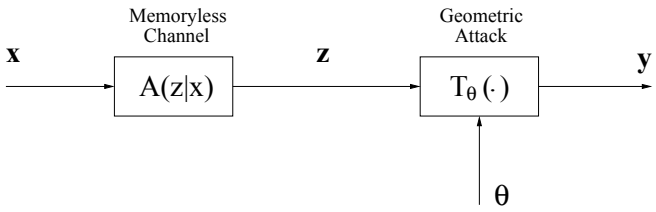

Fig. 2. Model for desynchronization attacks.

[19]. The setup of Fig. 1 is applicable both to authentication problems (in which case the set $\mathcal{M}$ of possible messages is typically small) and to data hiding (in which case $\mathcal{M}$ is large, typically exponentially increasing in the length of the sequence $\mathbf{s}$).

The decision rules used by blind watermarking decoders are generally minimum-distance rules or variations thereof. Such decoders – especially those used in conventional binning schemes for blind watermarking – typically fail when the channel introduces signal processing operations such as filtering, amplitude scaling, modulation, delays, warping, etc. The perceptual effects of such operations are normally quite weak, but the receiver is desynchronized, and its decoding performance can be catastrophic. In our communication model, desynchronization operations are absorbed as part of the channel model class $\mathcal{P}$. A typical formulation of the problem would include a parametric model, with desynchronization parameter $\boldsymbol{\theta}$. In the simplest setting, the dimensionality of $\boldsymbol{\theta}$ does not depend on N; more generally, $\boldsymbol{\theta}$ may be a sequence $\boldsymbol{\theta}(n), 1 \leq n \leq N$, that exhibits temporal coherence properties, i.e., it is slowly varying, with occasional jumps. An hypothetical decoder that is informed of the values of these parameters is a *coherent decoder*; the decoder that does not is a *noncoherent decoder*. We shall be interested in constructing good noncoherent decoders, and in estimating the noncoherent decoding penalty.

Basic examples of parametric desynchronization are:

- Amplitude scaling: $y(n) = \theta x(n)$, $\theta_{\min} \leq \theta \leq \theta_{\max}$.
- Gamma correction: $y(n) = x(n)^{\theta}$, $\theta_{\min} \leq \theta \leq \theta_{\max}$.
- Temporal shifts: $y(n) = x(n - \theta)$ for integer shift θ. If θ is not an integer,

$$y(n) = \sum_i h_i(\theta) x(n - i) \qquad (1)$$

is a resampled version of the shifted, interpolated signal $\mathbf{x}$, where $h_i(\theta)$ are the taps of the interpolation filter (would be a sinc for bandlimited interpolation). If θ is a constant, (1) is a particular linear time-invariant filter. If θ varies slowly over time (as is the case with warping attacks), (1) is a linear time-variant filter.

A variety of methods have been devised to resist simple types of desynchronization attacks [1]—[10]. These include the idea of invariant domains and use of pilots or training sequences, which convey information about $\boldsymbol{\theta}$ (but not about m) to the receiver [3, 4, 5]. A theoretically superior approach is to design a code that lends itself to resynchronization, without wasting resources communicating training sequences that are not information-bearing [20, 6, 7]. As an example of this approach, some of the best results to date for the blind embedding problem have been obtained by Balado *et al.* [9]. They explore the use of the EM algorithm for simultaneously decoding messages and estimating scale parameter or delay parameters. In more recent work [10], they explored the use of phased locked loops as an alternative to the EM algorithm. Two disturbing facts about their solution are that (1) very poor performance is obtained when the desynchronization is moderate or large, and (2) the parameter estimates are not consistent, i.e., the estimation errors do not tend to 0 as the host signal length N increases. Our recent work [11, 12] has proved the existence of universal decoders for such problems, i.e., noncoherent decoders whose error exponents are *identical* to those of the corresponding coherent decoder (which knows $\boldsymbol{\theta}$). This theoretical result suggests that the decoding performance of [9, 10] can in principle be substantially improved.

3. GRAPHICAL MODELS

A particularly exciting opportunity in the watermark decoding problem is the possibility to combine the Bayesian paradigm for optimal decision making with inference techniques on graphical models [14, 15, 16]. For instance, classical Kalman filtering (or Kalman smoothing) may be interpreted as an instance of probabilistic inference in a special Gaussian graphical model. The use of Bayesian recursive filters in lieu of Kalman filters is a natural extension to this technique to nonlinear state-space models.

The opportunity of graphical models in the context of watermark decoding consists of a way to embed a message with some redundancy in a host which can exhibit long range dependencies among the signal components. The final estimation can then be organized in such a way that the Bayesian estimator and an estimator for the data redundancy iteratively solve the probabilistic inference problem. This iterative approach to estimating data in noisy environments has been very successful in data transmission and is dubbed the "turbo" principle in a communication setup. In our context

we want to fully exploit the power of this approach even in hostile and very difficult environments as would be constituted by an active attack on the decoding scheme. As such, our approach is motivated by work on the probability propagation (sum-product) algorithm for iteratively decoding error-correcting codes, such as "turbo codes" [15]. Until recently, optimal decoding even on Gaussian channels was thought to be intractable. However, it turns out that probability propagation in a graphical model describing the code solves the problem for practical purposes.

The power of this approach is even more apparent in two-dimensional data sets as they naturally appear in watermarking of images or video sequences. In this case a graphical model may be used to estimate a distortion or alteration in the properties of the host data. Together with a powerful, interleaved code that protects the embedded data we obtain an efficient scheme for data embedding. Moreover such a scheme is computationally feasible due to its inherent divide-and-conquer philosophy. This approach has revolutionized much of communications in the last few years and we believe that it holds the potential to give similarly significant and practical improvements for the watermark decoding problem.

The "curse of dimensionality" is a problem that is efficiently addressed in graphical models. In fact, one might argue that graphical models were specifically invented to cope with inference problems in high dimensional setups. The essential trick is to find a decomposition of the posterior probability density function such that estimation and hypothesis testing has a tractable structure. The generic problem in our problem setup would be one where the adversary has K possible transformations (the first one being time warping, possibly using a multiscale representation for the warping process; the second one might be an amplitude modulation, again using a multiscale representation for the envelope; etc.) The key to coping with the dimensionality of such a model is to find (or model) a factorization of the probability density as is e.g. done in factor graphs [15, 16]. Once this is done, powerful inference algorithms such a the sum-product algorithm can effectively construct excellent approximations to the global objective function [15, 16].

4. AMPLITUDE MODULATION

We have developed a scheme that uses graphical models for the problem of recovering a watermark under an amplitude modulation attack. The complete probabilistic model is as follows. We seek the MAP estimate of m given $\mathbf{y}$.

$$p(\mathbf{s}) = \prod_{i\sim j} \psi_{ij}(s_i, s_j) \quad (2)$$

$$\mathbf{c} = \mathbf{c}(m) \quad (3)$$

$$x_i = F(s_i, c_i) \quad (4)$$

$$p(\mathbf{w}) = \prod_i p_W(w_i) \quad (5)$$

$$y_i = \theta_i (x_i + w_i) \quad (6)$$

$$p(\boldsymbol{\theta}) = \prod_{i\sim j} \phi_{ij}(\theta_i, \theta_j) \quad (7)$$

In (2) and (7), $\mathbf{s}$ and θ are modeled as MRFs with second-order cliques and potential functions $\psi_{ij}(s_i, s_j)$ and $\phi_{ij}(\theta_i, \theta_j)$, respectively. The products are over all pairs of neighbors $i \sim j$. The alphabet for the codeword symbols in (3) is $\{\pm\frac{\Delta}{4}\}$. The function F in (4) is the scalar QIM embedding function

$$F(s, c) = Q(\alpha s - c) + (1 - \alpha)s + c$$

where $\alpha \in (0, 1]$ is the distortion-compensation (Costa) parameter. The attacker adds white noise $\mathbf{w}$ from (5) to $\mathbf{x}$ and applies amplitude modulation to the sum, as described by (6). The resulting data are available to the decoder.

A factor graph modeling the probabilistic model described above is shown in Fig. 3. All variable and factor nodes are circled and shaded respectively. We attempt to compute the posterior distribution of m given $\mathbf{y}$, by applying the sum-product algorithm to the above (loopy) factor graph. The algorithm is initiated with all messages set to unity and updated thereafter according to the following variable-to-factor ($v \to F$) and factor-to-variable ($F \to v$) node message update equations [15]:

$$\mu_{v\to F}(v) = \prod_{G\in \text{ne}(v)\setminus F} \mu_{G\to v}(v) \quad (8)$$

$$\mu_{F\to v}(v) = \int_{\mathbf{u}} f(v, \mathbf{u}) \prod_{i=1}^{K} \mu_{u_i\to F}(u_i) \mathrm{d}\mathbf{u} \quad (9)$$

where, $\text{ne}(v)$ is the set of neighbors of v, $f(\cdot)$ is the function associated with factor node F and $\mathbf{u} = [u_1, u_2, \ldots u_K]$ is the vector containing all the neighbors of F, excluding v.

We compare the performance of the proposed decoder with a coherent decoder (that knows $\boldsymbol{\theta}$). We chose a white Gaussian host with each component having mean zero and variance $\sigma_s^2 = 50^2$. Note that any correlation present in the source (which normally is the case) can only help the decoder to perform better. The code is a simple block repetition code with block length equal to 8, hence the embedding rate is $\frac{1}{8}$ bits per sample. For embedding we used $\Delta = 17.32$, resulting in an embedding distortion $D_1 = \frac{\Delta^2}{12} = 25$. The noise $\mathbf{w}$ is Gaussian with mean zero and variance $D_2 = D_1$. For

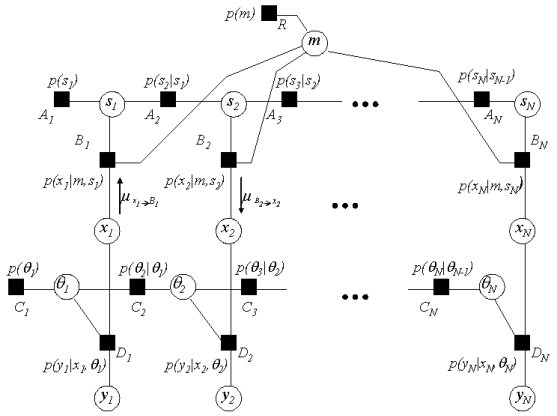

Fig. 3. Factor graph representation for the probabilistic model in (2) – (7)

the amplitude field $\boldsymbol{\theta}$, we explore three possibilities: (i) $\boldsymbol{\theta}$ has i.i.d. components, each drawn from a $\mathcal{N}(1, 0.01)$ distribution. This poses the toughest scenario; (ii) Components of $\boldsymbol{\theta}$ are not independent but have nearest neighbour dependencies through (7); the amplitude field is Gauss-Markov with mean 1, variance 0.01, and a normalized correlation coefficient ρ equal to 0.9 and (iii) All components of $\boldsymbol{\theta}$ are same, representing amplitude scaling as a special case of amplitude modulation.

We first evaluated the performance of the coherent decoder that knows the amplitude field $\boldsymbol{\theta}$ and can therefore invert the AM operation. The estimated error probability, evaluated from 10^4 Monte-Carlo simulations, was approximately $P_e = 2.7 \times 10^{-3}$ in all cases.

Next we evaluated the performance of our noncoherent decoder. For each simulation, the belief propagation step must be repeated until the estimates of m were stabilized. Again by performing 10^4 Monte-Carlo simulations, we obtained $P_e = 0.0231$, $P_e = 5.6 \times 10^{-3}$ and $P_e = 4.1 \times 10^{-3}$ for cases (i), (ii) and (iii) respectively.

An extension of the current results to two dimensions is quite straightforward. A more interesting extension would be to test the feasibility of reliable watermark detection, in the presence of desynchronization attacks, on image models which are less synthetic. We will explore that in the future.

5. REFERENCES

[1] M. Kutter, "Watermarking Resisting to Translation, Rotation and Scaling," *Proc. SPIE*, Boston, Vol. 3528, pp. 423—431, 1998.

[2] J. J. K. O'Ruanaidh and T. Pun, "Rotation, Scale and Translation Invariant Spread Spectrum Digital Image Watermarking," *Signal Processing*, Vol. 66, No. 3, pp. 303—317, 1998.

[3] S. Pereira and T. Pun, "Robust Template Matching for Affine Resistant Image Watermarks," *IEEE Trans. on Image Processing*, Vol. 9, No. 6, pp. 1123–1129, June 2000.

[4] M. Álvarez-Rodríguez and F. Pérez-González, "Analysis of Pilot-Based Synchronization Algorithms for Watermarking of Still Images," *Signal Processing: Image Communication*, Vol. 17, pp. 611-633, Sep. 2002.

[5] P. Moulin and A. Ivanović, "The Fisher Information Game for Optimal Design of Synchronization Patterns in Blind Watermarking," *Proc. IEEE Int. Conf. on Image Processing*, pp. II. 550—553, Thessaloniki, Greece, Oct. 2001.

[6] P. Moulin, "Embedded-Signal Design for Channel Parameter Estimation. Part I: Linear Embedding," *Proc. IEEE Statistical Signal Processing Workshop*, pp. 38—41, St Louis, MO, Sep. 2003.

[7] P. Moulin, "Embedded-Signal Design for Channel Parameter Estimation. Part II: Quantization Embedding," *Proc. IEEE Statistical Signal Processing Workshop*, pp. 42–45, St Louis, MO, Sep. 2003.

[8] P. Moulin, A. Briassouli and H. Malvar, "Detection-Theoretic Analysis of Desynchronization Attacks in Watermarking," *Proc. 14th Int. Conf. on Digital Signal Proc.*, pp. I. 77—84, Santorini, Greece, July 2002.

[9] F. Balado, K. M. Whelan, G. C. M. Silvestre, and N. J. Hurley, "Joint Iterative Decoding and Estimation for Side-Informed Data Hiding" *IEEE Trans. on Signal Processing*, 3rd supplement on Secure Media, Vol. 53, No. 10, pp. 4006—4019, Oct. 2005.

[10] K. M. Whelan, F. Balado, G. C. M. Silvestre, and N. J. Hurley, "PLL-Based Synchronization of Dither Modulation Data Hiding," *Proc. ICASSP*, Toulouse, France, May 2006.

[11] P. Moulin, "Universal Decoding of Watermarks Under Geometric Attacks" *Proc. IEEE Int. Symp. on Information Theory*, Seattle, WA, July 2006.

[12] P. Moulin, "On the Optimal Structure of Watermark Decoders Under Geometric Attacks," *Proc. IEEE Int. Conf. on Image Proc.*, Atlanta, GA, Oct. 2006.

[13] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, UK, 1996.

[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[15] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, MIT Press, Cambridge, MA, 1998.

[16] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms*, 2001.

[17] J. Yedidia, W. T. Freeman, and Y. Weiss "Generalized Belief Propagation" *Advances in Neural Information Processing Systems* vol. 13, 2001.

[18] M. I. Jordan, "Graphical Models," *Statistical Science*, Special issue on Bayesian statistics, Vol. 19, pp. 140—155, 2004.

[19] P. Moulin and R. Koetter, "Data-Hiding Codes," *Proc. IEEE*, Vol. 93, No. 12, pp. 2083–2127, Dec. 2005.

[20] A. Lapidoth and P. Narayan, "Reliable Communication Under Channel Uncertainty," *IEEE Trans. Information Theory*, Vol. 44, No. 6, pp. 2148—2177, Oct. 1998.

[21] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Information Theory*, Vol. 49, No. 3, pp. 563–593, March 2003.