Attacks on Kuribayashi's Fingerprinting Scheme

Hans Georg Schaathun

***Abstract*—The main threat against fingerprinting systems is collusion attacks. The attack most commonly assumed in the literature is a combination of averaging the collusion fingerprints and additive noise. In this correspondence we demonstrate that the recently proposed fingerprinting scheme of Kuribayashi's is very vulnerable to certain non-linear collusion attacks.**

Digital fingerprinting (FP) is used to trace unauthorised copies and identify the copyright violator. This is done by embedding a fingerprint in each copy distributed, identifying the authorised user. One of the most recently proposed FP schemes is due to Kuribayashi [1], [2]. The FP code and basic decoding algorithm is described in [1]. The second paper [2] refines the decoding algorithm, iteratively removing interference from fingerprints already detected, in order to identify more users. This refinement is generic, but only described and evaluated using the one scheme.

Kuribayashi evaluated his FP scheme against the averaging attack combined with JPEG compression. Citing [3], he claimed that a *number of nonlinear collusions such as interleaving attacks can be well approximated by averaging collusion plus additive noise*. We have been unable to find the foundation for this claim, and we shall show that the system is extremely vulnerable to the Moderated Minority Extreme (MMX) attack [4] and the uniform attack [5].

I. The Fingerprinting Scheme

Kuribayashi's FP scheme uses additive watermarking. Given user i with assigned fingerprint $\mathbf{w}_i$ of length ℓ and a host $\mathbf{x}$, the user's copy is $\mathbf{y}_i = \mathbf{x} + \mathbf{w}_i$. Both $\mathbf{x}$ and $\mathbf{w}_i$ are continous-valued signals. The scheme is hierarchical, so each user is identified by a group index g and a user index u, where g and u are natural numbers. Two approaches are suggested for combining group and user information, but we will only discuss the one recommended by Kuribayashi, superimposing the two watermarks. Each component is the element-wise product of two elements: the i-th basis vector $\mathbf{z}_i$ of the DCT transform and a member $\mathbf{p}_i$ of a family of Gold sequences [6]. Thus the fingerprint for user (g, u) is the floating-point signal

$$\mathbf{w}_{g,u} = \mathbf{p}_s \otimes \mathbf{z}_g \cdot \beta_\mathrm{g} + \mathbf{p}_g \otimes \mathbf{z}_u \cdot \beta_\mathrm{u},$$

where $\otimes$ denotes element-wise product, s is a secret key, β_g and β_u are scaling factors, and $\mathbf{z}_i = \mathrm{DCT}(\mathbf{i}_i)$ where $\mathbf{i}_i$ is the vector with 1 in the ith position and 0 elsewhere, and DCT is a DCT transform.

Only a subset of samples from the host signal $\mathbf{x}$ are used for embedding, and the samples are randomly ordered. The subset and the ordering are determined by a pseudo-random generator seeded by a secret key. Kuribayashi did not discuss secrecy of the indices u and g, but we assume that the user does not know their indices.

Intercepting a copy $\mathbf{y}'$, we subtract the host to get the received watermark $\mathbf{w}'$. The group detector sequence is calculated as

$$\mathbf{d} = (d_1, d_2, \ldots, d_\ell) = \mathrm{DCT}^{-1}(\mathbf{p}_s \otimes \mathbf{w}').$$

High values for d_g indicate suspicious groups g. Similarly, the user detector sequence is calculated for each suspicious group g as $\mathbf{d}' = \mathrm{DCT}^{-1}(\mathbf{p}_g \otimes \mathbf{w}')$. A group or user is deemed suspicious if the corresponding element in the detector sequence is above the threshold

$$T = \sqrt{2\sigma^2} \cdot \mathrm{erfc}^{-1}(2\epsilon),$$

where σ^2 is the variance of the detector sequence, ϵ is a target false positive probability, and

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt.$$

Interference removal [2] can be used to subtract the watermarks of detected users and iterating to to identify further pirates. Due to page constraints, the reader is referred to the original paper for the details. The implementation used for this paper is available [7].

II. The Attacks

Let P be a set of fingerprints of a pirate collusion, and write $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ for any $\mathbf{x} \in P$. Well-known attacks include:

Average:	$\bar{x}_i = \frac{1}{N} \sum_{\mathbf{x} \in P} x_i.$
Minimum:	$x_i^{\min} = \min_{\mathbf{x} \in P} x_i.$
Maximum:	$x_i^{\max} = \max_{\mathbf{x} \in P} x_i.$
Midpoint:	$x_i^{\mathrm{mid}} = (x_i^{\min} + x_i^{\max})/2.$

None of these attacks are particularly effective by themselves, but we use them to define the following two attacks.

Definition 1 (Moderated Minority Extreme Attack) *Let $D_i = \bar{x}_i - x_i^{\mathrm{mid}}$. The MMX attack for a given threshold θ outputs the hybrid signal $\mathbf{x}^{(\theta)}$, where*

$$x_i^{(\theta)} = \begin{cases} x_i^{\min} \text{ if } D_i \geq \theta, \\ \bar{x}_i \text{ if } \theta > D_i > -\theta, \\ x_i^{\max} \text{ if } D_i \leq -\theta, \end{cases}$$

The rationale for this attack is to use, for each sample, a value which incriminates as few pirates as possible. If no value is particularly good in that respect, we use the average to minimise overall distortion.

Definition 2 (Uniform Attack) *The uniform attack with scaling factor s ($0 \leq s \leq 1$) constructs the hybrid signal $\mathbf{x}^{U(s)}$, by drawing $x_i^{U(s)}$ uniformly at random from the range $x_i^{\mathrm{mid}} \pm s \cdot d$ where $d = (x_i^{\max} - x_i^{\mathrm{mid}})/2$.*

The parameters s and θ should be tuned to get a competitive level of distortion, while preventing detection. Note that MMX degenerates to an averaging attack when θ is sufficiently large. Similarly, the uniform attack degenerates to midpoint for $s = 0$. Details can be sought in [5]. Only the scaling factor s is new in this paper.

Remark 1 *The uniform and MMX attacks can be applied to the entire watermarked files, without knowledge of the embedding region. For any sample unused by the embedding, every pirate will see the same value x, and either attack will return this value x unmodified.*

Remark 2 *The uniform and MMX attacks can be equivalently applied either to the fingerprints $\mathbf{w}_u$ or to the fingerprinted signals $\mathbf{y}_u$. We have $\mathbf{y}^{(\theta)} = \mathbf{x} + \mathbf{w}^{(\theta)}$ and $\mathbf{y}^{U(s)} = \mathbf{x} + \mathbf{w}^{U(s)}$.*

III. The Experiments

Following Kuribayashi [2], we fix $\beta_\mathrm{g} = 400$ and $\beta_\mathrm{u} = 600$, with 2^{20} users. The fingerprint length is $\ell = 2048$, and we consider a coalition size of $c = 10$ pirates, a number which gives virtually complete detection in all experiments reported in [1], [2]. Three attacks are considered: the uniform attack, the MMX attack, and the averaging attack with JPEG compression. Justified by Remarks 1–2, the attacks are applied to the FP signals $\mathbf{w}$, without actually

The author is with Ålesund University College, Pb. 1517, N-6025 Ålesund, Norway. Email: ⟨georg@schaathun.net⟩

	User		Group		Power	
Attack	TP	FP	TP	FP	Received	Embed'd
JPEG QF35	10.00	0.14	9.95	1.90	55606	520037
JPEG QF100	10.00	0.01	9.96	2.00	55681	520091
MMX (1.9)	0.00	0.01	0.05	15.23	853109	520090
MMX (3.55)	0.00	0.00	0.06	14.11	511907	520069
MMX (6.0)	0.00	0.00	0.90	6.37	201755	520067
UNI (1.0)	0.00	0.00	1.72	2.13	435581	520028
UNI (0.5)	0.73	0.00	7.82	3.25	140610	519952
UNI (0.1)	9.68	0.10	9.95	12.27	46616	520086

(a) Distinct groups

	User		Group		Power	
Attack	TP	FP	TP	FP	Received	Embed'd
JPEG QF35	7.88	0.00	1.00	2.09	173138	520111
JPEG QF100	7.85	0.00	1.00	2.10	172973	520002
MMX (1.9)	0.00	0.00	1.00	1.94	519405	519801
MMX (3.55)	0.00	0.00	1.00	2.03	358565	519958
MMX (6.0)	0.00	0.00	1.00	2.18	235798	519981
UNI (1.0)	0.00	0.00	1.00	2.08	374433	520140
UNI (0.5)	0.00	0.00	1.00	2.16	218292	520040
UNI (0.1)	0.00	0.00	1.00	2.49	168140	519955

(b) Single group

Table I: True positive (TP) and false positive (FP) rates for the original FP scheme.

	User		Group		Power	
Attack	TP	FP	TP	FP	Received	Embed'd
JPEG QF35	10.00	0.00	9.95	1.96	55601	520014
JPEG QF100	10.00	0.00	9.94	1.91	55714	519889
MMX (1.9)	0.00	0.04	0.07	15.18	851763	519887
MMX (3.55)	0.00	0.04	0.06	14.21	512779	519976
MMX (6.0)	0.00	0.01	0.90	6.33	201866	520018
UNI (1.0)	0.00	0.00	1.72	2.02	435068	519859
UNI (0.5)	1.94	0.09	7.71	3.20	140836	519985
UNI (0.1)	9.98	0.15	9.95	12.38	46529	520107

(a) Distinct groups

	User		Group		Power	
Attack	TP	FP	TP	FP	Received	Embed'd
JPEG QF35	9.96	0.00	1.00	2.10	172936	520036
JPEG QF100	9.94	0.00	1.00	2.09	173199	520175
MMX (1.9)	0.00	0.02	1.00	2.03	519046	519923
MMX (3.55)	0.00	0.00	1.00	2.05	359167	519734
MMX (6.0)	0.00	0.02	1.00	2.14	235642	520158
UNI (1.0)	0.00	0.00	1.00	2.23	374588	519960
UNI (0.5)	0.27	0.02	1.00	2.17	218217	519967
UNI (0.1)	9.94	1.61	1.00	2.40	168203	519981

(b) Single group

Table II: True positive (TP) and false positive (FP) rates for interference removal.

embedding. For the JPEG attack, the average fingerprint is embedded in the frequency domain of a Lena image as prescribed in [2]. The resulting image is then compressed and decompressed with JPEG, before reextracting the fingerprint.

We run two sets of tests, one drawing a single random group g and c random pirates from this group (single group), and another drawing c random groups and a single random pirate per group (distinct groups). Each experiment takes the average over 1000 trials.

As a measure of distortion, we use the power (squared Euclidean norm) of the embedded and the received watermark. Attacks which give received watermarks of lower power than the embedded watermark are considered reasonable. What is good enough for the authorised user ought to be good enough for the pirate.

Test 1 (Original Scheme). Experimental results for the original scheme [1] are shown in Table I. These results are consistent with [1], which only considered detection rates for distinct groups. We note that the JPEG attacks give pirate watermarks of very low power, which means that the compression noise caused by JPEG must have very limited impact on the watermark.

We show the MMX and uniform attacks with various parameter choices (θ and s). We observe that it is possible to tune the parameters to give zero detection, and simultanously less power than the embedded watermark. In some cases we get false positives, although in insufficient numbers to draw any conclusions.

Test 2 (Interference Removal). Experimental results with interference removal [2] are shown in Table II. We can see that this improves detection in some cases, but it has still zero detection against MMX and for the uniform attack with scaling factor $s = 1$.

Test 3 (Group Detectability). Finally, we studied the behaviour of the group detection statistics. We tested with ten pirates from ten distinct groups. For MMX ($\theta = 3.55$), the mean detection statistic is lower for guilty groups (-2.7) than for innocent groups (0.2). Taking the average variance over 1000 trials, we get 256 for guilty groups and 277 for innocent ones. This is compared to between 11.5 and 13 for the JPEG attacks. Innocent and guilty users are thus statistically indistinguishable, and this problem cannot easily be overcome by increasing the embedding strength or refining the detection algorithm. A similar effect can be expected at the user level.

The uniform attack does not have the same clear effect as MMX, but it does give a significantly increased variance to the detection heuristic. For instance, for $s = 1$, we get a variance of 193 for guilty groups and 208 for innocent ones. In contrast, the mean remains larger for guilty groups.

IV. Conclusions

We have demonstrated that Kuribayashi's FP scheme can be effectively attacked by either the uniform or the MMX attack, which are much more effective than the more well-known attacks combining averaging with additive noise. It is essential that these nonlinear attacks be considered in future works in the area.

It is doubtful if additive watermarking can provide FP schemes secure against large collusions. Both this and previous papers [5] show that the MMX attack give lower mean score for guilty users than for innocent ones against several known detection heuristics. Future solutions may have to take an entirely different approach.

References

[1] M. Kuribayashi, "Hierarchical spread spectrum fingerprinting scheme based on the cdma technique," *EURASIP Journal on Information Security*, vol. 2011, no. 1, p. 502782, 2011.

[2] ——, "Interference removal operation for spread spectrum fingerprinting scheme," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 403–417, 2012.

[3] H. Zhao, M. Wu, Z. J. Wang, and K. J. R. Liu, "Forensic analysis of nonlinear collusion attacks for multimedia fingerprinting," *IEEE Trans. Image Proc.*, pp. 646–661, 2005.

[4] H. G. Schaathun, "Attack analysis for He&Wu's joint watermarking/fingerprinting scheme," in *The 6th International Workshop on Digital Watermarking*, ser. Lecture Notes in Computer Science, vol. 3304, 2007.

[5] ——, "Novel attacks on spread-spectrum fingerprinting," *EURASIP J. Information Security*, vol. 2008, 2008.

[6] R. Gold, "Maximal recursive sequences with 3-valued recursive cross-correlation functions (corresp.)," *Information Theory, IEEE Transactions on*, vol. 14, no. 1, pp. 154–156, 1968.

[7] H. G. Schaathun, "The pyfp library," 2013, http://www.ifs.schaathun.net/pyfp/.

Hans Georg Schaathun was born in Bergen, Norway, in 1975. He is Cand.Mag. 1996 (Mathematics, Economics, and Informatics), Cand.Scient. 1999 (Industrial and Applied Mathematics and Informatics), and Dr.Scient. 2002 (Informatics – Coding Theory), all from the University of Bergen, Norway. He was lecturer in coding and cryptography at the University of Bergen 2002 and Post.Doc. 2003-2006. As a lecturer and senior lecturer in computer science at the University of Surrey, England 2006-2010, his research focused on multimedia security. He joined Ålesund University College in Norway in January 2011 and became professor of computing. Current research interests span machine learning, model-driven engineering, and health care technology. His book on *Machine Learning in Image Steganalysis* was published by John Wiley & sons in 2012.