COLLUDING FINGERPRINTED VIDEO USING THE GRADIENT ATTACK

Shan He [1]*, Darko Kirovski* [2] *and Min Wu* [1]

[1]Department of ECE, University of Maryland, College Park, MD 20742, USA
[2]Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA

ABSTRACT

Digital fingerprinting is an emerging tool to protect multimedia content from unauthorized distribution by embedding a unique fingerprint into each user's copy. Although several fingerprinting schemes have been proposed in related work, disproportional effort has been targeted towards identifying effective collusion attacks on fingerprinting schemes. Recent introduction of the gradient attack has refined the definition of an optimal attack and demonstrated strong effect on direct-sequence, uniformly distributed, and Gaussian spread spectrum fingerprints when applied to synthetic signals. In this paper, we apply the gradient attack on an existing well-engineered video fingerprinting scheme, refine the attack procedure, and demonstrate that the gradient attack is effective on Laplace fingerprints. Finally, we explore an improvement on fingerprint design to thwart the gradient attack. Results suggest that Laplace fingerprint should be avoided. However, we show that a signal mixed of Laplace and Gaussian fingerprints may serve as a design strategy to disable the gradient attack and force pirates into averaging as a form of adversary collusion.

Index Terms– Multimedia fingerprinting, collusion resistance, gradient attack.

1. INTRODUCTION

Significantly increased levels of multimedia piracy over the last decade have put the movie and music industry under pressure to deploy a standard anti-piracy technology. In a typical scenario that uses multimedia marking for forensic purposes, studios create a uniquely marked content copy for each individual user request. User-specific distinct watermarks are commonly denoted as fingerprints. The fingerprinted copy is securely distributed to the user who plays the content using a media player which is unmodified compared to modern media players. Certain users may chose to illegally distribute this content. To address this problem, the media studios deploy search robots in order to find content copies on the Internet. Illegally distributed content is retrieved and based upon the known user database as well as the original clip, media studios use forensic analysis tools to identify the pirates.

A major problem for fingerprinting systems is the collusion attack. To launch such an attack, an adversarial group of malicious users colludes their copies in order to create a copy which is statistically clean of any fingerprint traces (e.g., the original) or a copy that incriminates another innocent user. Various collusion attacks have been studied in literature [1]. Results show that a number of non-linear collusion attacks based on order statistics can be well modelled by collusion via averaging plus an additive noise under Gaussian fingerprint construction. Recent introduction of the gradient attack [2] showed that a new class of attacks could be built by deploying two instead of one estimate of the original signal. If one estimate is provably and possible probabilistically better than the other, the gradient attack identifies an attack vector that is key to reducing the expected correlation between any fingerprint and the attacked copy to zero. To date, the gradient attack has been examined on direct-sequence, uniform, and Gaussian fingerprints and synthetic signals. In this paper, we examine the effectiveness of the gradient attack on a video fingerprinting scheme proposed in [3, 4], and found that it is very effective. We also found that, due to the underlying Laplace fingerprints, the observed order of the estimates used for the gradient attack is exactly opposite to what has been shown on Gaussian fingerprints. Based on this observation, we propose a fingerprint construction combining these two distributions that improves the robustness to gradient attack.

2. A VIDEO FINGERPRINTING SCHEME

The video fingerprinting scheme that is examined in this paper is proposed in [3, 4]. It marks the content by designing a content-adaptive watermark signal via solving an optimization problem. The fingerprint is embedded into the DC band of the DWT (Discrete Wavelet Transform) domain. The algorithm packs these coefficients into a 3D cuboid $\mathbf{x}$, where the third dimension represents the frame index. Based upon a unique user key, the fingerprint embedding algorithm selects pseudo-randomly, in terms of positions and sizes, a collection of sub-cuboids $\mathbb{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\} \subset \mathbf{x}$ that may

Contact information: shanhe@eng.umd.edu

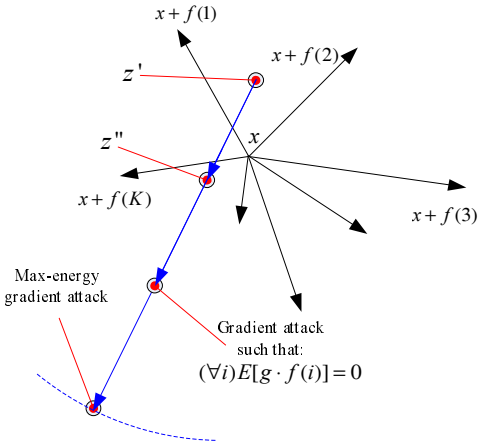

Fig. 1. Illustration of the gradient attack. The dotted line depicts the bound of imperceptive noise.

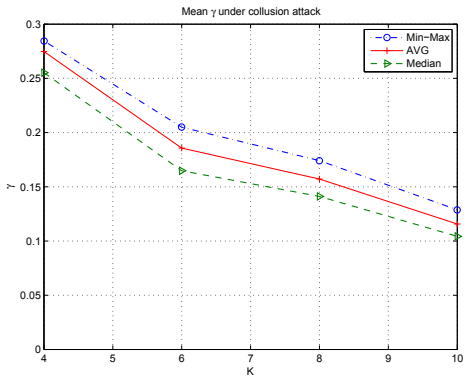

Fig. 2. Colluding video fingerprints: detection statistic γ vs. number of colluders K under averaging, min-max and median attacks.

overlap. Then, the coefficients in each sub-cuboid $\mathbf{p}_j \in \mathbb{P}$ are weighted using a smooth weighting cuboid $\mathbf{u}_j$. The weighting cuboids are generated pseudo-randomly using a user-specific secret key. Finally, the algorithm computes the feature vector $\mathbf{g} = [g_1, ..., g_n]$ of the host signal with g_j as the mean value of $\mathbf{p}_j \cdot \mathbf{u}_j$. The embedding process quantizes the feature using a private quantizer $q(\mathbf{g})$ to get the feature vector $\hat{\mathbf{g}}$ of the fingerprinted signal. The equivalent fingerprint in the feature domain would be $\hat{\mathbf{g}} - \mathbf{g}$, which is spread among the pixels of the containing cuboid in such a way that the introduced distortion is minimized.

Given a received video signal $\mathbf{z}$, the detector extracts the feature vector $\mathbf{g}_z$ in the same way as the embedding process using a suspect user's key. The extracted test fingerprint in the feature domain would be $\mathbf{g}_z - \mathbf{g}$. It then employs a correlation based detection to identify the existence of a watermark as follows:

$$\gamma = \frac{(\mathbf{g}_z - \mathbf{g}) \cdot (\hat{\mathbf{g}} - \mathbf{g})}{||\hat{\mathbf{g}} - \mathbf{g}||^2} \lessgtr T, \qquad (1)$$

If γ is greater than a certain threshold T, the detector concludes that $\mathbf{z}$ is marked with the fingerprint generated using the suspect user key; otherwise, no fingerprint is detected.

3. THE GRADIENT ATTACK

Traditional collusion attacks considered in literatures estimate the original content in order to remove the traces of each participant [1]. As estimates get closer to the original, the detection statistic for each participant, usually the correlation of the colluded signal with each user's fingerprint, becomes lower. In the gradient attack, instead of estimating the original, attackers try to find a spot in the perceptively similar neighborhood of the original, that is expected to be orthogonal to the fingerprints of all colluders. This is indeed the paramount objective in attacking fingerprinting systems.

In a fingerprinting system, user i's copy $\mathbf{y}^{(i)}$ is generated as $\mathbf{y}^{(i)} = \mathbf{x} + \mathbf{w}^{(i)}$ with $\mathbf{x}$ being the original signal and $\mathbf{w}^{(i)}$ being user i's fingerprint sequence. The marked copies for users correspond to different points in the multi-dimension space as shown in Fig. 1. To launch a gradient attack, a collusion group $\{1, ..., K\}$ designs two attacks on their fingerprinted copies, with one yielding always a better expected estimate $\mathbf{z}''$ of the original content than the other one $\mathbf{z}'$. Signals $\mathbf{z}'$ and $\mathbf{z}''$ can be obtained via averaging or median attacks as examined in [1]. These two attacks identify a direction $\mathbf{z}'' - \mathbf{z}'$ in the space opposite to the direction of all fingerprints in the collusion group. The adversary can move the estimate $\mathbf{z}''$ along this direction to a point $\mathbf{z}$ in which every colluder's fingerprint is mostly removed. That is, $\mathrm{E}[d(\mathbf{z} - \mathbf{x}, \mathbf{w}^{(i)})] \approx 0$ for each participant, where $d(\mathbf{z} - \mathbf{x}, \mathbf{w}^{(i)})$ is the corresponding detection statistic for user i. Note that colluders also have to maintain the visual quality of the attack signal, which imposes a maximum energy constraint on the gradient attack, shown as an arc in Fig. 1.

Mathematically, the gradient attack is defined as: $\mathbf{z} = \mathbf{z}'' - \beta(\mathbf{z}' - \mathbf{z}'')$, where β is such that $\mathrm{E}[c(\mathbf{z} - \mathbf{x}, \mathbf{w}^{(i)})] \approx 0$ for each user i in the collusion group $\mathcal{K}$ and the pirated vector is perceptually close to the original so that $||\mathbf{z} - \mathbf{x}|| \leq \delta\sqrt{N}$. Hence, the value of β that can be applied in the gradient attack satisfies:

$$\beta \leq \frac{-\sigma_m^2 + c + \sqrt{(\sigma_m^2 - c)^2 - (\sigma_m^2 - \delta^2)(\sigma_m^2 + \sigma_a^2 - 2c)}}{\sigma_m^2 + \sigma_a^2 - 2c}$$

where σ_a^2 and σ_m^2 are the variances of $\mathbf{z}'$ and $\mathbf{z}''$ respectively, and c is the covariance between $\mathbf{z}'$ and $\mathbf{z}''$.

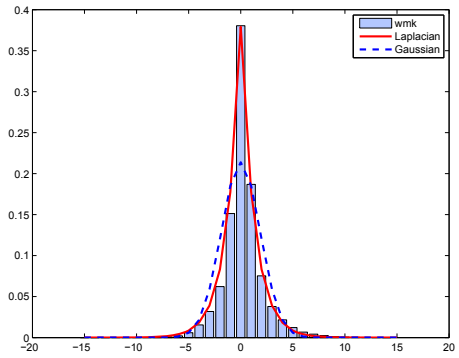

Fig. 3. Histogram of equivalent fingerprints of the video fingerprinting scheme along with the Gaussian and Laplace approximation.

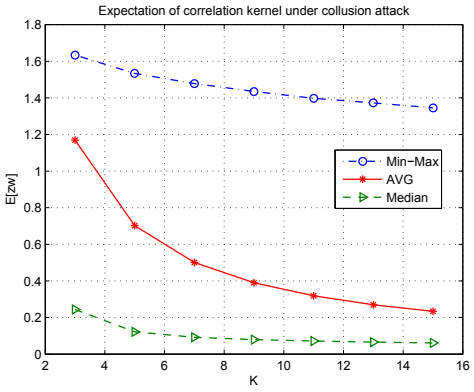

Fig. 4. Analytical results on Laplace fingerprints under averaging, min-max and median attacks.

3.1. Estimating Fingerprints

To apply the gradient attack onto the target video fingerprints, we first find two estimates of the original content in which one is always better than the other. In this paper, we start with three attacks in the pixel domain, namely, averaging, min-max and median attack. For a colluder group S_c, the averaging attack generates $\mathbf{z}_{AVG} = \frac{1}{K}\sum_{i\in S_c}\mathbf{y}^{(i)}$; the min-max attack generates $\mathbf{z}_{MM} = \frac{1}{2}(\min_{i\in S_c}\mathbf{y}^{(i)} + \max_{i\in S_c}\mathbf{y}^{(i)})$; and the median attack generates $\mathbf{z}_{Med} = \text{median}_{i\in S_c}\mathbf{y}^{(i)}$. The means of the detection statistics for colluders under these three attacks are shown in Fig. 2. We can see that the median attack yields an estimate that has the smallest detection statistic, while min-max has the largest detection statistic for all examined collusion sizes. Thus, we have the order of Min-Max $>$ Averaging $>$ Median in terms of the detection statistic, which is exactly opposite to the case with Gaussian fingerprints reported in [5]. This inspired us to examine the underlying distribution of the fingerprints in the examined video fingerprinting system.

3.2. Laplace Approximation

We extract the equivalent fingerprint of the video fingerprinting system in the pixel domain and plot the histogram in Fig.3 along with the approximation using Gaussian and Laplace distribution. We see that the equivalent fingerprint can be well approximated by Laplace distribution rather than Gaussian distribution. Based on this observation, we model the fingerprints to follow a zero-mean bounded Laplacian:

$$f(x) = \begin{cases} \frac{\exp(-|x|/b)}{2b[F(\delta)-F(-\delta)]}, & |x| \le \delta, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $F()$ is the $c.d.f.$ of an unbounded zero-mean Laplace distribution with parameter b. We denote this distribution as $\bar{L}(b,\delta)$. The variance of the bounded Laplace distribution is:

$$\sigma_w^2 = \frac{2b^2 - [(\delta+b)^2+b^2]\exp(-\delta/b)}{1-\exp(-\delta/b)}. \tag{3}$$

With this model, we are able to derive the expected correlation kernel $E[\mathbf{z}\cdot\mathbf{w}_i]$ as an indicator of the detection statistic γ. Then we numerically evaluate $E[\mathbf{z}\cdot\mathbf{w}_i]$ for three attacks, and show the results in Fig. 4. We can see that the analytical results of the correlation kernel for three attacks are consistent with the simulation results, which further demonstrates that the underlying fingerprints are Laplacian.

3.3. Experimental Results

As seen from Fig. 2, the Median attack generates a copy that always has a smaller detection statistic than the Min-Max attack. Thus, we choose $\mathbf{z}_{MM}$ as $\mathbf{z}'$ and $\mathbf{z}_{Med}$ as $\mathbf{z}''$ to apply the gradient attack on the target video fingerprinting scheme. For simplicity, β is set at 3 to keep the visual quality acceptable. After the gradient attack, the average PSNR of the attack video frames is about 37dB. Fig. 5 shows the detection results after the gradient attack. We can see that the expected detection statistic for colluders is significantly reduced and with the current visual quality of the colluded content, only $K = 8$ colluders are able to defeat the system regardless of object size. The results suggest that Laplacian fingerprints are vulnerable to the gradient attack and should be avoided in fingerprint design.

4. IMPROVED FINGERPRINT CONSTRUCTION

In this section, we discuss how to design fingerprints to improve the collusion resistance against the gradient attack. We observe that the order of the detection statistics under

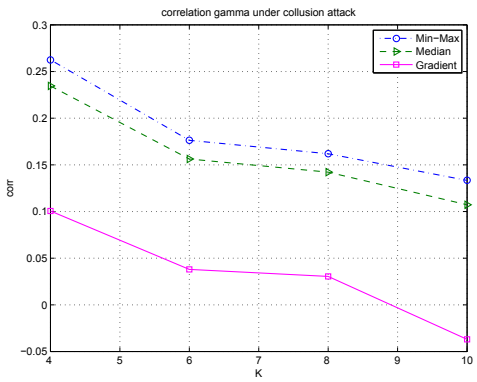

Fig. 5. Detection statistic γ of the video fingerprinting [3] vs. number of colluders K under gradient attack.

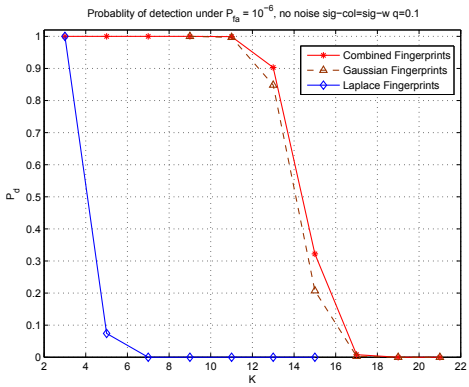

Fig. 6. Experimental results on combined fingerprints under the gradient attack.

the three considered attacks are exactly opposite for Gaussian and Laplace fingerprints. Note that the effectiveness of the gradient attack comes from the gap between the two estimates $\mathbf{z}'$ and $\mathbf{z}''$ or the direction pointed by $\mathbf{z}' - \mathbf{z}''$. If we can reduce or eliminate the gap, the gradient attack will be less effective. To achieve this, we propose to combine Gaussian and Laplacian fingerprints. As a preliminary exploration, we linearly combine fingerprints generated from these two distributions to get each user's fingerprint. For each user i, we generate two sequences: $\mathbf{w}_1^{(i)}$ following the Laplace distribution and $\mathbf{w}_2^{(i)}$ following the Gaussian distribution. The final fingerprint sequence $\mathbf{w}^{(i)}$ for user i is obtained as:

$$\mathbf{w}^{(i)} = \sqrt{q}\mathbf{w}_1^{(i)} + \sqrt{1-q}\mathbf{w}_2^{(i)}. \quad (4)$$

where q is a parameter to adjust the weight for sequences $\mathbf{w}_1^{(i)}$ and $\mathbf{w}_2^{(i)}$.

We examine the performance of the combined fingerprinting through simulation. We choose $\delta = 7$ and $\sigma_w^2 = 4.7$ for both distributions. Combination parameter q is set to be 0.1. With this parameter settings, we choose $\mathbf{z}''$ to be the signal by the Min-Max attack and $\mathbf{z}'$ the signal obtained by the median attack. β is chosen such that the distortion introduced by the collusion attack is comparable to that by fingerprint embedding. We measure the probability of catching all colluders P_d given the probability of a false alarm to be no higher than 10^{-6}. Fig. 6 shows the results for Laplace, Gaussian and combined fingerprints. Since we measure the probability of catching all the colluders, the curves drop to zero quickly as K increases. From the results, we can see that the combined fingerprints offer significant improvement in P_d over Laplace fingerprints and around 10% increase over Gaussian fingerprints. Although the improvement over Gaussian fingerprints is limited, the results on the simple linear summation suggest the potential of combining two distributions for fingerprint construction. We believe other combination approaches, such as interleaving sequences from two distributions, could lead to better performance. How to combine and where to put sequence from which distribution to achieve the highest collusion resistance are open problems, which we plan to explore in our future work.

5. CONCLUSION

In this paper, we have examined the effectiveness of the gradient attack on an existing video fingerprinting scheme. The results show that the video fingerprinting scheme is vulnerable to the gradient attack, whereby as few as 8 colluders are able to defeat the system regardless of object size. The vulnerability comes from the Laplace distribution of the fingerprints, which suggests that one should avoid using Laplace distribution in fingerprint design. We explored a countermeasure for fingerprint construction against the gradient attack by combing Gaussian and Laplace distributions. Results suggest great potential for the new breed of mixed-distribution fingerprints.

6. REFERENCES

[1] H. V. Zhao, M. Wu, Z. J. Wang and K. J. R. Liu, "Forensic Analysis of Nonlinear Collusion Attacks for Multimedia Fingerprinting," *IEEE Trans. on Image Proc.*, vol. 14, no. 5, pp. 646–661, May 2005.

[2] D. Schonberg and D. Kirovski. "Fingerprinting and Forensic Analysis of Multimedia," ACM *Multimedia*, 2004.

[3] O. Harmanci and M.K. Mihcak. "Complexity-Regularized Video Watermarking via Quantization of Pseudo-Random Semi-Global Linear Statistics". *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2005.

[4] O. Harmanci and M.K. Mihcak. "Motion Picture Watermarking Via Quantization of Pseudo-Random Linear Statistics". *Visual Communications and Image Processing Conference*, 2005.

[5] D. Kirovski and M.K. Mihcak, "Bounded Gaussian Fingerprints and the Gradient Collusion Attack," *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. 1037–1040, 2005.