# RISK-DISTORTION ANALYSIS FOR VIDEO COLLUSION ATTACK

*Yan Chen, W. Sabrina Lin, and K. J. Ray Liu*

## ABSTRACT

Collusion attack is a cost-effective attack against digital fingerprint. To develop an efficient collusion-resistant fingerprint scheme, it is very important for the detector to study the behavior of the colluders and the performance of collusion attack. Although several prior works have been proposed in the literature to analyze the performance of collusion attack, few effort has been made to explicitly study the relationship between risk, i.e., the probability of the colluders to be detected, and distortion of collusion attack. In this paper, we investigate the risk-distortion relationship of the linear video collusion attack with Gaussian fingerprint. We formulate the optimal linear collusion attack as an optimization problem, where the colluders try to minimize the distortion subject to a risk constraint. For any fixed risk constraint, the optimal distortion can be found using numerical optimization methods. By varying the risk constraint, we can obtain the risk-distortion model. We also conduct experiments to verify the proposed risk-distortion model using real video data.

***Index Terms***— Risk-distortion, collusion attack, CCCP.

## 1. INTRODUCTION

Nowadays, video sharing over public networks becomes more and more popular. This causes a critical problem to digital contents providers since their copyrighted contents can be easily duplicated and distributed without authorization. Digital fingerprinting is an important technique used for tracing the distribution of video content and protecting them from unauthorized redistribution [1]. It embeds a unique identification information into each distributed copy of video signal. When a copy is redistributed without authorization, the content providers can extract the embedded fingerprint to trace back the source of the leak.

Collusion attack is a common and effective attack against digital fingerprinting [2], where the attackers combine information from different copies to remove or attenuate the embedded fingerprints. Most of the existing methods focused on image collusion attack, where the source signals are the same. In this case, if no post-processing techniques such as blurring and sharpening are performed, the collusion copy usually has equal or even better quality than the distributed copy. However, this is not the case of collusion attack on video. Video data have a unique characteristic that the temporally adjacent frames are similar but not the same, due to which distortion would be introduced during the attack. Therefore, for video collusion attack, there exists a tradeoff between the fingerprint remained in the colluded copy, i.e. the probability of being detected, which is the risk for the colluders, and the quality of the colluded copy, i.e. distortion. And the question arises: what is the best tradeoff between risk and distortion?

To answer the above question, in this paper, we conduct a theoretical analysis of the risk-distortion relationship for the linear video collusion attack with Gaussian fingerprint. By modelling the residue as a Gaussian distribution, we express the risk and distortion as functions of the temporal filter coefficients, and formulate the collusion attack as an optimization problem of finding the optimal coefficients to minimize the distortion subject to a given risk constraint. We show that, under a fixed false alarm probability $\alpha$, when the risk is not larger than $\alpha$, the globally optimal coefficient can be found by solving a convex optimization problem. When the risk is larger than $\alpha$, the problem is not convex. However, the locally optimal coefficient can be found by the constrained concave-convex procedure (CCCP) [3]. Using the optimal coefficient, the risk-distortion model can be obtained. Finally, we conduct several experiments to verify the proposed risk-distortion model using real video data. To the best of our knowledge, this is the first work on risk-distortion analysis of video collusion attack.

The rest of this paper is organized as follows. Section II provides a brief background on fingerprint embedding and extracting. In section III, we conduct a theoretical analysis of the linear collusion attack and derive the risk-distortion relationship. Section IV shows the experimental results on real video signals. Finally, we draw a conclusion in Section V.

## 2. BACKGROUND INFORMATION

### 2.1. Fingerprint Embedding

For the $k^{th}$ user, the fingerprint embedding process is:

$$f_k(t) = I(t) + W_k(t), \tag{1}$$

where $I(t)$ and $f_k(t)$ are the $t^{th}$ frame of the original and fingerprinted video, respectively. And $W_k(t) = \alpha(t) \cdot \times w_k(t)$, where $w_k(t)$ is the $t^{th}$ frame of fingerprint signal and $\alpha(t)$ is a parameter used to control the energy of the embedded fingerprint to achieve the imperceptibility.

Since the fingerprint is unique for each user, orthogonal fingerprint modulation is used. Moreover, to resist intra collusion attack, the fingerprint $w_k$ between neighboring frames for the same user $k$ are correlated with each other [4].

Yan Chen, W. Sabrina Lin and K. J. Ray Liu are with Dept. ECE, University of Maryland, College Park. E-mail:{yan, wylin, kjrliu}@umd.edu.

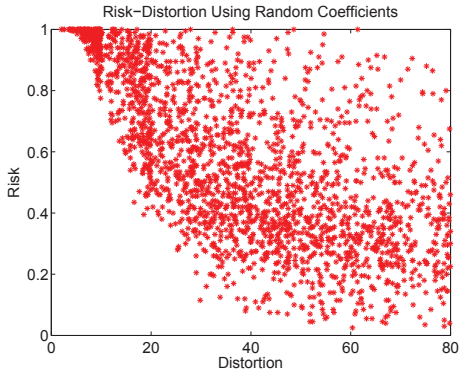



**Fig. 1**. Risk-Distortion Using Random Coefficient.

### 2.2. Fingerprint Extracting

Once the content owners found a suspicious copy, he/she can use correlation-based fingerprint detection to identify the attackers. Here, to make the analysis simple, we assume that frame-based detection is used. Similar analysis can be done if sequence-based or GOP-based detection is used. For each frame $\hat{f}(t)$, the fingerprint can be extracted using:

$$\hat{W}(t) = \hat{f}(t) - I(t). \tag{2}$$

Then, for each user $k$ who had received frame $t$, compute the detection statistics using:

$$TN_k(t) = \frac{W_k^T(t)\hat{W}(t)}{\sqrt{W_k^T(t)W_k(t)}}. \tag{3}$$

Finally, given a threshold $h$, the estimated attacker set for frame $t$ is $\hat{SC} = \{i : TN_i > h\}$.

## 3. THE PROPOSED MODEL

Let $M$ be the number of the colluders. Each attacker first performs intra attack by applying temporal filtering on the temporally adjacent frames. Then, all the attackers would collude together to perform inter attack. Since the fingerprint in every frame for each user is i.i.d, the weight allocated for the intra and inter attack would be the same for all the attackers. Therefore, the colluded copy of $t^{th}$ frame can be expressed as:

$$\hat{f}(t) = \sum_{k=1}^{M} \frac{1}{M}\Big[\sum_{i=-n}^{n} a_i f_k(t+i)\Big], \tag{4}$$

where $a_i$ is the weight of the $i^{th}$ frame, and $\sum_{i=-n}^{n} a_i = 1$.

### 3.1. Distortion

Let $d(t)$ be the difference between the colluded frame $\hat{f}(t)$ and the original frame $I(t)$, then,

$$d(t) = \hat{f}(t) - I(t) = (\mathbf{I}_r + \frac{1}{M}\sum_{k=1}^{M}\mathbf{W}_k)\mathbf{a}, \tag{5}$$

where $\mathbf{I}_r = [I(t-n) - I(t), ..., I(t+n) - I(t)]$, $\mathbf{W}_k = [W_k(t-n), ..., W_k(t+n)]^T$, and $\mathbf{a} = [a_{-n}, ..., a_n]^T$.

So, the distortion $D$, the mean square of $d(t)$, is:

$$D = E(d^T(t)d(t)) = \mathbf{a}^T\mathbf{K}_1\mathbf{a}, \tag{6}$$

where $\mathbf{K}_1 = E[\mathbf{I}_r^T\mathbf{I}_r] + \frac{1}{M}E[\mathbf{W}_1^T\mathbf{W}_1]$.

### 3.2. Risk of Being Detected

The detector can extract the fingerprint $\hat{W}(t)$ by:

$$\hat{W}(t) = \hat{f}(t) - I(t) = n_r + \sum_{j=1}^{M}\sum_{i=-n}^{n}\frac{a_i}{M}W_j(t+i), \tag{7}$$

where $n_r = \sum_{i=-n}^{n} a_i[I(t+i) - I(t)]$ is the linear combination of the residue. If we assume $[I(t+i) - I(t)] \sim N(0, \sigma_i^2)$, then $n_r \sim N(0, ||\Lambda\mathbf{a}||_2^2)$, where $\Lambda = diag\{\sigma_{-n}, ..., \sigma_n\}$.

By Eqn. (3) and (7), the detection statistic becomes:

$$TN_k(t) = \frac{W_k^T(t)[n_r + \sum_{j=1}^{M}\sum_{i=-n}^{n}\frac{a_i}{M}W_j(t+i)]}{\sqrt{W_k^T(t)W_k(t)}}. \tag{8}$$

So, $TN_k(t) \sim N(\mu_n, ||\Lambda\mathbf{a}||_2^2)$ with the mean $\mu_n = \frac{1}{M}\mathbf{p}^T\mathbf{a}$, where $\mathbf{p} = \left[E\left[\frac{W_k^T(t-n)W_k(t)}{\sqrt{W_k^T(t)W_k(t)}}\right], ..., E\left[\frac{W_k^T(t+n)W_k(t)}{\sqrt{W_k^T(t)W_k(t)}}\right]\right]^T$.

Let the risk $R$ be the probability of an attacker to be detected. Given a pre-defined threshold $h$, according to Eqn. (8), the risk $R$ can be computed by:

$$R = Prob(TN_k(t) > h) = Q(\frac{h - \frac{1}{M}\mathbf{p}^T\mathbf{a}}{||\Lambda\mathbf{a}||_2}), \tag{9}$$

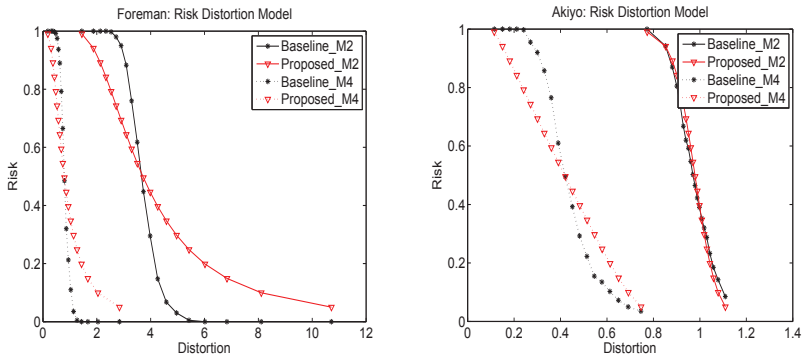
where $Q(x)$ is the Gaussian tail function $\int_x^\infty \frac{1}{\sqrt{2\pi}}\exp^{-\frac{x^2}{2}}dx$.

Similarly, the detection statistic of an innocent user satisfies Gaussian distribution $N(0, ||\Lambda\mathbf{a}||_2^2)$. Therefore, the probability of an innocent to be falsely detected as an attacker is $P_{fa} = Q(\frac{h}{||\Lambda\mathbf{a}||_2})$. If we fix the $P_{fa}$ to be $\alpha$, then $h = Q^{-1}(\alpha)||\Lambda\mathbf{a}||_2$ and the risk $R$ becomes:

$$R = Q(\frac{Q^{-1}(\alpha)||\Lambda\mathbf{a}||_2 - \frac{1}{M}\mathbf{p}^T\mathbf{a}}{||\Lambda\mathbf{a}||_2}). \tag{10}$$

### 3.3. The Risk-Distortion Relationship

From Eqn. (6) and (10), we can see that both the distortion and risk are determined by the coefficient $\mathbf{a}$. As shown in Fig. 1, for any fixed risk, there are many different $\mathbf{a}$, which lead to different distortion. The rational attackers will use the optimal $\mathbf{a}$ that minimizes the distortion to attack the fingerprinted videos. And the




(a) (b)

**Fig. 2**. The Risk-Distortion Model for Foreman and Akiyo Sequences: (a) Foreman; (b) Akiyo.

problem of finding such optimal **a** is:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \frac{1}{2}D = \frac{1}{2}\mathbf{a}^T\mathbf{K}\mathbf{a} \\ s.t. \quad & R = Q(\frac{Q^{-1}(\alpha)||\Lambda\mathbf{a}||_2 - \frac{1}{M}\mathbf{p}^T\mathbf{a}}{||\Lambda\mathbf{a}||_2}) \leq R_0; \\ & \mathbf{1}^T\mathbf{a} = 1. \end{aligned} \tag{11}$$

Obviously, the above optimization problem is not convex due to the quadratic term $||\Lambda\mathbf{a}||_2$ in the denominator. However, since $Q(x)$ is monotonic decreasing, the optimization problem can be re-written as:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \frac{1}{2}D = \frac{1}{2}\mathbf{a}^T\mathbf{K}\mathbf{a} \\ s.t. \quad & [Q^{-1}(R_0) - Q^{-1}(\alpha)]||\Lambda\mathbf{a}||_2 + \frac{1}{M}\mathbf{p}^T\mathbf{a} \leq 0; \\ & \mathbf{1}^T\mathbf{a} = 1. \end{aligned} \tag{12}$$

The optimization problem above is a Quadratically Constrained Quadratic Program (QCQP) problem. If $Q^{-1}(R_0) \geq Q^{-1}(\alpha)$, i.e. $R_0 \leq \alpha$, the problem is convex. We can find the optimal solution using numerical method.

If $Q^{-1}(R_0) < Q^{-1}(\alpha)$, i.e. $R_0 > \alpha$, the problem is non-convex. In general, a non-convex QCQP problem is a NP-hard problem. It is very difficult to find the global optimal solution. Fortunately, we can find the local optimal solution by CCCP [3], in which the concave term will be replaced with its first-order Taylor expansion. And the relaxed optimization problem becomes:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \frac{1}{2}D = \frac{1}{2}\mathbf{a}^T\mathbf{K}\mathbf{a} \\ s.t. \quad & [Q^{-1}(R_0) - Q^{-1}(\alpha)]\frac{\mathbf{a}^{(t)^T}\Lambda^T\Lambda\mathbf{a}}{||\Lambda\mathbf{a}^{(t)}||_2} + \frac{1}{M}\mathbf{p}^T\mathbf{a} \leq 0; \\ & \mathbf{1}^T\mathbf{a} = 1. \end{aligned} \tag{13}$$

Given an initial $\mathbf{a}^{(0)}$, CCCP computes $\mathbf{a}^{(t+1)}$ from $\mathbf{a}^{(t)}$ using Eqn. (13).

According to Eqn. (12) and (13), the optimal **a** that minimizes the distortion subject to a pre-defined risk constraint can be found using numerical method. Then, the corresponding distortion and risk can be computed using Eqn. (6) and (10). In this way, the risk-distortion relationship can be obtained, based on which the colluders can choose the optimal way to attack the fingerprint given any fixed risk constraint.

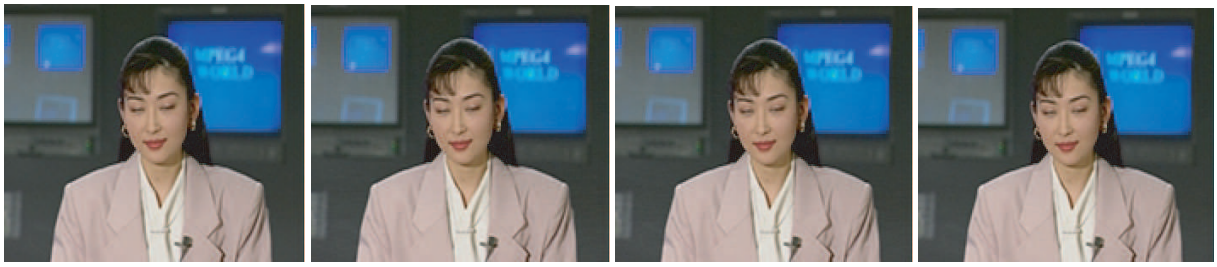
## 4. EXPERIMENTAL RESULTS

In order to evaluate the proposed risk-distortion model, we conduct the experiments on real video data. Two video sequences (akiyo, and foreman) in QCIF format, which represent slow and medium motion respectively, are tested. We use the human visual model based spread spectrum embedding in [5], and embed the fingerprint in the DCT domain. We generate independent vectors (length-$N$, with $N = 176 \times 144$) from Gaussian distribution $N(0, 1)$, and then apply Gram-Schmidt orthogonalization to produce fingerprint strictly satisfying $E[w_i(t)^T w_j(t)] = \delta_{i,j}$. Then, we scale the fingerprint to let the variance be $\sigma_w^2$. Finally, we perform inverse Gram-Schmidt orthogonalization to let the fingerprint of neighboring frame correlates with each other. We assume that the collusion attacks are also in the DCT domain. At the detector's side, a non-blind detection is performed where the host signal is first removed from the colluded copy. And the detector uses the detection statistics showed in Eqn. (3) to identify the attackers. In all the following experiments, the parameter $n$ in Eqn. (4) is set to be 5, which means that the 10 temporally adjacent frames are involved in the intra attack process for each attacker. The false alarm probability is set to be $\alpha = 10^{-3}$. $M$ at 2 and 4 are tested. $\sigma_w^2$ is set to be 20 and 30 for foreman and akiyo respectively. We compare the proposed risk-distortion model with the baseline curve, which is the experimental risk-distortion curve.

The risk-distortion curves are shown in Fig. 2. We can see that the proposed risk-distortion model is very accurate, almost identical to the baseline curve, especially for akiyo sequence. From Fig. 2, we can also see that with $M = 4$, the attackers can perfectly

(a) (b) (c) (d)

**Fig. 3**. Subjective Visual Quality Comparison for "Foreman" Sequence: (a) Original frame; (b) Fingerprinted frame; (c) Attacked frame with M=2 attackers where the corresponding (risk, distortion) are (0.07, 4.58); (d) Attacked frame with M=4 attackers where the corresponding (risk, distortion) are (0.11, 1.03).



(a) (b) (c) (d)

**Fig. 4**. Subjective Visual Quality Comparison for "Akiyo" Sequence: (a) Original frame; (b) Fingerprinted frame; (c) Attacked frame with M=2 attackers where the corresponding (risk, distortion) are (0.085, 1.11); (d) Attacked frame with M=4 attackers where the co rresponding (risk, distortion) are (0.035, 0.75).

attack the fingerprinted video with distortion (MSE) smaller than 2 (0.8) for foreman (akiyo). Even for the case $M = 2$, the minimal distortion with zero risk is smaller than 6 (1.2) for foreman (akiyo). This shows that the attackers can easily break the system using linear attack with the optimal coefficient derived by the proposed risk-distortion model.

The subjective visual quality of the attacked video are also examined in Fig. 3 and 4, where (a) and (b) are the original and fingerprinted frame respectively. Fig. 3 (c) shows the attacked frame for foreman using the proposed method with $M = 2$, where the corresponding (risk, distortion) are (0.07, 4.58). We can see that the quality is good except the hallucinating artifacts on the face. When $M = 4$, this artifacts are greatly reduced, as shown in Fig. 3 (d). The corresponding (risk, distortion) of Fig. 3 (d) are (0.11, 1.03). Since akiyo is a slow motion sequence, intra attack is very efficient, due to which the quality of the attacked frame are almost the same as the original frame, as shown in Fig. 4 (c) and (d).

## 5. CONCLUSIONS

In this paper, we provided a theoretical analysis on the risk-distortion relationship of the linear video collusion attack with Gaussian fingerprint, and conducted several experiments on real video sequences to verify the proposed risk-distortion model. From the experimental results, we could see that the attackers could easily break the fingerprint with a very small distortion when the source sequence is a slow motion sequence such as Akiyo. Even when the source sequence is a medium motion sequence, e.g. Foreman, the detector could not catch any colluders with a reasonable small distortion when the number of the colluders is 4. Therefore, the detector may need to use both the fingerprint of current frame and those of the neighboring frames to improve the detection performance.

## 6. REFERENCES


[1] F. Zane, "Efficient watermark detection and collusion security," in *Proc. Financial Cryptography*, 2000, pp. 21–32.

[2] H. V. Zhao, M. Wu, Z. J. Wang, and K. J. Ray Liu, "Forensic analysis of nonlinear collusion attacks for multimedia fingerprinting," vol. 14, pp. 646–661, 2005.

[3] Pak-Ming Cheung and James T. Kwok, "A regularization framework for multiple-instance learning," in *Proc. International Conference on Machine Learning*, 2006.

[4] K. Su, D. Kundur, and D. Hatzinakos, "Statistical invisibility for collusion-resistant digital video watermarking," vol. 7, pp. 43–51, 2005.

[5] C. Podilchuk and W. Zeng, "Image adaptive watermarking using visual models," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 525–540, 1998.