COLLUSION-RESILIENT FINGERPRINTING USING RANDOM PRE-WARPING

Mehmet U. Celik[a], *Gaurav Sharma*[a], *A. Murat Tekalp*[a,b]

[a] Electrical and Computer Engineering Dept., University of Rochester, Rochester, NY, 14627-0126
[b] College of Engineering, Koc University, Istanbul, Turkey
{celik,tekalp}@ece.rochester.edu, g.sharma@ieee.org

ABSTRACT

Fingerprinting of audio-visual content using digital watermarks is an effective means of determining the originators of unauthorized copies and fighting piracy in digital distribution networks. In particular, watermarks embedded within the content help trace the traitor responsible for the piracy. A group of users may, however, collude and collectively escape identification by creating an average of their individually watermarked copies that appears unwatermarked. We propose a novel collusion-resilience mechanism, wherein the host signal is warped randomly prior to watermarking. As each copy undergoes a distinctive warp, collusion through averaging either yields low-quality results or requires substantial computational resources to undo random warps. The proposed method is independent of the watermarking scheme used and does not impose any restrictions on the watermark signal that are required by some collusion resistant watermarking schemes. We demonstrate the effectiveness of this approach on digital images.

1. INTRODUCTION

The term *fingerprinting* or *traitor tracing* refers to addition of a unique mark on each copy of distributed content. If and when an unauthorized copy of the fingerprinted content is made and distributed, the mark (fingerprint) embedded in the leaked copy uniquely identifies the traitor. Historically, fingerprinting has been mostly utilized for secret documents. The small number of users in traditional settings and relatively high value of the documents have allowed for manual embedding and detection of fingerprints, typically by making minor, but distinct, edits to the document. In large scale digital distribution networks, however, manual editing is prohibitive .

Digital watermarking [1, 2] offers an efficient method for automatic fingerprinting of content for distribution over large networks. Typically, the digital watermark (fingerprint) is a pseudo-noise pattern representing the identity of the user and it is superimposed on the host signal—the content to be distributed, e.g. the image. Later, when an unauthorized copy is found, the presence of a particular watermark pattern reveals the identity of the traitor who has compromised the content.

One particular strategy for the traitors to escape identification upon unauthorized distribution is working collectively in order to disable the watermark. In particular, a group of traitors collude and use each of their fingerprinted copies to obtain a copy within which the fingerprint patterns cannot be detected reliably. Collusion often involves averaging several distributed copies and results in a high-quality untraceable copy.

Collusion-secure watermarks proposed earlier [3, 4] impose restrictions on the construction of the watermark pattern to thwart collusion attacks. In these algorithms, the watermark patterns—or payloads—are partially correlated and any subset of these patterns bears a static component. This static component not only survives the averaging attack, but also uniquely identifies a particular subset of watermarks that were averaged—revealing the group of traitors. Nevertheless, security against collusion often comes in the expense of the capacity and/or robustness of the original watermarking algorithm [4, 5] and its performance degrades as the number of traitors increases.

An alternative approach to collusion-resilient fingerprinting is proposed here. The method imposes no restrictions on the watermark pattern and may therefore be used in conjunction with any watermarking method. In contrast with schemes that detect collusion and identify colluders, the proposed scheme is aimed at preempting collusion by preventing traitors from obtaining a high-quality copy through collusion. In our approach, the geometry of the host-signal is randomly and imperceptibly distorted prior to watermark addition and distribution. If a group of users collude and average multiple distinctively warped copies of the content, the mis-registration, i.e. the mismatch between the underlying geometries, makes the resulting averaged version inferior in quality.

2. FINGERPRINTING AND COLLUSION ATTACK

We present a brief overview of the use of digital watermarking for fingerprinting purposes and describe the collusion attack. The watermarking scheme described here is a specific method which is based on additive spread-spectrum techniques. Alternative watermarking methods, such as quantization based algorithms, may be also used for fingerprinting purposes. In general, both the collusion attack presented in this section and the solution proposed in the next section are valid, regardless of the selected watermarking method.

Let us assume that the content which will be distributed is a continuous signal, such as an audio recording. We refer to this signal, upon which a digital fingerprint will be imposed, as the host signal and denote it by S. The unique fingerprint (watermark) which is unique to a particular user is denoted by W_i, where subscript i is linked to the identity of that user. The watermarking procedure superimposes the watermark pattern W_i on the host signal and yields the fingerprinted signal S_i.

$$S_i = S + W_i \tag{1}$$

for all $i \in \{1, ..., N\}$, where N is the total number of users.

When an unauthorized copy of the host signal S_u is found, the presence of the particular watermark is checked using a correlation

0-7803-7750-8/03/$17.00 ©2003 IEEE.

detector. That is W_i is present in the signal if

$$\langle S_u, W_i \rangle > Threshold \quad (2)$$

It is desirable to design watermark patterns W_i such that they are uncorrelated with the host signal and with each other. If $S_u = S_i$, the correlation is randomly distributed around $||W_i||$ and around zero otherwise. The $Threshold$, determines the trade-off between the false positives and misses.

To escape identification and prevent detection of watermark patterns, a group of users may obtain an average signal using their individually watermarked copies. For K colluding users, the average signal is

$$S_{avg} = \frac{1}{K}\sum_{i=1}^{K} S_i = S + \frac{1}{K}\sum_{i=1}^{K} W_i \quad (3)$$

It should be noted that the correlation between the average signal S_{avg} and a particular watermark pattern W_j, $\langle S_{avg}, W_j \rangle$ decreases linearly with K, the number of copies used. In general, it is possible to prevent detection by forcing the correlation value under the threshold using more and more copies. Moreover, the quality of the of the averaged signal is often superior when compared with the fingerprinted signal,

$$d(S_{avg}, S) < d(S_i, S) \quad (4)$$

where $d(\cdot,\cdot) = || \cdot - \cdot ||$ is the Euclidean distance.

3. COLLUSION-RESILIENCE WITH RANDOM PRE-WARPING

Consider a set of warping functions Φ such that for all functions $\phi_i(\cdot) \in \Phi$, $\phi_i(S)$ is perceptually identical to S, but the Euclidean distance between two signals is significantly large ($d(\phi_i(S), S) >> 1$). We further require that the Euclidean distance between different warped versions of the same signal is large, i.e. $d(\phi_i(S), \phi_j(S)) >> 1$ for $i \neq j$. The set of functions Φ may be populated, for instance, by functions that correspond to small local distortions on the geometry of the signal. In a majority of cases, the human perceptual system is highly tolerant of such manipulations and the overall effect is imperceptible, despite the large mean squared error (MSE) distortion among signals. This property has been previously exploited in Stirmark [6] where it forms the basis of de-synchronization attacks on a number of watermarking schemes. Here instead we exploit the same property to provide collusion resilience.

3.1. Collusion-Resilience for Oblivious Watermarking

We first discuss pre-warping for oblivious watermarking schemes, where the watermark signal is detected without any reference to the original host signal. Our method is based on applying a different, randomly selected warping function $\phi_i(\cdot)$ to the host signal prior to addition of the watermark signal W_i. That is, the i^{th} watermarked signal is formed as

$$S_i = \phi_i(S) + W_i. \quad (5)$$

Eqn. 5 captures a number of different cases where the present technique may be applied. For instance, in the case of a video stream where S is a function of 2-D spatial coordinates x, y and time t, the geometric warping is applicable in the 3-D spatio-temporal space (x, y, t) and Eqn. 5 may be written as

$$\begin{aligned} S_i(x,y,t) &= S(x',y',t') + W_i & (6) \\ x' = \phi_i^x(x,y,t);\ y' &= \phi_i^y(x,y,t);\ t' = \phi_i^t(x,y,t) & (7) \end{aligned}$$

The requirement of imperceptible visual distortion can then be imposed as a smoothness requirement on the 3-D coordinate transformation in Eqn. 7.

Watermark detection is performed on a suspected copy S_u, through a correlation detector as earlier (Eqn. 2). As the warped host signal bears statistical characteristics similar to the original host signal, performance of the detection is not affected by the warping.

As in the earlier scheme, several users may try and collude and obtain an average host signal in order to thwart the watermark. The average signal from a collusion attack by K users is given by

$$S_{avg} = \frac{1}{K}\sum_{i=1}^{K} S_i = \frac{1}{K}\sum_{i=1}^{K} \phi_i(S) + \frac{1}{K}\sum_{i=1}^{K} W_i \quad (8)$$

Note that by comparing the averaged watermark terms in Eqn. 8 and Eqn. 3, we can see that the averaging attack still impairs the watermark detection to the same degree as the prior schemes: the correlation value decreases linearly with the number of copies averaged. However, if we compare the corresponding "signal" terms in Eqn. 8 and Eqn. 3 one can see a very significant difference. The signal terms averaged in Eqn. 3 are identical and therefore cause no degradation of the signal, whereas due to the warping the signal terms averaged in Eqn. 8 are different and this typically produces an averaged copy often of significantly inferior quality. Averaging of the differently warped images produces several artifacts in the signal ($d(S_{avg}, S) >> 1$) which are typically both perceptually significant and disturbing. For instance, if the host signal is a digital image the average appears to have multiple ghost images or it is a blurred version. In either case, the colluded copy is a rather low quality version.

Note that the proposed scheme does not guarantee absolute security against collusion. Given a number of copies, it is possible to undo the warping to align all signals to a common geometry and then average the registered copies. This in effect undoes the individual warping operations ϕ_i. Nonetheless, this requires significant additional effort on the part of the colluders in terms of computation time and custom software.

3.2. Collusion-Resilience for Non-Oblivious Watermarking

The method described above can also be used with non-oblivious watermarks, which utilize the original signal during detection. In non-oblivious watermarking, the warping can be performed either *before* or *after* the addition of the watermark. If the warping is applied prior to watermarking, the warped unwatermarked signal has to be present at the decoder. This can be achieved either by storing the warped signals or the warping parameters along with the original signal. On the other hand, if the warping is applied after watermarking, it has to be undone before detection to prevent loss of synchronization. Although, undoing random geometric manipulations is challenging, this process can be simplified significantly by utilizing a pre-determined subset of all possible warping parameters. Unless the watermarked signal is warped again by a third party, detector performs a limited search in this small subset.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

We demonstrate the transparency and effectiveness of the proposed anti-collusion algorithm on digital images. In this context, we utilize the Stirmark tool [6] (version 3.1.79), which we use exclusively to apply geometric distortions while disabling its other attack features. A spatial domain spread-spectrum watermarking system with minimal complexity is chosen to simplify the analysis. The watermark is an additive white Gaussian pseudo-noise with variance σ_W^2. (In our experiments watermarking strength is set to $\sigma_W^2 = 9$, which corresponds to 38.6 dB embedding distortion.) The detection is performed through normalized correlation (Eqn. 10) which is preceded by Wiener filtering (Eqn. 9).

$$\hat{W} = \frac{\sigma_W^2}{\sigma_S^2 + \sigma_W^2}(S_u - \mu_{S_u}) \tag{9}$$

$$NC = \frac{\langle \hat{W}, W \rangle}{\sqrt{\langle \hat{W}, \hat{W} \rangle \langle W, W \rangle}} \tag{10}$$

If the watermark estimate perfectly matches the watermark, normalized correlation equals to 1. However, the watermark estimate is often corrupted with noise interference from the image $\hat{W} = W_i + N$. Similarly, $\hat{W} = W_i/K + N$ when K-copies of the content are averaged. If we assume that the noise and the watermark patterns are uncorrelated ($\langle N, W_i \rangle = 0$) and the noise-to-watermark ratio is constant, i.e. $c_0 = \frac{\langle N,N \rangle}{\langle W_i, W_i \rangle}$, then Eqn. 10 simplifies to

$$NC = \frac{1}{\sqrt{1 + c_0 K^2}}. \tag{11}$$

The visual impact of the proposed scheme is first demonstrated through an example. The original gray-scale image (800×600) is shown in Fig. 1. Watermarked images are obtained by distorting the image geometry with Stirmark (using a random seed) and adding a key-dependent pseudo-noise watermark pattern. A representative watermarked image is seen in Fig. 2. While the distortion produces a rather low peak signal to noise ratio (PSNR) (31.4 dB) with respect to the original, the perceptual quality of the image is largely unchanged. The result of a simulated collusion averaging attack employing two images is shown in Fig. 3. Note that the visual quality of the image is rather poor showing clear ghosting around the edges, despite a similar PSNR value of 30.7 dB. (The visual quality of images can be assessed better in electronic versions compared to hard-copy prints.) The result of the collusion attack using nine watermarked images is shown in Fig. 4. In this case, the average image is blurred and its perceptual quality is significantly degraded (PSNR = 29.3 dB). The normalized correlation coefficients for Figs. 2, 3, 4 are 0.39, 0.26, and 0.11, respectively.

We also test the performance of the watermark with and without the proposed pre-warping scheme in a set of 23 gray-scale images (size 768×512) from Kodak Photo-CD. Figures 5 and 6 summarize this performance as averaged results over all images and for 10 iterations using different seeds for the random-number generator. In Fig. 5, we observe that in the absence of pre-warping the quality of colluded copy improves significantly (up to 6 dB over the watermarked image) with increasing number of colluders. The corresponding curve with pre-warping is also included in the same figure. Note however that the PSNR metric based on pixel-wise differences is not meaningful in the presence of geometric deformations.

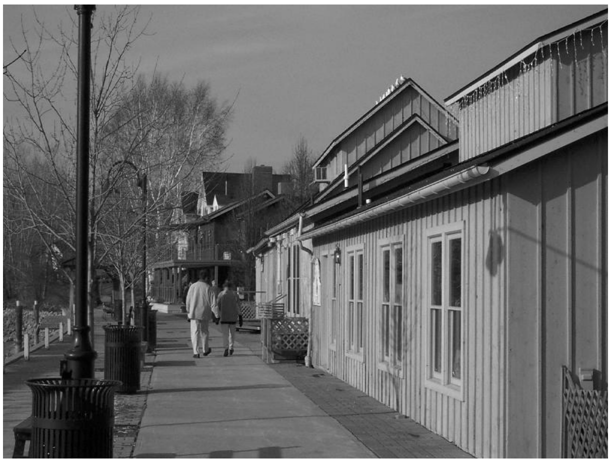

Fig. 1. Original image.

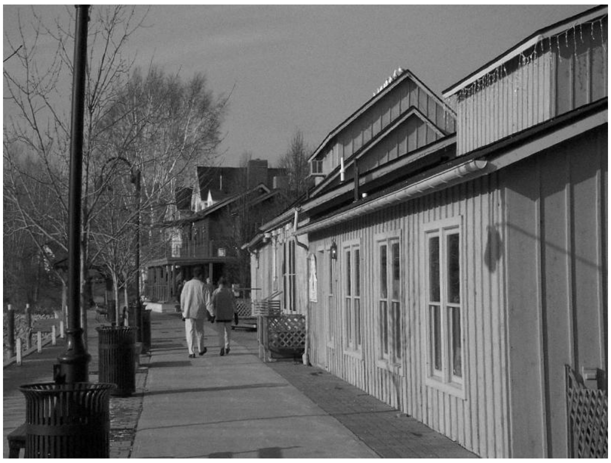

Fig. 2. Image after warping by Stirmark [6] and watermark addition. The distortion is perceptually tolerable.

In Fig. 6, normalized correlation values for the classical and proposed schemes are plotted. In addition, normalized correlation values in case of collusion are estimated according to Eqn. 11 (solid line). As expected, we observe the sharp drop in the detector's response when multiple watermarked copies are averaged. This shows that the collusion attack significantly reduces the normalized correlation, potentially limiting the capacity/robustness of the underlying fingerprinting system. Surprisingly, the proposed anti-collusion feature offers an improvement in watermark detection—in addition to the degradation in visual quality of the colluded copy. Averaging independently warped copies of an image produces a smooth (blurred) image that has a lower variance. Consequently, the image can be filtered out more effectively by the Wiener filter. The residual noise—thus the noise-to-watermark ratio c_0—is reduced and the normalized correlation values are improved with respect to the classical case. It is worth mentioning that this improvement is an unintended consequence of the proposed method in the given watermarking system and these results cannot be generalized to other watermark embedding and detection systems.

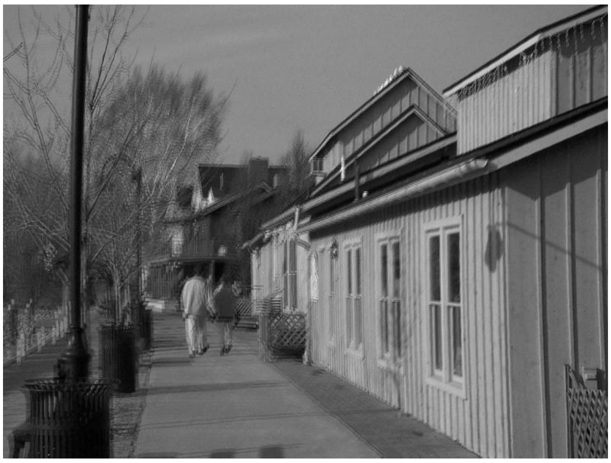

Fig. 3. Collusion result when two warped images are averaged. Ghosting around edges is visible and disturbing.

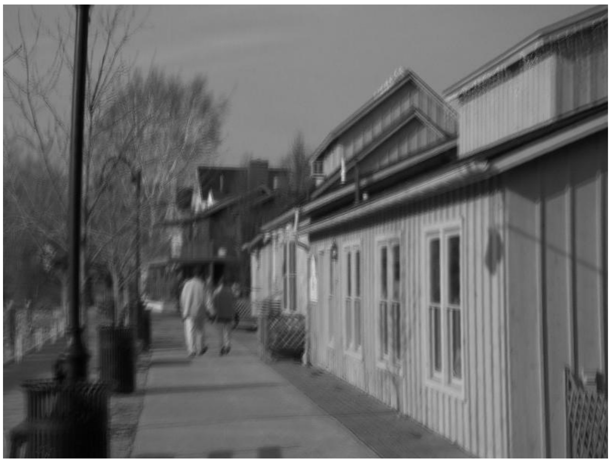

Fig. 4. Collusion result when nine (9) warped images are averaged. The image is blurred and bears little commercial value.

5. CONCLUSION AND DISCUSSION

We have presented an alternative method for collusion-resilient watermarking. Our approach is based on randomly and uniquely pre-warping each copy of the host-signal prior to distribution. Human perception is quite tolerant of the geometric distortion and the warping therefore does not significantly affect the perceived quality of the watermarked signal. On the other hand, as the geometry of each copy is distorted independently, a collusion attack yields a low-quality signal. We have shown that collusion even with only two copies results in disturbing distortions—ghost edges—and visual quality does not improve when the number of copies is increased. Higher-quality collusion is only possible by undoing the warping which requires special software and substantial computational resources. Finally, the solution does not adversely affect the performance (capacity/robustness) of the underlying watermarking scheme. The approach can either replace existing collusion resistant watermarks or it can be applied in conjunction to identify even traitors who choose to use the low-quality average signals.

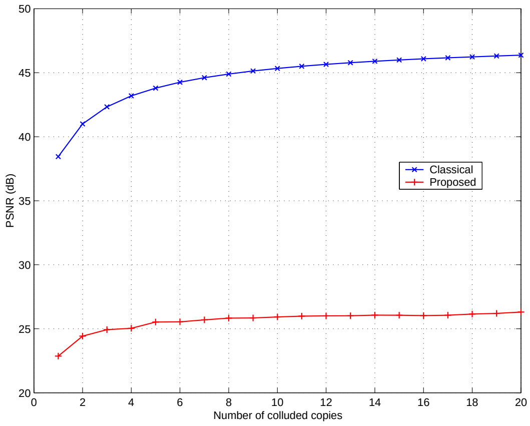

Fig. 5. PSNR of the colluded copy vs. number of colluders (without pre- warping).

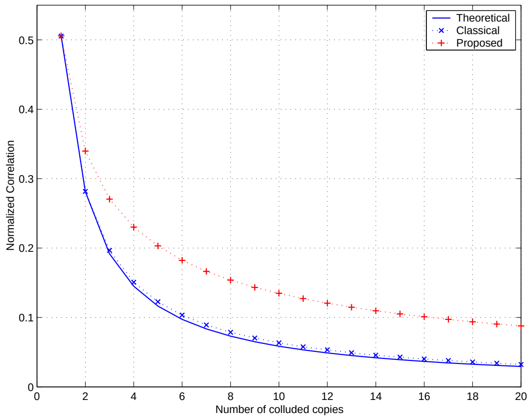

Fig. 6. Normalized correlation vs. number of colluders (with and without pre-warping).

6. REFERENCES

[1] R. L. G.C. Langelaar, I. Setyawan, "Watermarking digital image and video data," *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 20–46, Sept. 2000.

[2] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. of the IEEE*, vol. 87, no. 7, pp. 1079–1107, July 1999.

[3] J. Dittmann, A. Behr, and M. Stabenau, "Combining digital watermarks and collusion secure fingerprints for digital images," *Proc. of SPIE Sec. and Watermarking of Multimedia Cont. I*, vol. 3657, no. 13, Jan. 1999.

[4] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. on Information Theory*, vol. 44, no. 5, pp. 1897–1905, Sept. 1998.

[5] J. Su, J. Eggers, and B. Girod, "Capacity of digital watermarks subjected to an optimal collusion attack," in *EUSIPCO 2000*, Tampere, Finland, Jan. 2000.

[6] F. Petitcolas, R. Anderson, and M. Kuhn, "Attacks on copyright marking systems," in *Information Hiding Workshop, IH'98*, Portland, OR, USA, Apr. 1998, pp. 219–239.