Resistance of Digital Watermarks to Collusive Attacks

Joe Kilian
NEC Research Institute[1]

F. Thomson Leighton[6]
MIT LCS and Department of Mathematics[3]

Lesley R. Matheson[6]
InterTrust STAR Lab[2]

Talal G. Shamoon[6]
InterTrust STAR Lab[2]

Robert E. Tarjan[6]
InterTrust STAR Lab[2] and Princeton University[4]

Francis Zane[6]
UCSD[5] and InterTrust STAR Lab[2]

In *digital watermarking* (also called digital fingerprinting), extra information is embedded imperceptibly into digital content (such as an audio track, a still image, or a movie). This extra information can be read by authorized parties, and other users attempting to remove the watermark cannot do so without destroying the value of the content by making perceptible changes to the content. This provides a disincentive to copying by allowing copies to be traced to their original owner. Unlike cryptography, digital watermarking provides protection to content that is in the clear.

It is not easy to design watermarks that are hard to erase, especially if an attacker has access to several differently marked copies of the same base content. Cox et al. [1] have proposed the use of additive normally distributed values as watermarks, and have sketched an argument showing that, in a certain theoretical model, such watermarks are resistant to collusive attacks. Here, we fill in the mathematical justification for this claim.

In our model, the original content or document is an n-dimensional vector V, with components independently distributed according to $N(0,1)$, the normal distribution with mean zero and variance one. We assume that V' is perceptibly different from V if $\|V' - V\| > \delta\sqrt{n}$, where the norm is Euclidean and δ is a suitable constant.

To insert a watermark, we first produce a random n-dimensional vector X whose components are independently distributed according to $N(0,\alpha)$ for a suitable α. We then add X to V to form a watermarked copy $W = V + X$. If $\alpha < \delta$, this change is imperceptible with high probability.

To search for a watermark in a document V^* that may have been modified by an adversary, we first compute an *implied watermark* $X^* = V^* - V$. Then, we test X^* against each watermark X previously used by computing a similarity measure $s(X^*, X) = X^* \cdot X/\|X^*\|$ (a normalized correlation). If this similarity measure exceeds a threshold t, we claim that the watermark X is present in V^*, and thus the user given X was responsible for producing V^*.

Our main result is that, if m is the total number of watermarks distributed, the scheme is secure against k-way collusive attacks with $k = O(\sqrt{n/\ln m})$. In a k-way collusive attack, an adversary obtains k differently marked copies of the same document V, and from them, by any means, prepares a malicious copy V^*. The security of the scheme has two components: First, we require that the adversary is highly unlikely to falsely implicate other users. Second, with high probability, a V^* produced by the adversary is perceptibly different from V, or the test determines that at least one of the watermarks owned by the adversary is present in V^*.

As long as n is sufficiently large compared to the number of users, then the adversary is unlikely to guess a document whose implied watermark correlates with another user, since the watermarks of the other users were generated independently. Proving that the adversary cannot disguise his own watermarks is more difficult. To do this, we look at the posterior distribution of the original document, given the k watermarked copies. We show that the variation in this distribution is sufficiently large that there is no document the adversary can chose which is, with nonnegligible probability, both unwatermarked and close to the original.

References

[1] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia", *IEEE Trans. on Image Processing* **6** (12). pp. 1673-1687 (1997).

[1] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

[2] STAR Lab, InterTrust Technologies, 460 Oakmead Parkway, Sunnyvale, CA 94086

[3] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

[4] Department of Computer Science, Princeton University, Princeton, NJ 08544

[5] Department of Computer Science, University of California, San Diego, La Jolla, CA 92093.

[6] Some of this work done at the NEC Research Institute