DANGER OF LOW-DIMENSIONAL WATERMARKING SUBSPACES

Gwenaël DOERR and Jean-Luc DUGELAY

Department of Multimedia Communications
Eurécom Institute, Sophia-Antipolis, FRANCE
http://www.eurecom.fr/∼image

ABSTRACT

The security issue has been neglected for a long time in digital watermarking. Recent results for video watermarking have pointed out that existing watermarking schemes were not secure i.e. hostile intelligence succeeds in removing the hidden watermarks. In particular, for a given secret key, many watermarking schemes embed watermarks which lie in the same low-dimensional subspace whatever the host data is. In this article, it will be shown that this subspace can be quite easily estimated with an efficient Principal Component Analysis (PCA). For storage convenience, an online Expectation-Maximization (EM) algorithm will be considered. Once this watermarking subspace has been estimated, an attacker only has to project incoming data onto the orthogonal of this subspace to remove the watermark.

1. INTRODUCTION

Recent technological advances have permitted to initiate the transition from the analog world to the digital one. Unfortunately, this has raised many concerns regarding copyright protection since digital data can be easily and perfectly duplicated and rapidly redistributed at a large scale. Encryption is usually used to make digital data completely useless for people not having the correct decryption key. However, encrypted data has to be decrypted sooner or later to be eventually presented to a human observer/listener i.e. its protection falls during content presentation. As a result, digital watermarking [1] has been introduced in the 90's as a second line of defense. Basically, a key dependent secret signal is embedded into digital data in a robust and invisible way. Moreover, this underlying signal is closely tied to the host data so that it survives D/A conversion. There exists a complex trade-off between conflicting parameters (data payload, invisibility, robustness and security) and a compromise has to be found, which usually depends on the targeted application.

In its infancy, digital watermarking has been extensively studied for still images and the research effort was mainly devoted to achieve enhanced robustness, additional payload or less visible watermarks. The security issue was almost ignored apart from works defining secure protocols for watermarking [2]. When a user wants to insert some copyright information in digital data, a Trusted Third Party (TTP) computes the payload to be embedded, e.g. a hash string depending of the encrypted copyright information and the host data, and keeps it in a repository with a timestamp. The user embeds then this payload with a given watermarking scheme. As a result, a user with fixed secret key and copyright information always embed a different watermark in alternative multimedia documents, which may first sound appealing to resist hostile attacks such as the copy attack [3] or watermark estimation collusion attack [4]. Security raised more concerns when results for still images were extended to video watermarking. Most of the proposed schemes have indeed relied on a frame by frame approach which has leaded to non-secure algorithms.

Watermark security has been defined as *the inability by unauthorized users to have access to the raw watermarking channel* [5]. In particular, unauthorized users should not be able to estimate and/or remove the embedded watermarks. In Section 2, it will be reminded that many algorithms come down to an additive watermark which lies in a low-dimensional subspace. As a result, an hostile attacker can gather several works watermarked with the same secret key to estimate this watermarking subspace via PCA as described in Section 3 for later removal. The experimental results reported in Section 4 will further demonstrate the danger of such an attack in practice. Eventually, security issues will be discussed in Section 5 and possible tracks to survive this attack will be proposed.

2. ADDITIVE WATERMARKING

Many algorithms have been proposed to hide a secret watermark in digital multimedia content. Most of them can be reduced to a simple additive scheme as expressed below:

$$\mathbf{c}_{\mathrm{w}} = \mathbf{c}_{\mathrm{o}} + \sum_{i=1}^{p} \lambda_i \mathbf{w}_i(K) \tag{1}$$

where $\mathbf{c}_{\mathrm{o}}$ is the original digital content and $\mathbf{c}_t extrmw$ its watermarked version. The $\mathbf{w}_i$'s are orthogonal pseudo-random watermark patterns dependent of a secret key K and the λ_i's their associated mixing coefficients. On the detector side, the hidden bits are then obtained back by correlating the watermarked digital content with each one of the pseudo-secret watermark patterns:

$$\rho_i = \mathbf{c}_{\mathrm{w}} \cdot \mathbf{w}_i(K) \tag{2}$$

where $\cdot$ is a correlation operator e.g. linear correlation.

Code Division Multiple Access (CDMA [6]). The watermark patterns $\mathbf{w}_i$'s have the same dimensions than the host digital content $\mathbf{c}_{\mathrm{o}}$. Furthermore, they are normally distributed with zero mean and unit variance and are orthonormalized, thanks for example to a Gram Schmidt procedure. The mixing coefficients λ_i are set equal to $\pm\alpha$ where α is a fixed embedding strength. The sign of the λ_i's is chosen accordingly to the bits to be hidden e.g. if the i^{th} bit is equal to 0 (resp. 1), the associated λ_i should be negative (resp.

Contact author: dugelay@eurecom.fr

positive). The distortion introduced during the embedding process is then equal to $\alpha^2 p$ in terms of Mean Square Error (MSE). On the detector side, the hidden bits are retrieved with respect to the sign of the correlation scores ρ_i's.

Time Division Multiple Access (TDMA [7]). The procedure is exactly the same than for CDMA modulation except that each watermark pattern $\mathbf{w}_i$ does not cover anymore the whole host digital content. In fact the host digital content is partitioned into p distinct chunks $\mathbf{c}_{\text{o}}^i$. Each watermark pattern $\mathbf{w}_i$ is then generated in such a way that it is normally distributed with zero mean and unit variance on $\mathbf{c}_{\text{o}}^i$ and equal to 0 elsewhere. In other terms, the watermark patterns do not "overlap" and are *de facto* orthogonal. The embedding process now introduces a distortion equal to α^2 in terms of MSE.

Dither Index Modulation (DIM [8]). Non overlapping watermark patterns are used as in TDMA. The key idea is then to chose the mixing coefficients λ_i so that the correlation scores ρ_i after quantization define a point on a regular lattice which encodes the hidden message. Alternative points of the lattice can encode the same message and the embedder selects the one which introduces the lowest distortion in terms of MSE. Assuming that Δ is the difference between two quantized values which encode the same symbol, the mixing coefficients λ_i vary uniformly between $-\Delta/2$ and $+\Delta/2$ and the introduced distortion is equal in average to $\Delta^2/12$ in terms of MSE. On the detector side, the hidden message can be extracted back according to the quantized values of the correlation scores.

Embedding Strength Modulation (ESM [9]). Introduced in the specific context of video watermarking, this approach uses overlapping watermark patterns as in CDMA modulation and relies on time-dependent mixing coefficients. Temporally zero-mean λ_i's are used for security reasons and phase differences between them are introduced to encode the payload. For example, for a given frame, the mixing coefficient can be written $\alpha \sin(\Phi + \Phi_i)$ where α is a fixed embedding strength. As a result the introduced distortion is equal to $\alpha^2 p/2$ in terms of MSE. On the detector side, the correlation scores are computed for each frame and the phase differences are estimated to obtain back the hidden payload.

Other watermarking algorithms can be reduced to the simple additive scheme defined in (1). In all those algorithms, the secret key K defines a *private watermarking subspace* which has usually fewer dimensions than the whole media space and, whatever the message is, whatever the host content is, the embedded watermark is bounded to this private low-dimensional subspace. The danger due to the use of such reduced subspaces will be investigated in the next section.

3. COLLUSION ATTACK

In previous work, it has been shown that, when a watermark is embedded in a digital document, it can be roughly estimated and remodulated [10]. This approach has been further extended to video to obtain a refined estimation of a redundantly embedded watermark [4]. In other terms, the original attack has become a collusion attack i.e. several watermarked digital contents (each video frame) are gathered to produce unwatermarked digital content. In this article, this collusion approach will be further developed in a slightly more general framework: several alternative watermarks are embedded in distinct digital contents but they all belong to the same low-dimensional subspace. Since the embedded watermarks can be randomly distributed in this watermarking subspace, the goal is to estimate this subspace rather than the individual watermarks and then to remove any residual watermark signal. The process can be divided into three steps: individual watermarks estimation, watermarking subspace estimation and finally watermark removal.

3.1. Watermark Estimation

In the proposed approach, the first task of an attacker is to collect several individual watermark estimations from several watermarked contents. The ideal estimator would consist in computing the difference between the watermarked content and the associated original one i.e. $\mathbf{w} = \mathbf{c}_{\text{w}} - \mathbf{c}_{\text{o}}$. However, the attacker does not have access to the original digital content in practice and should estimate the underlying watermark in a blind manner. Digital watermarks are usually located in high frequencies. As a result, a rough estimation can be obtained thanks to denoising techniques, or more simply by computing the difference between the watermarked digital content and its low-pass filtered version:

$$\tilde{\mathbf{w}} = \mathbf{c}_{\text{w}} - \mathcal{L}(\mathbf{c}_{\text{w}}) \tag{3}$$

where $\mathcal{L}(.)$ is a low-pass filter operator. With such a strategy, some samples are doomed to be badly estimated e.g. around discontinuities. An additional thresholding operation is consequently performed to isolate non-pertinent samples and ignore them for the remaining of the attack. For example, estimated samples whose magnitude is greater than τ_{discard} can be discarded.

3.2. Watermarking Subspace Estimation

At this stage, the attacker has a collection of n individual watermark estimates of size s. Knowing that the watermarks are all contained in a subspace having p dimensions ($p \ll s$), he/she wants to find p vectors $\mathbf{e}_i$ which generate the same subspace than the one created by the secret watermark patterns $\mathbf{w}_i$:

$$\mathcal{W} = \text{span}(\mathbf{w}_i) = \text{span}(\mathbf{e}_i) = \mathcal{E} \tag{4}$$

With this end in view, Principal Component Analysis (PCA) is performed as it is an optimal dimension reduction technique. Let $\tilde{\mathbf{W}}$ be a $s \times n$ matrix whose columns are the previously computed watermark estimates $\tilde{\mathbf{w}}$. The goal is then to find a $s \times p$ matrix $\mathbf{C}$ and a $p \times n$ matrix $\mathbf{V}$ which minimize the norm $\|\tilde{\mathbf{W}} - \mathbf{CV}\|$. A column of matrix $\mathbf{V}$ can be regarded as the coordinates of the associated watermark estimate in matrix $\tilde{\mathbf{W}}$ in the principal subspace generated by the vectors defined by the columns of matrix $\mathbf{C}$.

A major shortcoming of standard approaches to PCA is that high dimensional data are difficult to be dealt with. Troubles can arise in the form of computational complexity, storage requirements or data scarcity. Since the dimension s is likely to be large in the context of digital watermarking, e.g. the size of a typical video frame is $s = 576 \times 704$, an approach based on the expectation-maximization (EM) algorithm will be exploited [11]. The PCA procedure is then reduced to an iterative algorithm using the following two steps:

$$\textbf{e-step:} \quad \mathbf{V} = (\mathbf{C}^{\text{T}}\mathbf{C})^{-1}\mathbf{C}^{\text{T}}\tilde{\mathbf{W}} \tag{5}$$

$$\textbf{m-step:} \quad \mathbf{C} = \tilde{\mathbf{W}}\mathbf{V}^{\text{T}}(\mathbf{V}\mathbf{V}^{\text{T}})^{-1} \tag{6}$$

where $.^{\text{T}}$ denotes the transposition operation. A major asset of this approach is that it can be performed *online* using only a single watermark estimate at a time, which reduces significantly storage

requirements. Moreover, the EM framework allows to deal with missing data, in our case watermark estimates having some non pertinent values that should be discarded as introduced in Subsection 3.1. The e-step only has to be slightly generalized so that missing information is also estimated accordingly to the current principal subspace estimation. For each incomplete watermark estimate $\tilde{\mathbf{w}}$, missing information is completed so that the distance to the current principal subspace is minimized i.e. the unique pair of points $\mathbf{v}^*$ and $\tilde{\mathbf{w}}^*$ - such that $\mathbf{v}^*$ lies in the current principal subspace, $\tilde{\mathbf{w}}^*$ lies in the subspace defined by the reliable information in $\tilde{\mathbf{w}}$ and the norm $\|\tilde{\mathbf{w}}^* - \mathbf{C}\mathbf{v}^*\|$ is minimized - is computed. The corresponding column of $\tilde{\mathbf{W}}$ (resp. $\mathbf{V}$) is then set to $\tilde{\mathbf{w}}^*$ (resp. $\mathbf{v}^*$) for the following m-step. At this point, the principal subspace has been estimated and a p-dimensional ortho-normalized basis $\{\mathbf{e}_i\}$ can be found e.g. the eigenvectors of matrix $\mathbf{C}$.

3.3. Watermark Removal

Once the PCA iterations are finished, the obtained vectors $\{\mathbf{e}_i\}$ span a subspace $\mathcal{E}$ which is assumed to be close to the watermarking subspace $\mathcal{W}$. At this point, the attacker basically wants to drain any energy contained in this estimated subspace $\mathcal{E}$. This can be easily ensured as follows:

$$\begin{aligned} \mathbf{c}_{\bar{\mathrm{w}}} &= \mathbf{c}_{\mathrm{w}} - \sum_{i=1}^{p} (\mathbf{c}_{\mathrm{w}} \cdot \mathbf{e}_i)\mathbf{e}_i \\ &\approx \mathbf{c}_{\mathrm{o}} + \sum_{i=1}^{p} \lambda_i \Big(\mathbf{w}_i - \sum_{j=1}^{p} (\mathbf{w}_i \cdot \mathbf{e}_j)\mathbf{e}_j \Big) \end{aligned} \quad (7)$$

where $\mathbf{c}_{\bar{\mathrm{w}}}$ is the resulting attacked digital content. If the watermarking subspace $\mathcal{W}$ has been finely estimated as defined in (4), the terms $\mathbf{d}_i = \mathbf{w}_i - \sum_{j=1}^{p} (\mathbf{w}_i \cdot \mathbf{e}_j)\mathbf{e}_j$ are null. In other terms, the attacker retrieves almost the original digital content $\mathbf{c}_{\mathrm{o}}$ and no pertinent watermark can be retrieved.

4. EXPERIMENTAL RESULTS

Experiments have been conducted with a collection of 500 images of size 512×512 to verify the previous theoretical assertion. This database has been watermarked with the schemes described in section 2 and the parameters have been chosen so that the embedding process introduces a distortion equal to 9 in terms of MSE i.e. a PSNR around 38 dB. Next, for the watermark estimation, a simple 5×5 averaging filter has been used for $\mathcal{L}(.)$ and practical results have shown that the threshold $\tau_{\text{discard}} = 8$ gives good results to isolate non-pertinent estimated watermark samples. Experimentally, the iterative approach was found to converge quite rapidly and 20 iterations were performed to obtain the vectors $\{\mathbf{e}_i\}$. Eventually, the average norm D of the vectors $\{\mathbf{d}_i\}$ has been computed to assert the effectiveness of the attack. If the watermarking subspace $\mathcal{W}$ has been perfectly estimated, D should be equal to 0. Alternatively, if the estimated subspace $\mathcal{E}$ is completely orthogonal to $\mathcal{W}$, the vectors $\mathbf{d}_i$ are equal to $\mathbf{w}_i$ and D is equal to 1. In other terms, the lower D is, the finer the watermarking subspace has been estimated and the more effective is the attack.

A typical experiment consists in watermarking the database with a given value for the dimension p of the watermarking subspace. Then, on the attacker side, a p-dimensional subspace is estimated[1] with the presented attack using only n images from the watermarked database and the average norm D is computed. The results obtained with the CDMA algorithm have been gathered in Figure 1. It basically illustrates how evolves the average norm D depending on the dimension p of the considered watermarking subspace and the number n of images considered to estimate the watermarking subspace $\mathcal{W}$. The darker the figure is, the greater the average norm D is. For a given subspace dimension, the more images are taken into account, the finer the private watermarking subspace is estimated i.e. the more efficient is the attack. It is a common feature of collusion attacks since observing more watermarked contents permits to extract more information about the hidden parameters, here the watermarking subspace $\mathcal{W}$. On the other hand, for a fixed number of images, the greater the dimension of the watermarking subspace is, the harder it is to have a good estimate of the watermarking subspace, which is also quite a natural result. However it should be noted that when 60 watermarks have been used, the proposed attack still permits to cut the average norm D down to 37% even if each watermark is then embedded with a quite low strength. In other terms, at least 63% of the watermark signal energy has been removed, which is usually enough to trap most detectors.

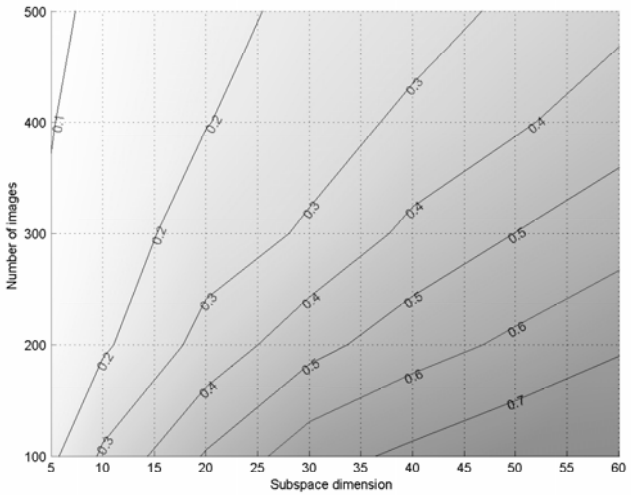

Fig. 1. Percentage of residual CDMA watermark energy after attack for several watermarking subspace dimensions p and a varying number n of images considered for collusion. The darker the image is, the more watermark energy is left i.e. the less efficient is the proposed attack.

The results for the other three watermarking schemes exhibit the same behavior even if they are not reported in this paper. Moreover, the impact of perceptual shaping has also been investigated. Digital watermarks can indeed be slightly modified according to the human perceptual system so that the inserted watermark is less perceptible. In the context of still images watermarks, the embedding strength can for example be made dependent of the local variance of the image. However such a perceptual shaping does not

[1] It has been assumed here that the attacker knows the dimension p of the watermarking subspace. If it is not the case, the attack can be performed with an arbitrary large value for p. Next, only eigenvectors associated with high eigenvalues are kept.

modify drastically the watermark to be embedded and the resulting shaped watermark is still highly correlated with the reference one. As a result, perceptual shaping does not strongly interfere with the estimation of the private watermarking subspace $\mathcal{W}$. The proposed attack is consequently still pertinent even if a small decrease of effectiveness is observed in comparison with the results obtained without perceptual shaping

5. CONCLUSION AND PERSPECTIVES

Many watermarking schemes rely on the same approach. First, a secret key is used to build a set of pseudo-random reference watermarks, which can be considered as a key-dependent private watermarking subspace. The watermark to be embedded is then constructed as a linear combination of those reference watermarks according to some parameters e.g. the payload to be hidden. Eventually, the resulting watermark is embedded with an additive scheme, possibly with perceptual shaping. The weakness of such watermarking schemes is that if an attacker succeeds in estimating the private watermarking subspace, he/she can completely remove the underlying watermarks. A collusion-based approach has been presented in this paper to achieve this goal, which relies on the principal component analysis of several watermark estimates obtained from distinct watermarked contents.

The reader might argue that the attacker needs to gather a large collection of contents watermarked with the same secret key, which can be unfeasible in practice. However, in a video context, each video frame can be regarded as a distinct watermarked content. This assumption is all the more pertinent since watermarking video is often considered as watermarking a sequence of still images [12]. Thus, considering a 1h30 watermarked movie, if the attacker draws a frame every 10 seconds, he/she can collect 540 frames watermarked with the same secret key and subsequently perform the attack. This emphasizes the issue of *intra-video* collusion in video watermarking. Some watermarking applications require a very high level of security and use secure cryptographic codes [13] to prevent users from colluding and producing unprotected content. However, it is useless to embed watermarks encoding such secure codes if the watermark signal can be stirred out with a signal processing attack.

Once again, this attack points out the need for informed watermarking and in particular informed coding. The weakness of the presented watermarking schemes is that, for a given secret key, the watermarking space is fixed once for all whatever the host content to be watermarked is. A possible countermeasure would be to have a watermarking subspace which is dependent on the host content e.g. dirty-paper watermarks. In this case, the dimension of the watermarking subspace is very large which virtually prevents an estimation attack. On the other hand, the detector has to search in a high dimensional space and an effective method has to be designed [14]. An alternative approach consists in modifying the watermark, drawn from a low-dimensional subspace, according to a key-dependent image signature [15]. As a result, since the detector can invert this modification, nothing has changed and the watermark seems to be in a low-dimensional subspace. On the other hand, when the attacker estimates the watermarks from several works, he/she sees them as if they were uniformly distributed over the whole image space.

6. REFERENCES

[1] I. Cox, M. Miller, and J. Bloom, *Digital Watermarking*, Morgan Kaufmann Publishers, 2001.

[2] N. Memon and P. Wong, "A buyer-seller watermarking protocol," *IEEE Transactions on Image Processing*, vol. 10, no. 4, pp. 643–649, April 2001.

[3] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *Security and Watermarking of Multimedia Contents II*, January 2000, vol. 3971 of *Proceedings of SPIE*, pp. 371–380.

[4] K. Su, D. Kundur, and D. Hatzinakos, "A novel approach to collusion resistant video watermarking," in *Security and Watermarking of Multimedia Contents IV*, January 2002, vol. 4675 of *Proceedings of SPIE*, pp. 491–502.

[5] T. Kalker, "Considerations on watermarking security," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, October 2001, pp. 201–206.

[6] J. Ó Ruanaidh and S. Pereira, "A secure robust digital image watermark," in *Electronic Imaging: Processing, Printing, and Publishing in Color*, May 1998, vol. 3409 of *Proceedings of SPIE*, pp. 150–163.

[7] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Processing*, vol. 66, no. 3, pp. 283–301, May 1998.

[8] B. Chen and G. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," in *Security and Watermarking of Multimedia Contents*, January 1999, vol. 3657 of *Proceedings of SPIE*, pp. 342–353.

[9] G. Doërr and J.-L. Dugelay, "Secure video watermarking via embedding strength modulation," in *Proceedings of the Second International Workshop on Digital Watermarking*, To be published, Lecture Notes in Computer Science.

[10] C. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner, and T. Pun, "Generalized watermarking attack based on watermark estimation and perceptual remodulation," in *Security and Watermarking of Multimedia Contents II*, January 2000, vol. 3971 of *Proceedings of SPIE*, pp. 358–370.

[11] S. Roweis, "EM algorithms for PCA and SPCA," *Neural Information Processing Systems*, vol. 10, pp. 626–632, 1998.

[12] G. Doërr and J.-L. Dugelay, "A guide tour of video watermarking," *Signal Processing: Image Communication*, vol. 18, no. 4, pp. 263–282, April 2003.

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

[14] M. Miller, G. Doërr, and I. Cox, "Applying informed coding and informed embedding to design a robust, high capacity watermark," *IEEE Transactions on Image Processing*, To be published.

[15] D. Delannay and B. Macq, "Method for hiding synchronization marks in scale and rotation resilient watermarking schemes," in *Security and Watermarking of Multimedia Contents IV*, January 2002, vol. 4675 of *Proceedings of SPIE*, pp. 548–554.