A CRYPTANALYTIC METHOD FOR EMBEDDING VIDEO WATERMARKS

Qian Zhang and Nigel Boston

Dept of Electrical and Computer Engineering
University of Wisconsin Madison
1415 Engineering Drive, Madison, WI 53706
Email: qianz@cae.wisc.edu, boston@engr.wisc.edu

ABSTRACT

Digital video watermarking is increasingly important. Video signals are very susceptible to attacks like frame averaging,dropping,swapping,collusion. This paper presents a new video watermarking method. The most important information to be watermarked such as a company's name is hidden in a sequence of statistical data, which is then embedded into each frame. Every segment of the statistical data is typically enough for the key watermark extraction. A very detailed example of this method is given, based on the cryptanalysis. The most important information is used as the key to encrypt a plaintext. The plaintext can be some less important information about the product. The ciphertext is then embedded in the video frames. With statistical knowledge of the English language, we can quickly recover the key. Mathematical analysis and simulation results are given in the paper.

1. INTRODUCTION

Still image watermarks have been extensively investigated, and many algorithms proposed.These still image watermarking schemes can be adapted to digital video watermarking. Videos, however, have large data redundancy, and successive frames are very similar, and so are more vulnerable to attacks like collusion, frame dropping, frame swapping, etc.. New watermarking schemes have been therefore proposed.[1] presents a scene-based and video watermarking scheme. The video sequence is segmented into the scenes, and the wavelet transform is employed along the temporal axis of the scene. The watermark can be detected without the knowledge of the location of the frames in the video scene. It is robust to colored noise, MPEG coding, frame dropping and printing and scanning, but it involves a large amount of buffering and computation.[2] proposes a method for video synchronization. The watermarking scheme consists of watermark embedder, temporal redundancy control, feature extractor, and key generator. The period α and the repeat β are defined first. In every period, the embedding procedure is to read a frame of the input video; send the current key K to the watermark embedder to watermark the current frame; repeat to apply the same key K to the next $\beta - 1$ frames; Then the feature extractor and key generator are used to create a new key K for subsequent frames. The same key K_E is embeded at the first frame of every period . The detection is based on the key K_E and the feature extraction. This method relies greatly on the feature extraction, and so may not be robust to some temporal attacks like frame average. A lot of computation is also required. [3] gives a theoretical framework for the linear collusion analysis of the watermarked video sequence. It says that the video watermark is statistically invisible if and only if the correlation coefficient between any two host frames is equal to that between the two corresponding watermarked frames. The two watermark design principles are that the second moment of the watermark scaling factors should be adapted proportionally to the variance of the host video frames; and that the correlation of the watermarks embedded into each pair of video frames should be matched to the correlation of the host frames themselves. The paper doesn't give a practical example satisfying the design principles. For instance, suppose we have a sequence of host video frames U_k, $k = 1, \ldots, n$, and watermarked frames $X_k = U_k + \alpha_k W_k$. We are trying to embed one watermark into this sequence, i.e. $\rho(W_a, W_a) = 1$, where $\rho(A, B) = \frac{cov(A,B)}{\sqrt{var(A)var(B)}}$. The question of how to generate the scaling factors α_k to satisfy: $\rho(U_a, U_b) = \frac{E\alpha_a\alpha_b}{E\alpha^2}$ for $\forall a, b \in \{1, 2, \ldots, n\}$ is not an easy matter.

This paper propses a new video watermark algorithm, which spreads a sequence of statistical data through the temporal axis. The key information is extracted from a sequence of statistical data. One example uses cryptanalysis. We have two levels of watermarking. The most important watermark such as the company's name or some data, time stamp is used as the key. Lesser information is used as the plaintext. Some encryption method like a shift cipher is employed. The ciphertext is embedded into the video sequence. At the decoder side, we extract the ciphertext first,

based on which we recover the key. This does not require much computation, and is robust to collusion attack, frame averaging, frame dropping, etc.. The paper is organized in the following way. A detailed description of our new watermarking system(WMS) is presented in section 2. Section 3 gives a mathematical analysis for attacks like frame dropping, frame averaging, collusion. Section 4 extends the watermarking system to multiple letters. Simulations are done in section 5, followed by the conclusion and further work in section 6.

2. SHIFT-CIPHER-BASED WMS

Suppose for simplicity that the most important watermark is one letter. The basic idea is to have it be the key to a shift cipher [4]. For example if the company U produces a video advertisement for Olympus cameras, it can use lesser information about itself or the video as the plaintext, its name "U" as the key, and so obtain the ciphertext. As well as claiming ownership through the main watermark, viewers can learn more information about the company or the video through the plaintext after decryption. Fig. 2 shows the flowchart.

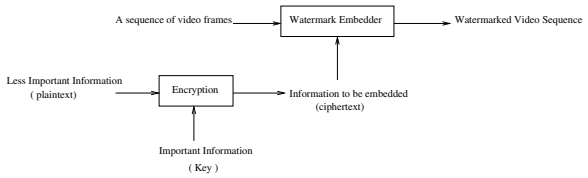

Fig. 1. The Flowchart of the watermarking system based on shift ciphers

We start with any watermarking scheme such as the spread spectrum watermark method, to embed the ciphertext into each video frame by this method [5] is robust to image scaling, JPEG compression, dither distortion, clipping, etc.. Here is the embedding process:

1. The Original video frame at time i in the video sequence is denoted as F_i.

2. The watermark set $\mathcal{W}$ consists of 26 orthogonal pseudo random noise signals, $\mathcal{W} = \{\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_{25}\}$. $\mathbf{W}_i$ has to satisfy

a). $\rho(\mathbf{W}_i, \mathbf{W}_j) = 0$ when $i \leq j$, $\rho(\mathbf{W}_i, \mathbf{W}_i) = \text{const} >> 0$.

b). $\rho(F_i, \mathbf{W}_i) = \frac{1}{X} << \rho(\mathbf{W}_i, \mathbf{W}_i)$

where ρ is defined as

$$\rho(X, Y) = \frac{1}{MN} X \cdot Y = \frac{1}{MN} \sum_{i,j} X(i,j) Y(i,j)$$

where X, Y are $M \times N$ matrices.

3. Encrypt the plaintext by the key watermark. The ciphertext sequence $c_1 c_2 \ldots c_n$ is embedded into each frame of the video sequence.

4. The watermarked video frame at time i is denoted as Fw_i. $Fw_i = F_i + \alpha W_{c_i}$,where α is the amplitude modulation parameter.

The decoding process is

1. The letter embedded in the frame is recovered by

$$w_i = \arg \max_{k \in \{0,\ldots,25\}} \rho(Fw_i, \mathbf{W}_k) \tag{1}$$

2. Analysize the sequence $\{w_0, w_1, \ldots\}$, and get the distribution of the letters in the decoded ciphertext, say $\{q_0, q_1, \ldots, q_{25}\}$. Suppose the distribution of the letters in plaintext is $\{p_0, p_1, \ldots, p_{25}\}$, The shift is recovered by

$$s = \arg \max_{k \in \{0,\ldots,25\}} \sum_{i=0,\ldots,25} p_i q_{i+k} \tag{2}$$

where $i + k$ is taken mod 26.

How long does the ciphertext have to be in order to successfully find the key? [4] Shannon proposes a concept of unicity distance. Empirical studies estimate that for English language the redundancy is 75%. The unicity distance for shift cipher is about 1.33.

3. ATTACKS

3.1. Collusion

Collusion is in general the use of more than one frame to obtain the watermark or the original data. In detail there are two types of collusions.

Collusion type I in [6] : The same watermark is embedded into different copies of different data. The collusion can estimate the watermark from each watermarked frame and obtain a refined estimate of the watermark by linear combination, e.g. the average, of the individual estimations.A good estimate of the watermark permits us to obtain unwatermarked data with a simple subtraction.

Collusion type II in [6]: Different watermarks are embedded into different copies of the same data. The collusion only has to make a linear combination of the different watermarked data, e.g. the average, to produce unwatermarked data. Indeed, generally, averaging different watermarks converges toward zero. We can recover the data, and so remove the watermark.

Shift-cipher based watermarking avoids collusion type I. The ciphertext conveys information about the key but different watermarks, i.e. ciphertexts are actually embedded into the different data. In video it is easy to put the linear combination of the n adjacent so similar frames into

the video sequence without any visual notice. $\bar{F}w \approx F + \alpha \sum_{\gamma \in \mathcal{E}} p_\gamma \mathbf{W}_\gamma$. The English letters are not uniformly distributed, there is no way to get the unwatermarked data so that type II collusion is avoided.

If we linearly combine a small number of the adjacent so similiar frames, i.e. n is small, then Denote by $\mathcal{A}$ the set of the English letters embedded in the n frames.

$$\begin{aligned} \rho(\bar{F}w, \mathbf{W}_\beta) &= \rho(F, \mathbf{W}_\beta) + \frac{\alpha}{n} \sum_{\gamma \in \mathcal{A}} n_\gamma \rho(\mathbf{W}_\gamma, \mathbf{W}_\beta) \\ &= \begin{cases} \frac{1}{X}\text{Const} + \alpha \frac{n_\beta}{n}\text{Const} & \text{if } \beta \in \mathcal{A}; \\ \frac{1}{X}\text{Const} & \text{if } \beta \notin \mathcal{A}. \end{cases} \end{aligned} \quad (3)$$

For instance if we linearly combine 7 frames, where OLYMPUS are embedded in these 7 frames. $\mathcal{A} = \{OLYMPUS\}$, $n_O = 1$,..., $n_S = 1$.

In the decoding process,we will see 7 peaks and so we are able to get the embedded letters.

3.2. Frame Dropping

By shift cipher cryptanalysis it is clear to see that to recover the key we only need a proportion of the ciphertext, and so dropping frames is not a successful attack. For the ciphertext, if we drop some frames we are still able to recover most of the information due to the redundancy of English. We can model the English language X as a first-order Markov Chain with transition matrix P,the deletion channel as a K-ary deletion channel. Let $X = \{x_1, \ldots, x_n\}$ be an input sequence, where $x_i \in \{1, \ldots, K\}$. The deletion process is i.i.d. distributed (with $P(D_i = 1) = p_d$) for the binary sequence D_i. The receiver receives $Y = \{y_1, \ldots, y_m\}$, where $m \leq n$.

Now consider the deletion process, denoted as Z. Actually the deletion process Z can be regarded as a Markov Chain. Every x_i has $2K$ states: $\{1, \ldots, K, \text{Del } 1, \ldots, \text{Del } K\}$. The transition matrix could be written as

$$P_z = \begin{bmatrix} (1-p_d)P & p_d P \\ (1-p_d)P & p_d P \end{bmatrix}$$

Assume that almost surely the process doesn't get stuck in deleted states, i.e. every recurrent class contains at least one non-deleted state. Then Y is a Markov chain, and its transition matrix is

$$P_Y = (1-p_d)P + p_d(1-p_d)P(I - p_d P)^{-1}P$$

The channel capacity is then $I(X;Y) = H(Y) - H(Y|X)$. It is very hard to calcuate $H(Y|X)$. [7] prosposed to utilize the reduced state trellis technique to upper bound $H(Y|X)$. [8] considers a symmetric 2-ary first order markov chain input sequence. Given an i.i.d. deletion channel, and a binary input alphabet, the lower bound of the capacity is obtained. [9] extends [8] by allowing the codewords of length N to consist of zeros or ones generated independently chosen from the distribution P, having length j with probability P_j.

3.3. Frame Swapping

It is clear that switching won't affect the statistical results of the English letter. At most of the switching is only limited in some successive frames, it is not hard to recover the less important information.

3.4. Substitution

Suppose an English plaintext of length N is encrypted by a shift cipher, and the ciphertext $c_1 c_2 \ldots c_N$ is embedded in the video. The distribution of the ciphertext letters is the same as the distribution of the plaintext letters, only shifted. In the N letters plaintext the frequencies of the letters $A, B, \ldots, Z$ are denoted as $N_0, N_1, \ldots, N_{25}$, where $N_i \approx p_i N$. In the N letters ciphertext we decode from the video frames the frequencies of $A, B, C, \ldots, Z$ are denoted $N'_0, N'_1, \ldots, N'_{25}$. Define

$$I_m = \frac{N_0 N'_{0+m} + N_1 N'_{1+m} + \ldots + N_{25} N'_{25+m}}{N^2}$$

The key is denoted as K.If all the embedded letters in the video frames are correctly decoded, we have $N'_{i+m} \approx p_{i+m-K} N$,$I_m = \sum_{i \in \{0,\ldots,25\}} p_i p_{i+m-K}$. Thus $I_{max} = I_K = \sum_{i \in \{0,\ldots,25\}} p_i^2$. If some attacks such as filtering, compression, addition of noise, cropping, quantization, A/D conversion, geometric distortion are applied to the watermarked video by the attackers, not all the embedded letters in the video frames can be correctly decoded. Suppose the probability of error for the embedded letter in each frame is p. The probability that the embedded letter i is decoded to another letter j is p_{ij}. We have$N'_{j+m} \approx p_{j+m-K} N(1-p) + \sum_{i \neq j+m} pN p_{i-K} p_{i(j+m)}$,

$$\begin{aligned} I(m) &= \sum_{i \in \{0,\ldots,25\}} p_i p_{i+m-K}(1-p) + \\ &\quad p \sum_{i \in \{0,\ldots,25\}} p_i \sum_{j \neq i+m} p_{j-K} p_{j(i+m)} \end{aligned}$$

$\sum p_i p_{i+m-K}$ is maximal when $m = K$. Whether we can successfully recover the key is decided by the sequence $\sum_{i \in \{0,\ldots,25\}} p_i \sum_{j \neq i+m} p_{j-K} p_{j(i+m)}$. In other words, the probability p_{ij} that the embedded letter i is decoded to another letter j under certain attacks plays a very important role here. For instance, if we can smartly design the watermark set $\mathcal{W}$, if one frame is not correctly decoded, the embedded letter is uniformly decoded to the other 25 letters, i.e. $p_{ij} = \frac{1}{25}$. $I(m) = \sum_{i \in \{0,\ldots,25\}} p_i p_{i+m-K}(1 - \frac{26}{24}p) + \frac{p}{25}$. It is obvious that $I_{max} = I_K$. It implies even if the

watermarked video undergoes some attack, we are still able to recover the key if we can properly design the watermark set $\mathcal{W}$.

4. VIGENERE-CIPHER-BASED WMS

The Shift-Cipher-based watermarking system can be extended to a watermarking system based on the Vigenere Cipher[4]. The multiples of the key length letters are embedded each video frame. The cryptanalysis methods of the Vigenere Cipher can be found in [4]. A statistical analysis similar to the shift-cipher-based watermarking system can be also applied here.

5. SIMULATION RESULTS

A plaintext with length 400 is encrypted by the key "U", and the ciphertext is embedded into a video which has 400 frames. The 400 frames undergo some attacks by Stirmark [10, 11]. Table. 5 lists some of the simulation results. The first column lists the attacks. The second column lists the percentage of the frames that are correctly decoded. The third column lists whether the key "U" is successfully decoded. It is easy to see that it is vulnerable to the median filter; but in those cases, the images after filters are blurred anyway. If frame dropping tests are run, where the frames

Attack	$p_c(\%)$	key Recover
NOISE_60	57	yes
NOISE_80	36	yes
NOISE_100	21	yes
MEDIAN FILTER_3	100	yes
MEDIAN FILTER_5	50.5	no
MEDIAN FILTER_7	18.75	no
JPEG_20	98.75	yes
JPEG_25	99.25	yes
PSNR_0	100	yes
PSNR_100	100	yes

Table 1. Stirmark attack

are randomly dropped, the key "U" can be successfully recovered even if only 100 frames are left. Frame swapping has no influence at all. If frame averaging is applied, where every three successive frames are averaged, the key "U" is correctly recovered.

6. CONCLUSION

This paper presents a new video watermarking system. It brings cryptanalytical, statistical methods into the watermarking field. It spreads the information along the temporal axis. The key information is hidden in a statistical data sequence, which is then embedded into the video frames. With the help of statistical analysis, we are then able to decode the key information. This article gives the simplest approach, employing shift ciphers, but other types of statistical embedding can be employed.

7. REFERENCES

[1] Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik, "Multiresolution scene-based video watermarking using perceptual models," *IEEE Journal on Special Areas in Communications 16(4)*, pp. 540–550, 1998.

[2] E. T. Lin and E. J. Delp, "Temporal synchronization in video watermarking," *IEEE Transactions on Signal Processing:Supplement on Secure Media*, to appear.

[3] D. Kundur K. Su and D. Hatzinakos, "Statistical invisibility for collusion-resistant digital video watermarking," *IEEE Transactions on Multimedia*, to appear.

[4] D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, 1995.

[5] Ingemar J. Cox, Joe Killian, Tom Leighton, and Talal Shamoon, "Secure spread spectrum watermarking for images, audio, and video," in *IEEE International Conference on Image Processing (ICIP'96) III*, 1996, vol. 4314, pp. 243–246.

[6] Gwenal Dorr and Jean-Luc, "Video watermarking overview and challenges," *Dugelay Multimedia Communications Image Group Eurcom Institute Sophia-Antipolis, France*.

[7] Aleksandar Kavcic and Ravi Motwani, "Insertion/deletion channels: Reduced-state lower bounds on channel capacities," in *ISIT Chicago U.S.A.*, July 2004.

[8] S. Diggavi and M. Grossglauser, "On transmission over deletion channels," in *Allerton Conference, Monticello, Illinois*, Oct 2001.

[9] Eleni Drinea and Michael Mitzenmacher, "On lower bounds for capacity of deletion channels," in *ISIT Chicago U.S.A.*, July 2004.

[10] Markus G. Kuhn Fabien A. P. Petitcolas, Ross J. Anderson, "Attacks on copyright marking systems," in *Information Hiding, Second International Workshop, IH98, Portland, Oregon*. April 1998, pp. 219–239, Springer-Verlag.

[11] Fabien A. P. Petitcolas, "Watermarking schemes evaluation," *I.E.E.E. Signal Processing*, vol. 17, pp. 58 – 64, September 2000.