A Novel Approach to Collusion-resistant Video Watermarking

Karen Su, Deepa Kundur and Dimitrios Hatzinakos

University of Toronto, Department of Electrical & Computer Engineering
10 King's College Road, Toronto, Ontario, Canada, M5S 3G4

ABSTRACT

This work considers the problem of frame collusion in video watermarking, one that is particularly relevant for this media due to the large collection of frames whose temporal inter-relationships may be exploited to facilitate estimation of the mark. Two new components are introduced: A mathematical framework for the statistical analysis of linear collusion and development of potential counterattacks; and a novel video watermarking approach employing the proposed strategies for robustness to collusion as well as other frame-as-image distortions. Experimental results demonstrating the performance of the proposed techniques against two types of collusion attacks are presented.

Keywords: Video watermarking, Collusion, Frame averaging, Feature extraction, Subframe watermarking.

1. INTRODUCTION

Although in their raw form video streams are simply sequences of image frames, the complexity of video processing and watermarking algorithms is greatly increased by the addition of the time dimension. However, the introduction of a third dimension also increases the flexibility and size of the solution space, and opens up a whole world of new ideas. Most of the academic and industrial interest in digital video watermarking has centered on the design of a copyright protection system for MPEG-2 coded video distributed on DVDs.[2] A video watermarking system had been designed by the Galaxy Group to complement the existing CSS that is part of the DVD standard; the technology is now called WaterCast and is being applied in the automatic monitoring of digital video broadcasts. Other applications of interest include digital TV transmission,[6] video on demand distribution,[8] and authenticating video surveillance for use as legal evidence.[1]

There remain many unexplored aspects of the video watermarking problem. In this work, we concentrate on resilience to an attack that is especially applicable to watermarks embedded into video sequences. Collusion occurs when collections of video frames are analyzed or combined with the ultimate goal of producing a mark-free copy of the original. The frames may form a temporally continuous subsequence, or come from greatly varied parts of the video. The key idea is the exploitation of temporal redundancy, either of the host video or the watermark, to estimate the redundant component. To date, the collusion attack has not been well studied, most probably because of the research focus on still image watermarking, where it does not arise in this form. However, its growing importance is evidenced by the publication of recent papers concentrating on collusion attacks.[7, 15]

We begin in Section 2 by studying video sequences from a frame-based statistical perspective and defining two types of collusion attacks. Analysis leading to the derivation of a counterattack to both types of collusion is then presented in Section 3. Based on this theoretical result, in Section 4 we develop and propose a novel approach to video watermark design. It is nicknamed SLIDE since it incorporates spatial localization with image-dependence. Two SLIDE watermark implementations are described in Section 5 and simulations demonstrating their performance are presented in Section 6. Finally the paper closes with a discussion of conclusions and potential areas of future work.

2. PROBLEM DEFINITION

In this section, we specify a frame-based statistical setup for analyzing video sequences and define the multiple frame collusion problem in this context. Section 2.1 describes the basic notation and in Section 2.2 two types of linear frame collusion attacks are studied.

Security and Watermarking of Multimedia Contents IV, Edward J. Delp III, Ping Wah Wong,
Editors, Proceedings of SPIE Vol. 4675 (2002) © 2002 SPIE · 0277-786X/02/$15.00

2.1. Nomenclature

The original or host video sequence is denoted $U_k(m,n)$, where k is a time or frame number index set, m and n are row and column indices respectively. The inputs to the embeddor are the host, a key K, and a binary data message vector V_i. The key is a sequence of bits encapsulating all parameters and secret components of the watermarking system, i.e., block sizes for block-based algorithms or seeds for random number generators. The embedded watermark signal $W_k(m,n)$ is defined over the same domain as the host $U_k(m,n)$, and is derived from the three inputs according to an embedding procedure.

The embeddor produces a watermarked video $X_k(m,n)$ sequence obtained by linear combination of the watermark with the host data $X_k(m,n) = U_k(m,n) + \alpha_k(m,n) \cdot W_k(m,n)$, where α represents a general scaling factor (i.e., local or global). Observe that no matter how the watermark is actually embedded, all data hiding procedures can be expressed in this form by defining the watermark as the difference between the watermarked and host signals (and setting α to 1 or some other appropriate function given the embedding algorithm).

The inputs to the blind watermark detector are the key K and a possibly watermarked signal $Y_k(m,n)$. We consider private-key watermarks, where the same key is used at both detector and the embeddor. The video sequence to be tested is expressed mathematically as $Y_k(m,n)=X_k(m,n)+E_k(m,n)$, where $E_k(m,n)$ is an error signal encapsulating both incidental and intentional distortions introduced into the video sequence. The watermark detector produces two outputs, C and $\hat{V}_i$, indicating the certainty of the watermark's presence and the extracted data message respectively. C is a value between 0 (watermark absent) and 1 (watermark present) indicating to what degree the watermark was detected. $\hat{V}_i$ is a binary vector of bits that can be compared to the embedded message V_i to measure the bit error rate of the system.

To study multiple frame collusion, we will be interested in the statistical relationships between the original host frames $U_k(m,n)$, watermark frames $\overline{W_k(m,n)}$, watermarked frames $X_k(m,n)$, and linear combinations defined in Section 2.2 as colluded frames $\overline{X(m,n)} = \sum_k \beta_k X_k(m,n)$. In particular we consider second-order statistics, i.e., the correlation coefficients among these entities, derive conditions on these that facilitate or prevent collusion attacks, and analyze how the watermark can be designed to achieve favourable coefficients.

2.2. Multiple Frame Collusion

What really makes video watermarking different from image watermarking is that there is much more data available both to the attacker as well as to the watermarker. Furthermore this data may be highly correlated; even making the assumption of spatially uncorrelated samples within each image frame, typical video sequences contain visual data that is strongly correlated along the temporal dimension. The class of attacks that applies in this case is known as *multiple frame collusion*. We define here two types of linear collusion attacks.

Type I linear collusion arises when large numbers of visually dissimilar video frames are marked via linear combination with a fixed watermark pattern. This is exactly the case for many existing video watermarks.[5, 9, 10] * Type II collusion arises when large numbers of visually similar frames are marked via linear combination with independent watermark patterns. This case is relevant, for instance, to video watermarks that use different 2D PN sequences to mark each frame.[11] We formalize the description of these attacks in the following definitions:

DEFINITION 2.1. *Given a set of watermarked video frames* $X_k = U_k + \alpha_k W_k,\ k = 1, \ldots, n,$ linear collusion *is the process of forming a linear combination of the frames*

$$\begin{aligned} \overline{X} &= \sum_{k=1}^{n} \beta_k X_k \\ &= \sum_{k=1}^{n} \beta_k U_k + \sum_{k=1}^{n} \beta_k \alpha_k W_k, \end{aligned}$$

such that $\overline{X}$ *provides an optimal MSE estimate of a) the watermark, or b) the host. In case of a), we refer to the attack as* Type I *collusion; if case of b), as* Type II *collusion.*

*Note: The watermarks do not have to be embedded in the spatial domain, the analysis presented here is relevant to watermark and host signals considered in their embedding domain.

We call this form of collusion *linear*, since it involves forming a linear combination of the watermarked video frames. Observe that Type I collusion is only possible if "the watermark" is a well-defined notion, i.e., if the same watermark is used to mark each video frame. Likewise, Type II collusion implies that "the host" is a well-defined notion; in this case we do not require that the host frames be identical, but they should be similar in the mean square sense, i.e., $\mathbf{E}[(U_a - U_b)^2] \approx 0$. Definition 2.1 encapsulates frame averaging attacks by setting $\beta_k = \frac{1}{n}$, as well as more sophisticated linear temporal filters by allowing β_k to take on arbitrary values. It also allows consideration of arbitrary sets of frames that are not necessarily in a temporally continuous sequence relative to the video.

Before moving on, we present a necessary condition for each type of linear collusion, in terms of the correlation between linear combinations of the watermarked and host frames. It will then become clear that protection from collusion can be achieved by imposing certain design criteria on the watermark. These criteria are formalized mathematically in Section 3 as statistical invisibility.

PROPOSITION 1. *Assuming that the watermarks W_k are independent of the host frames U_k, then a necessary condition for each of the two forms of linear collusion described in Definition 2.1 is given by*

$$\begin{aligned} \rho(\overline{X}, \overline{U}) &= 0 \text{ (Type I)} \\ \rho(\overline{X}, \overline{U}) &= 1 \text{ (Type II)}, \end{aligned}$$

where $\overline{U} = \sum_{k=1}^{n} \beta_k U_k$, i.e.,

$$\begin{aligned} \rho(\overline{X}, \overline{U}) \neq 0 \quad &\Longrightarrow \quad \text{Type I linear collusion is not possible, and} \\ \rho(\overline{X}, \overline{U}) \neq 1 \quad &\Longrightarrow \quad \text{Type II linear collusion is not possible.} \end{aligned}$$

Note that in the case of Type I collusion, if independent watermark patterns were used to mark each of the video frames, and if the necessary condition was met, then

$$\rho(\overline{X}, \overline{W}) = 1,$$

where $\overline{W} = \sum_{k=1}^{n} \beta_k \alpha_k W_k$. However, no information about the watermarks would be revealed since we could also write

$$\rho^2(\overline{X}, \overline{W}) = \sum_{k=1}^{n} \rho^2(\overline{X}, \beta_k \alpha_k W_k) = 1, \tag{2}$$

where all of the watermark terms on the right hand side of Equation 2 are unknown random variables.

Thus Proposition 1 gives only a necessary condition for Type I linear collusion attacks to be possible. The condition is not sufficient since if the watermark were designed such that independent patterns were embedded into each frame, then Type I collusion could be evaded.

When considering Type II collusion, we observe that if the same watermark pattern W were used to mark each of the video frames, and if the necessary condition was met, then

$$\rho(\overline{X}, \overline{U}) \approx \rho(\overline{X}, U) = 1,$$

where U is a representative of the set of host frames. (Recall that for Type II collusion, the "host" must be a well defined notion.) However, the colluded host would also include a scaled version of the watermark pattern:

$$\begin{aligned} \overline{X} &= \overline{U} + \overline{W} \\ &= U + W \sum_{k=1}^{n} \frac{\alpha_k}{n}. \end{aligned}$$

Therefore the collusion attempt would fail to separate the host from the watermark and a mark-free copy could not be obtained. Like in the case of Type I collusion, Proposition 1 gives only a necessary condition for collusion attacks to be possible. The condition is not sufficient since proper design of the watermark could provide protection against these attacks.

3. STATISTICAL INVISIBILITY

Having defined linear collusion, we now define statistical invisibility and show that it is precisely the property that is necessary to achieve robustness to Type I and Type II collusion attacks. The analysis leads to a theorem specifying the design of a statistically invisible collusion-resistant video watermark.

DEFINITION 3.1. *Given a sequence of host video frames U_k, $k = 1, \ldots, n$ and watermarked video frames $X_k = U_k + \alpha_k W_k$, we say that the video watermark W_k is* statistically invisible *if and only if the correlation coefficient between any two host frames a and b is equal to that between the two corresponding watermarked frames, i.e.,*

$$\rho(U_a, U_b) = \rho(X_a, X_b) \ \forall \ a, b \in 1, \ldots, n$$

Definition 3.1 states that given some correlation between two host frames U_a and U_b, the correlation between the two corresponding watermarked frames X_a and X_b should be the same. We refer to this property as statistical invisibility since an attacker analyzing the video sequence in a frame-by-frame manner does not observe any statistical difference between the host and watermarked sequences. It is exactly these statistical differences that could be exploited to form the colluded linear combination $\overline{X}$ in Proposition 1, thus enabling linear collusion.

Only the main points in the analysis will be highlighted here; full proofs can be found in.[14] Throughout, a number of assumptions may be made about the statistics of the watermark, host, and scaling factors:

(A1) the video frames share a common mean and variance (average power), i.e., $\mathbf{E}U_k = \mu_U$ and $\mathsf{var}(U_k) = \sigma_U^2$

(A2) the watermarks W_k are zero-mean and share a common non-zero variance $\sigma_W^2 > 0$

(A3) the scaling factors α_k share a second moment $\mathbf{E}\alpha^2$

(A4) the watermarks W_k are independent of the scaling factors α_k and the host U_k

We begin with an alternate interpretation of the statistical invisibility criterion.

PROPOSITION 3. *Under assumptions (A1), (A2), (A3), and (A4), a necessary and sufficient condition for the statistical invisibility of a video watermark is given by*

$$\rho(U_a, U_b) = \frac{\mathbf{E}\alpha_a \alpha_b}{\mathbf{E}\alpha^2} \rho(W_a, W_b) \ \forall \ a, b \in \{1, 2, \ldots, n\}. \tag{4}$$

Intuitively, Proposition 3 implies that in order for the watermarks embedded into two video frames to be statistically invisible, their correlation must differ from that of the host frames only by a scaling factor. In the trivial case, where a constant strength watermark is embedded into each frame, i.e., $\alpha_a = \alpha_b = A$, we require that $\rho(W_a, W_b) = \rho(U_a, U_b)$. In other words, highly correlated video frames should be watermarked with highly correlated watermark patterns, and vice versa. This is exactly a more precise mathematical statement of the hypothesis originally proposed by Swanson *et al.*[17]: that visually similar regions of video sequences should be marked with consistent watermarks.

Next, we present a sufficient condition for robustness to linear collusion; this expression will then enable us to show a direct relationship between statistical invisibility and collusion resistance.

PROPOSITION 5. *Under assumptions (A2) and (A4)*

$$\rho(\overline{X}, \overline{U}) = \rho(X_a, U_a) \ \forall \ a \in \{1, 2, \ldots, n\} \tag{6}$$

is a sufficient condition for robustness to linear collusion, i.e.,

$$\begin{array}{rcl} \rho(\overline{X}, \overline{U}) & = & \rho(X_a, U_a) \ \forall \ a \in \{1, 2, \ldots, n\} \\ & \Longrightarrow & \\ \forall \ \beta_1, \beta_2, \ldots, \beta_n \ s.t. \ \sum_{k=1}^{n} \beta_k \neq 0, & & \rho(\overline{X}, \overline{U}) \neq \{0, 1\}. \end{array}$$

Observe that Equation 6 implies the following condition:

$$\rho(X_a, U_a) = \rho(X_b, U_b) \ \forall \ a, b \in \{1, 2, \ldots, n\}. \tag{7}$$

At first glance this looks like quite a restrictive assumption. However, in Proposition 8 we show that not only is it easy to achieve, but that it is also a reasonable assumption given current trends in watermarking technique.

PROPOSITION 8. *Under assumptions (A2) and (A4)*

$$\begin{aligned} \rho(U_a, X_a) &= \rho(U_b, X_b) \\ &\Longleftrightarrow \\ \frac{\mathbf{E}\alpha_b^2}{\mathbf{E}\alpha_a^2} &= \frac{\mathit{var}(U_b)}{\mathit{var}(U_a)}. \end{aligned} \tag{9}$$

In other words, if the energy of the basic watermark signals W_k embedded into each frame are kept constant over the video sequence, modulating the per-frame embedding strengths $\mathbf{E}\alpha_k^2$ proportionally to the variances of the host frames ensures that the condition in Equation 7 is met. The idea of watermark strength adaptation according to some function of the image variance, both at global and local scales, is a popular rule of thumb used by many image watermarks.

Finally, we state our theorem on the relationship between statistical invisibility and multiple frame collusion. Theorem 3.2 also presents a practical criterion that exploits this relationship and enables the design of a collusion-resistant video watermark.

THEOREM 3.2. *Under assumptions (A1), (A2), (A3), and (A4), the following statements are equivalent:*

(1) $\rho(X_a, X_b) = \rho(U_a, U_b) \ \forall \ a, b \in \{1, 2, \ldots, n\}$,

(2) $\rho(U_a, U_b) = \frac{\mathbf{E}_{\alpha_a \alpha_b}}{\mathbf{E}_{\alpha^2}} \rho(W_a, W_b) \ \forall \ a, b \in \{1, 2, \ldots, n\}$, *and*

(3) $\rho(\overline{X}, \overline{U}) = \rho(X_a, U_a) \ \forall \ a \in \{1, 2, \ldots, n\}$.

Property (1) describes the statistical invisibility condition; property (2) defines a host-dependent watermark design criterion; and property (3) ensures that a watermark satisfying these criteria exhibits statistical resistance to Type I and Type II linear collusion attacks.

The main theoretical result is that to protect the watermark from linear collusion, the correlation of the watermarks embedded into each pair of video frames should be matched to the correlation of the host frames themselves. Although strictly correct only when the specified assumptions hold, we believe that the derivations presented here are still indicative of the behaviour to be expected under more general circumstances. Assumption (A2) does in fact hold for all existing spread spectrum watermarks, and assumption (A4) is reasonable as long as the watermark pattern is sufficiently large in terms of its spatial spread. Assumption (A3) can be more difficult to interpret, since it considers the statistics of the scaling factors. If a global scaling factor is used, it is clear that the assumption is valid. On the other hand, if the factors are locally derived from an image property like the NVF,[20] the analysis becomes more complicated.

4. SPATIALLY LOCALIZED IMAGE-DEPENDENT SUBFRAME WATERMARKING

Based on the concept of matched host-watermark correlation developed in the previous section, and given that it is also desirable that the watermark possess a relatively low complexity detection algorithm, and resistance to perceptually invisible geometric distortions, we have proposed a solution called Spatially Localized Image-DEpendent (SLIDE) subframe watermarking. In this section we will describe the essential novelty of the proposed framework and show how it achieves the desired goals.

4.1. Spatial Localization

The *spatial localization* property means that our watermark is specifically placed only into certain spatio-temporal regions of the video by design. We call the set of regions over which the watermark signal can take non-zero values the *watermark footprint*. The essence of our proposal is a spatially localized strategy, where the footprint does not cover all of the pixels in the video sequence. This is quite a different approach from that taken by most of the currently proposed schemes, in which the watermark's energy is spread globally throughout all of the available space. We distinguish this notion of a spatially localized watermark from another class of marks, also with localized footprints, known as *region of interest watermarks*. For instance, Su *et al.*[16] propose a watermark that is embedded only into regions of interest that must be pre-selected by the owner of the document. In contrast, the footprint of our spatially localized mark is automatically generated, which makes it appropriate for use in video watermarking, where there is a very large number of frames to be processed. The only known existing proposal for a spatially localized watermark is presented by Brisbane *et al.*[4] The approach taken to defining the watermark footprint is based on feature-oriented segmentation and region growing, which is believed to have a higher complexity than the proposed algorithm (due to the additional growing step).

To motivate our proposed watermarking framework, we make the following argument for why it might be advantageous to use a spatially localized footprint. To answer this question we must look at the correlation properties of watermarks with global and local footprints. When we consider the case of typical spread spectrum watermarks with global footprints,[9–11] we see that

$$\rho(W_a, W_b) = \begin{cases} 1, & W_a = W_b \\ 0, & \text{otherwise.} \end{cases}$$

Either the same PN pattern is used in each video frame, or an independent signal is generated for each. In these cases, the watermark's ability to adjust its correlation to match that of the host video is minimal.

Consider now the case of a spatially localized watermark, with a footprint structure that is comprised of non-overlapping subframes centered around a set of anchor points selected from the frame, and with the same basic PN watermark pattern W_s embedded into each subframe. The basic pattern is smaller in size than the full-frame watermark, but is otherwise the same, i.e., having zero-mean and variance σ_W^2. Taking two arbitrary frames, we can then see that the correspondence between the two sets of anchor points, A_s^a and A_s^b, plays a key role in controlling the correlation of the watermarks

$$\begin{aligned} \rho(W_a, W_b) &= \frac{\mathbf{E} W_a W_b}{\mathsf{var}(W)} \\ &= \frac{\sum_{l=1}^{L} \mathbf{E} W_s(m,n) W_s(m - [m_{l,b} - m_{l_a}], n - [n_{l,b} - n_{l,a}])}{L\sigma_W^2} \\ &= \frac{L'}{L}, \end{aligned} \tag{10}$$

where $L' = |A_s^a \cap A_s^b|$ is the cardinality of the intersection of the two sets of anchor points, i.e., the number of points at which $m_l^b - m_l^a = n_l^b - n_l^a = 0$ and therefore the corresponding numerator terms in Equation 10 will be non-zero, and $L = |A_s^a| = |A_s^b|$ is the total number of anchor points. For the moment, we assume that the two frames have the same number of anchor points.

Based on the number of subframes that are selected in each frame, the correlation of the overall watermark patterns can be adjusted. The resolution of these adjustments is limited to discrete steps, however it is clear that an improved collusion resistance can be obtained using a spatially localized framework, provided that the cardinality of the intersection set of the anchor points is directly proportional to the correlation of the underlying host frames. This requirement brings us to the *image-dependent* part of the proposed framework.

4.2. Image-dependent Subframe Placement

When we consider the notion of correlation between two host video frames, we can think of it as a statistical similarity measure. In order for the relative positions of the anchor points to vary according to the correlations between video frames, we propose that their locations be chosen based on image-dependent visual criteria. In this manner, similar sets of anchor points would be extracted from visually similar frames, resulting in a correspondence between large anchor point intersection sets and high host correlations. Conversely, with a good extraction algorithm, we expect that the sets of anchor points extracted from host frames with a low correlation will not share many common points.

We draw a link between the ideas behind the current work and a recent publication on an optimal image collusion attack.[13] In contrast to the definition of linear collusion proposed here, the authors define an image collusion attack as one where a number of copies of a watermarked document are obtained and a filtered linear combination of these is formed. The goal is to ensure that none of the originally marked documents can be identified by analyzing the attacked copy. A mechanism that is considered for combatting such an attack is *collusion-secure watermarking*. One of the key points in collusion-security is that when the modifications made to identical copies of a document are the same, no watermark information can be detected through further analysis. This result supports our goal of watermarking identical or highly similar video frames in the same manner.

In the design of a subframe watermarking scheme as proposed in Section 4, one of the main questions to consider is that of where to place the subframes. We have already alluded to the fact that the subframes should be located about image invariant features in order to provide collusion resistance. It is also desirable that we choose regions that have good properties for watermarking. We have developed an interpolation attack channel model that bounds the magnitude of the image-dependent additive spatial distortion signal arising from a certain class of attacks. We then propose a modified feature extraction algorithm that favours features located in regions where this distortion bound is small. Finally, we select the subframe anchor points from this set of extracted features. Due to space constraints, the reader is referred to[14] for more details.

5. PROPOSED WATERMARKING SYSTEM

Now, we turn our attention to the precise implementation details of two watermarks that we have developed based on the concepts introduced previously. The essential novelty of the work is twofold:

- First of all, the energy of the watermark is concentrated into a spatially localized footprint composed of regularly shaped subframes. The watermark payload is embedded into each subframe independently. This spatial diversity makes it more resilient to attacks that involve removing parts of the frame, such as cropping and row/column deletion. In comparison, a watermark like CDMA, with a global footprint, is very sensitive to any loss of spatial information.
- Secondly, the proposed framework uses image-dependent or content-based criteria to synchronize the subframes. The idea of locating the watermark relative to the content of the image is entirely new in watermarking. It eliminates the need for absolute spatial or temporal markers, like start of frame positions, and also enables the watermark to deal with attacks whose effects are not homogeneous throughout the frame. An example of such an attack is one that translates the pixels in the left half of the frame down by one position and those in the right half up by one. This imperceptible distortion would be much more effective against systems like JAWS, that require spatial synchronization, than against the proposed algorithms.

The watermark patterns themselves are designed using a number of concepts from the image watermarking literature. That these concepts can be so easily applied in a collusion-resistant video watermarking framework is another strength of the proposed approach.

5.1. The Watermark Embeddor

Our goal is to embed a watermark that is resilient to linear collusion and also robust to attacks that treat video frames as images. As is implied by the analysis, we consider a frame-by-frame approach to processing the video, and use visual content to modulate the placement of the watermark. To achieve this, the first step in the embedding process is footprint generation. It begins by extracting a set of anchor points about which watermark subframes are then embedded. The next task is basic pattern generation, in which one of the two watermark subframe patterns that we have tested are constructed:

- The first is a simple spatial domain spread spectrum pattern. Because of its subframe-oriented nature, there are clear analogies between this approach and the JAWS system. The main difference is that in JAWS, the subframes are regularly tiled, whereas in the proposed approach, their locations are synchronized according to the visual content of the frame.
- The second proposal is designed to achieve better robustness to geometric distortions. To achieve this goal, we construct the watermark in the DFT magnitude domain, where the effect of affine transformations applied to the spatial representation is well-defined.[3] In particular, a rotation in the spatial domain results in a rotation of the same angle and in the same direction of the coefficients in the DFT magnitude domain.

After the basic watermark pattern has been generated, it is convolved with the set of extracted feature points to form a frame-sized watermark. The extraction algorithm ensures that the features are appropriately spaced to avoid subframe overlap. Then spatial masking is applied to the watermark frame to modulate its strength locally, according to the properties of the video frame itself. Note that in the case of the second proposed pattern, the watermark is not embedded in the DFT domain. The pattern is constructed in the DFT domain, and then its inverse transform is computed and perceptual masking is performed in the spatial domain. Finally, the scaled watermark is embedded by addition to the host. The five main steps of the proposed embedding algorithm are illustrated graphically in Figure 1.

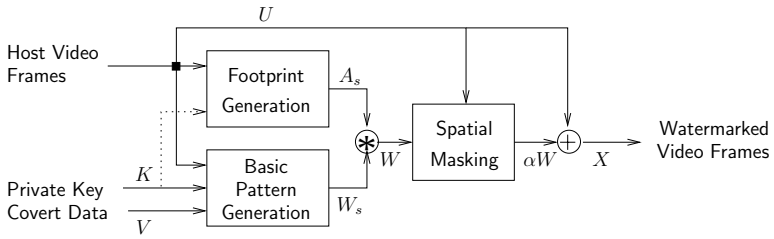

Figure 1: Block diagram of proposed watermark embeddor.

5.2. The Watermark Detector

The first step in the detection process is to estimate the locations of the subframes so that we can then proceed with watermark detection. To this end, we begin by forming the watermark footprint as at the embeddor. Observe that the detected footprint is denoted $\hat{A}_s$ to emphasize that it will not necessarily be identical to the footprint selected at the embeddor. Next, we recall that the NVF was used to locally scale the watermark's strength after generating the full-frame pattern. Thus at the detector we estimate the scaling factors from the watermarked frame and attempt to unscale the pattern to facilitate detection. From a communications perspective, the local scaling factors act as a multiplicative noise and the unscaling operation corresponds to a deconvolution. After generating the reference component of the basic watermark pattern, we can proceed with detection and extraction. The five main steps in the proposed algorithm are illustrated in Figure 2.

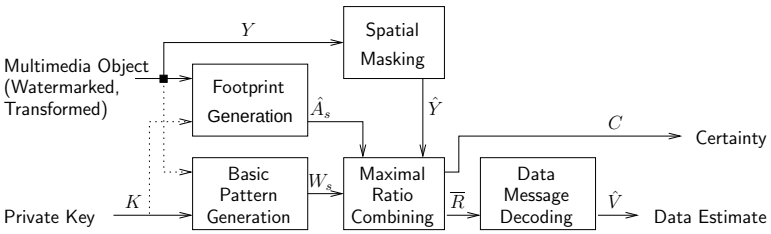

Figure 2: Block diagram of proposed watermark detector.

6. SIMULATION RESULTS

To test the collusion-resistant properties of the proposed schemes, we consider an accelerated Type I collusion attack in Section 6.1 and a Type II attack in Section 6.2. In the first of these, we estimate the noise in each watermarked frame and then combine these estimates to form an enhanced overall estimate of the watermark. The algorithms are tested against our implementation of JAWS,[10] chosen because of its current use in commercial applications, as well as the similarity between the embedding concepts employed. For instance, all three methods apply the watermark in a frame-by-frame manner, dividing each video frame spatially into subframes (or tiles). However, whereas in JAWS the subframes are regularly tiled and synchronized relative to structural properties, i.e., the top left corner of the frame, in our proposal their centres are synchronized relative to visual features and hence they are irregularly located. The second attack is conducted by averaging adjacent frames of a real video sequence after watermarking to obtain a mark-free copy of the host. JAWS is resistant to this attack, therefore we contrast the performance of the proposed schemes to that of the CDMA watermark in this case. The results demonstrate that the proposed schemes are robust against both Type I and II linear collusion.

Because of the limited resolution of the DFT domain approach (discussed in[14]), a $k = 11$ bit payload is used in testing both proposed schemes. Also, a subframe side width $s = 81$ was found experimentally to give a good performance tradeoff between robustness and data rate. Our implementation of JAWS uses tile sizes of $M = 128$, and a detection threshold of $T = \frac{5}{M}$.[10] It also has a maximum payload size of $k = 8$ bits. We found that this detector resulted in decoding failures when more than two correlation peaks exceeded the threshold in magnitude. Therefore, results obtained using a threshold of $T = \frac{15}{M}$ and a *no detection threshold* version are also shown for illustrative purposes. In this case, only the maximum and minimum correlation coefficients are considered, regardless of how many others also exceed the threshold. Thus the best case performance is achieved. However, we note this approach leads to an impractical false positive rate of 1. Finally, to ensure a fair comparison, the strengths of all watermarks are adjusted to a PSNR of 38 dB after embedding all of the implementations have comparable false positive rates lower than $1.0(10^{-6})$.

6.1. Type I Linear Collusion

In this section we consider a Type I linear collusion attack applied to a sequence of frames over which the visual content varies greatly. Recall that for this attack the sequence need not be temporally continuous. We also note that in a typical video there are 25 to 30 frames per second, and for instance in an action film, the scenes or shots may change dramatically every 0.5 seconds, thus making such sequences easy to construct.[19] Having gathered such a sequence, we first attempted to obtain an estimate of the watermark by averaging the N frames directly. This sort of attack is expected to be particularly effective against schemes like JAWS, in which the same watermark pattern is embedded additively into each frame. What we found was that even averaging over frame sets of size $N = 250$, the watermark pattern could not easily be estimated. One reason for this result may be its very small power compared to that of the content of the frames themselves, i.e., the component that we are trying to average out.

Therefore, since the purpose of Type I collusion attacks is to estimate the watermark, we propose a modified attack in which the watermark W_i is first estimated from each frame Y_i using a procedure which will be described in the next paragraph. We assume that these frame-based estimates are not of a sufficient accuracy, such that their subtraction from the frame would result in failure of watermark detection. However, we can enhance

these estimates and form a colluded approximation of the watermark pattern $\overline{W}$ by averaging them over N frames. For each frame to be attacked, this signal is then modulated according to the details of the embedding algorithm, i.e., using the NVF for the proposed approaches and a high-pass filter for JAWS. Finally, we scale the power of the modulated signal (globally) to achieve a distortion PSNR of 38 dB and *subtract* it from the original frame to obtain an attacked copy $\tilde{Y}$. A block diagram illustrating the steps in this modified attack is presented in Figure 3.

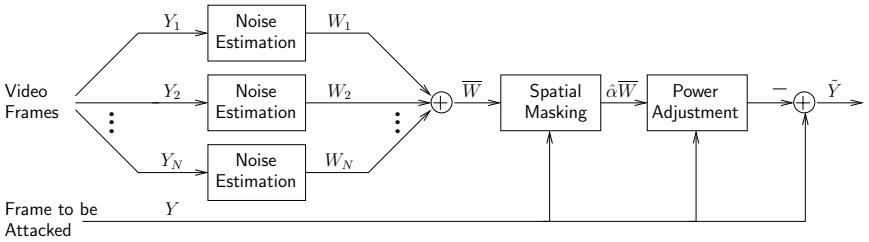

Figure 3: Block diagram of modified Type I collusion attack.

Since watermarks are noise-like signals, one simple way to attempt to remove them from an image is by denoising; this general concept was first put forward for applications in watermarking attacks and design by Voloshynovskiy *et al.*[20] Anisotropic diffusion is a popular image processing technique that can be used to remove noise from an image.[12, 18] The image is treated as the initial condition to a heat equation and "cooled" according to a set of image-dependent conduction coefficients. To achieve noise reduction while preserving edges, larger coefficients are used in windows with low gradients (faster cooling), and smaller coefficients in those with high gradients (slower cooling). Specifically, a 3×3 window is used for computing the local gradients,[12] and the coefficients are chosen heuristically from a discrete set based on these gradients. In the modified collusion attack, we diffuse each frame Y_i to obtain a more noise-free copy, and then take the difference between this copy and the watermarked frame W_i to be an estimate of the watermark itself.

Finally, we present results from the test applied to the fish_c2 video sequence is used; every 5^{th} frame was extracted to form the actual set of test frames. We watermarked this sequence using both JAWS and the two proposed schemes. Then the modified Type I collusion attack was applied to obtain an attacked copy of the first frame of the video sequence. The resulting bit error rates are shown in Figure 4. We can see that as the number of frames being combined increases, the performance of the JAWS system degrades. This behaviour occurs since with each additional estimate, the strength of the component of the overall estimate that corresponds to the non-time-varying watermark pattern is being enhanced. In contrast, in both proposed schemes, the best estimate of the watermark is obtained when only one frame is used in the estimation procedure. This behaviour can be attributed to the time-varying nature of the watermark. As successive estimates are added to the combination, the overall estimate becomes more like noise than like any of the actual watermark patterns.

6.2. Type II Linear Collusion

The Type II collusion attack that we consider is basically a two-tap unweighted MA filter operating along the temporal axis. We begin with a sequence of 10 consecutive frames extracted from the beginning of the hawk3 test video. These frames correspond to a relatively still scene, which makes them appropriate for frame averaging. In a more general implementation of a collusion attack, the attack module may choose to work with collections of consecutive frames rather than only two. It should also compute some metric comparing the content of these collections in order to determine whether Type I or Type II collusion would be more effective. In our simple attack, the sequence of frames is watermarked using the two proposed and CDMA schemes. Then the filter is applied, resulting in an attacked video sequence that is 9 frames in length. Finally, we attempt to detect the watermark from the attacked video. As is expected from considering the mathematical principles underlying the three schemes, the CDMA watermark is effectively removed by frame averaging, while both proposed watermarks remain intact.

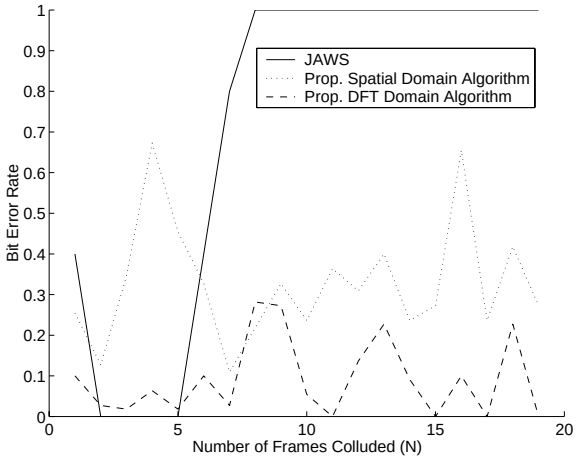

Figure 4. Bit error vs. number of frames used for collusion in modified Type I attack for the proposed DFT and spatial domain algorithms, and JAWS. The embedding and attack PSNRs are fixed at 38 dB.

In Figure 5, we show a pair of frames from the video, as well as their CDMA watermarked copies, and the averaged frame from which detection failed. This illustration shows that the frame averaging attack successfully removes the watermark without significantly damaging the visual quality of the video frames. Although frame averaging is not suitable for all pairs or collections of frames, we can see that for still scenes and certain types of video watermarks, it can be a simple yet effective attack.

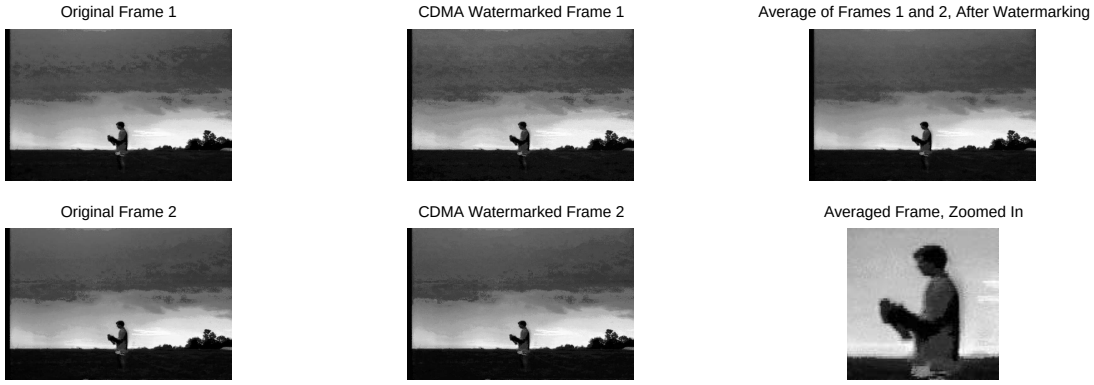

Figure 5. Illustration of sample test frames. Original frames 1 and 2 from hawk3 video (top and bottom left), frames 1 and 2 watermarked using CDMA (top and bottom middle), averaged frame constructed from watermarked frames 1 and 2 (top right), and zoom in of averaged frame to show that the attack does not significantly damage visual quality although the watermark is removed (bottom right).

7. CONCLUSIONS

In this paper, we present a novel spatially localized image-dependent framework for collusion-resistant video watermarking. A few useful components are added to the watermarking research toolkit:

- the development of a statistical analysis of collusion and the derivation of equations describing properties that a watermark should possess in order to resist such attacks, i.e., *statistical invisibility*,
- the notion of a watermark's *footprint*, the spatio-temporal co-ordinates over which its energy is spread,

- the use of a spatially localized footprint with a compact description based on a set of feature points - we note however, that there is a tradeoff in terms of spatial redundancy with non-global watermarks, and
- the mechanism of *content-based synchronization*, whereby image-dependent properties rather than absolute spatio-temporal markers, are used to perform watermark synchronization automatically.

Finally, we have proposed two new video watermarking schemes that are distinguished by their ability to be embedded and extracted using frame-based algorithms, while resisting collusion. Two different noise-like patterns are tested for robustness to a variety of attacks. The first is a PN sequence directly embedded into the spatial domain and the second is constructed by taking the inverse transform of a pattern of peaks from the DFT magnitude domain; the resulting signal is a sum of 2D DFT basis functions. We find that the spatial domain approach out-performs the DFT for severe JPEG compression. However, the DFT domain approach is more robust to general attacks, such as small-angle rotations.

REFERENCES

1. Richard Barnett. Digital watermarking: Applications, techniques, and challenges. *Electronics and Communication Engineering Journal*, 11(4):173–183, August 1999.
2. Jeffrey A. Bloom, Ingemar J. Cox, Ton Kalker, Jean-Paul M. G. Linnartz, Matthew L. Miller, and C. Brendan S. Traw. Copy protection for DVD video. *Proceedings of the IEEE*, 87(7):1267–1276, July 1999.
3. Ronald N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill Series in Electrical and Computer Engineering. McGraw-Hill Companies, Inc., 2000.
4. Gareth Brisbane, Rei Safavi-Naini, and Philip Ogunbona. Region-based watermarking for images. *Lecture Notes in Computer Science*, 1729:425–435, October 1999.
5. Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. on Image Processing*, 6(12):1673–1687, December 1997.
6. V. Darmstaedter, J.-F. Delaigle, D. Nicholson, and B. Macq. A block based watermarking technique for MPEG2 signals: Optimization and validation on real digital TV distribution links. In *Proc. Euro. Conf. on Multimedia Applications, Services and Techniques*, pages 190–206, 1998.
7. Frédéric Deguillaume, Gabriela Csurka, and Thierry Pun. Countermeasures for unintentional and intentional video watermarking attacks. In *Proceedings of the SPIE*, volume 3971, January 2000.
8. Carsten Griwodz, Oliver Merkel, Jana Dittmann, and Ralf Steinmetz. Protecting VoD the easier way. In *ACM Multimedia*, pages 21–28. 1998.
9. Frank Hartung, Peter Eisert, and Bernd Girod. Digital watermarking of MPEG-4 facial animation parameters. *Computers and Graphics*, 22(4):425–435, July-August 1998.
10. Ton Kalker, Geert Depovere, Jaap Haitsma, and Maurice Maes. A video watermarking system for broadcast monitoring. In *Proceedings of the SPIE*, volume 3657, pages 103–112, January 1999.
11. Bijan G. Mobasseri. Exploring CDMA for watermarking of digital video. In *Proceedings of the SPIE*, volume 3657, pages 96–102, January 1999.
12. Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, July 1990.
13. Jonathan K. Su, Joachim J. Eggers, and Bernd Girod. Capacity of digital watermarks subjected to an optimal collusion attack. In *Proceedings European Signal Processing Conference*, 2000.
14. Karen Su. Digital video watermarking principles for robustness to collusion and interpolation attacks. Master's thesis, University of Toronto, September 2001.
15. Karen Su, Deepa Kundur, and Dimitrios Hatzinakos. A content-dependent spatially localized video watermark for resistance to collusion and interpolation attacks. In *Proc. Int'l Conf. on Image Processing*, 2001.
16. Po-Chyi Su, Houng-Jyh Mike Wang, and C.-C. Jay Kuo. Digital image watermarking in regions of interest. In *Proceedings IS&T Processing/Image Quality/Image Capture Systems (PICS)*. April 1999.
17. Mitchell D. Swanson, Bin Zhu, and Ahmed T. Tewfik. Multiresolution scene-based video watermarking using perceptual models. *IEEE J. on Sel. Areas in Comm.*, 16(4):540–550, May 1998.
18. F. Torkamani-Azar and K. E. Tait. Image recovery using the anisotropic diffusion equation. *IEEE Trans. on Image Processing*, 5(11):1573–1578, November 1996.
19. Nuno Vasconcelos and Andrew Lippman. Statistical models of video structure for content analysis and characterization. *IEEE Trans. on Image Processing*, 9(1):3–19, January 2000.
20. Sviatoslav Voloshynovskiy, A. Herrigel, N. B., and Thierry Pun. A stochastic approach to content adaptive digital image watermarking. *Lecture Notes in Computer Science*, 1768:212–236, September 2000.