Spatially Localized Image-Dependent Watermarking for Statistical Invisibility and Collusion Resistance

Karen Su, *Student Member, IEEE*, Deepa Kundur, *Senior Member, IEEE*, and Dimitrios Hatzinakos, *Senior Member, IEEE*

***Abstract*—In this paper, we develop a novel video watermarking framework based on the collusion-resistant design rules formulated in a companion paper. We propose to employ a spatially-localized image dependent approach to create a watermark whose pairwise frame correlations approximate those of the host video. To characterize the spread of its spatially-localized energy distribution, the notion of a *watermark footprint* is introduced. Then we explain how a particular type of image dependent footprint structure, comprised of subframes centered around a set of visually significant anchor points, can lead to two advantageous results: pairwise watermark frame correlations that more closely match those of the host video for statistical invisibility, and the ability to apply image watermarks directly to a frame sequence without sacrificing collusion-resistance. In the ensuing overview of the proposed video watermark, two new ideas are put forward: synchronizing the subframe locations using *visual content rather than structural markers* and exploiting the inherent spatial diversity of the subframe-based watermark to improve detector performance. Simulation results are presented to show that the proposed scheme provides improved resistance to linear frame collusion, while still being embedded and extracted using relatively low complexity frame-based algorithms.**

***Index Terms*—Image feature extraction, linear collusion, robust digital video watermarking, statistical invisibility.**

I. Introduction

In this work, we motivate and theoretically address the problem of collusion-resistant video watermarking. Our analysis leads to a new theorem whose implications can be summarized by the following two assertions: In order for a video watermark to be resistant to linear collusion attacks, it must be *statistically invisible*. To achieve statistical invisibility, the watermark frames can be constructed such that their inter-correlation coefficients are matched to those of the underlying video frames, i.e., $\rho(U_a, U_b) = \gamma\rho(W_a, W_b)$ where U_k is the host, W_k is the watermark to be embedded in U_k, and γ is a constant related to the watermark scaling parameters discussed in [1].

The scope of this paper is restricted to the development of a framework that is more robust to collusion resistance in order to more fundamentally identify the issues that characterize this problem. We focus on the design and implementation of a novel video watermarking algorithm that achieves better statistical invisibility than a number of existing schemes. That is, its pairwise watermark frame correlations correspond more closely to those of the video frames into which the marks are embedded. Although, our watermarking proposal is a straightforward implementation based on spread spectrum watermarking to demonstrate the validity of our framework, we believe that our methodology is flexible enough to allow for improved robustness to other types of attacks such as geometric distortions.

We begin by considering the statistical invisibility properties of a few well-known video watermarks. Next, in Section III we define the notion of a watermark footprint and propose a new spatially localized structure based on irregularly aligned nonoverlapping subframes. By using such an approach, the correlation coefficient between two watermark frames can be modulated by changing the alignment of the subframe locations selected in each frame. Further, by applying image-dependent criteria in selecting the subframe locations, the correlations are then modulated according to the visual content of the underlying frames. Section IV presents a new watermarking proposal, built within this Spatially Localized Image-DEpendent (SLIDE) framework. The performance of the watermark is evaluated in Section V; we describe simple Type I and Type II collusion attacks and show how they can be used to estimate the watermark or host respectively. Applying these tests, results comparing the error rates achieved by the watermark to those obtained by a few existing approaches are shown. The paper concludes in Section VI by discussing some areas of future work in improving the effectiveness of the SLIDE approach.

II. Previous Work

In this section, we study the statistical invisibility properties of some well-known video watermarks to provide a context for the proposed work. We note that many of the existing algorithms are not designed for collusion-resistance.

The first video watermark that we consider is the one-dimensional spread spectrum approach by Hartung *et al.* [2]. Here, the watermark is a pseudo-random sequence spread over the video frames by direct spatial domain addition. The watermark is repeatedly embedded throughout the video in a sequential manner. Because of the discrete nature of the spreading operation, the number of frames over which the watermark sequence is spread can be expressed as a rational number

$$m = \frac{\#\text{ pixels used to embed one copy of the watermark}}{\#\text{ pixels per frame}}.$$

Manuscript received October 24, 2002; revised July 11, 2003. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) and by the Communications and Information Technology Ontario (CITO). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anna Hac.

K. Su was with the University of Toronto, Toronto, ON, Canada. She is now with the Laboratory for Communication Engineering, University of Cambridge, Cambridge, U.K.

D. Kundur was with the University of Toronto, Toronto, ON, Canada. She is now with the Electrical Engineering Department, Texas A&M University, College Station, TX 77843-3128 USA (e-mail: deepa@ee.tamu.edu).

D. Hatzinakos is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada.

Digital Object Identifier 10.1109/TMM.2004.840614

Then, the watermark frame correlations $\rho(W_i, W_j)$ can take two possible values: 1 when $j - i$ is an integer multiple of m, and 0 otherwise. In particular, observe that if $m = 1/n$ (where $j - i$ is always an integer multiple of m), the correlation is always equal to 1. Like many other proposed schemes, the image content is not taken into consideration, beyond supporting an optional local scaling factor.

Higher dimensional pseudo-random patterns have also been proposed for use in video watermarking applications. Just Another Watermarking System (JAWS) is a two-dimensional (2-D) watermark proposed by Kalker *et al.* to enable monitoring sites to verify and track video data transmitted over broadcast links [3]. It marks each video frame by adding it to the same noise-like pattern, thus $\rho(W_i, W_j) = 1 \; \forall \, i, j$.

In [4], Mobasseri proposes a fundamentally different scheme based on replacement rather than addition. Each video frame is decomposed into bitplanes, e.g., for an 8-bit gray-scale video there would be 8 bitplanes per frame. The video is marked by replacing one of the four least significant bitplanes of each frame with a watermark plane (pseudo-random binary pattern). The bitplane replacement procedure can also be expressed as an addition as follows: Assuming that the values in each of the four least significant bitplanes and in the watermark plane are uniformly distributed, the additive watermark can be modeled statistically as $W(m, n) = 2^k \cdot \{-1, 0, +1\}$ with probabilities {1/4, 1/2, 1/4} respectively, where k is the bitplane number, counting from least to greatest significance and starting at 0. If we further assume that in any two video frames, the four least significant bitplane values are independent between the two frames, then the watermark frames are i.i.d. variables with the distribution given above. Thus the expected value of the pairwise correlations $\mathbf{E}\rho(W_i, W_j) = 0$. If the same watermark plane is used to mark two identical video frames, then the bitplanes are not independent between these two frames, nor are the watermark planes, and therefore $\rho(W_i, W_j) = 1$.

One of the first transform domain watermarks, upon which many variations have been based [5]–[7], is presented by Cox *et al.* [8]. The watermark is a normally distributed sequence of real numbers added to the full-frame Discrete Cosine Transform (DCT) of each video frame. Like the JAWS watermark, it uses the same noise-like pattern for each frame, thus $\rho(W_i, W_j) = 1 \; \forall \, i, j$. Note that in transform domain approaches such as this one, the video and watermark frame correlations are computed in the appropriate transform domain.

Another transform domain approach is the Discrete Fourier Transform (DFT)-based spread spectrum scheme by de Guillaume *et al.* [9]. It is based on a successful image watermark proposed by Pereira *et al.* [10] and also incorporates some support for collusion resistance by varying the key, and hence the watermark pattern, every L_w frames. Thus two watermark frame correlation coefficients are possible: $\rho(W_i, W_j) = 1$ if $\lceil i/L_w \rceil = \lceil j/L_w \rceil$, and 0 otherwise where $\lceil \cdot \rceil$ is the ceiling operator. The main idea behind this strategy is that consecutive frames, within the set of length L_w, are expected to contain visually similar content and should therefore be marked using the same key. Frames outside of this set are expected to contain visually different material, and hence different keys are used. However, due to the use of arbitrary scene lengths, as well as arbitrary scene boundaries, these properties may not be completely accurate for real video sequences. A major challenge is also the generation, storage, and synchronization of the numerous keys. Both the detector and embedder must agree on a set of potential keys to be used. The detector must then check for watermarks embedded using *any* of these keys in *each* of the frames. Its complexity is thus immediately increased by a factor of the number of keys K_U/L_w, which becomes quite significant as the length of the video K_U increases.

In [11], Swanson *et al.* propose a multiresolution video watermarking approach that uses a perceptual HVS model to embed a highly robust watermark. The algorithm is scene-based and partitions the video into logical instead of arbitrary temporal segments. It achieves robustness to multiple frame collusion by working in the TWT domain and constructing a temporally layered watermark that is embedded into each motion plane of the video. Thus during a static scene, the watermark frames are highly correlated, whereas during a dynamic scene, the watermark frame correlations decrease. Although difficult to verify analytically, we believe that this watermark supports good statistical invisibility. However, the double transformation adds to the complexity of the scheme and this is its main weakness for real-time applications.

Darmstaedter *et al.* propose a method called TALISMAN that embeds hidden data by manipulating average luminance intensities, in sub-regions of each frame [12]. The embedding procedure is comprised of three operations: Block classification and separation of the pixels into zones; further subdivision of each zone into categories defined by a secret key-based grid; and embedding data bits by manipulating the mean energy of the pixels in each category according to some difference thresholds. Because of its consideration of local picture composition, reasonable resistance to multiple frame collusion is expected.

Finally, relevant work by Fridrich [13], [14] involves the computation of a robust visual hash function that is "similar" for correlated video frames and "different" for highly disparate video content. This hash is then used to generate a watermark pattern as an approximate sum of Gaussian random sequences; the correlation properties between resulting watermark patterns match those of the hash sequences employed to produce them. Hence, we believe, that Fridrich's scheme, designed with host-watermark correlation matching in mind achieves good statistical invisibility.

III. Spatially Localized Image-Dependent Subframe Watermarking

Based on the concept of matched host-watermark correlation developed in our previous work, and given that it is also desirable that the watermark possess the following characteristics:

- a relatively low complexity detection algorithm
- 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. The key idea underlying SLIDE is that of modulating frame-wise correlations by dividing the watermark frames into smaller subframes, and

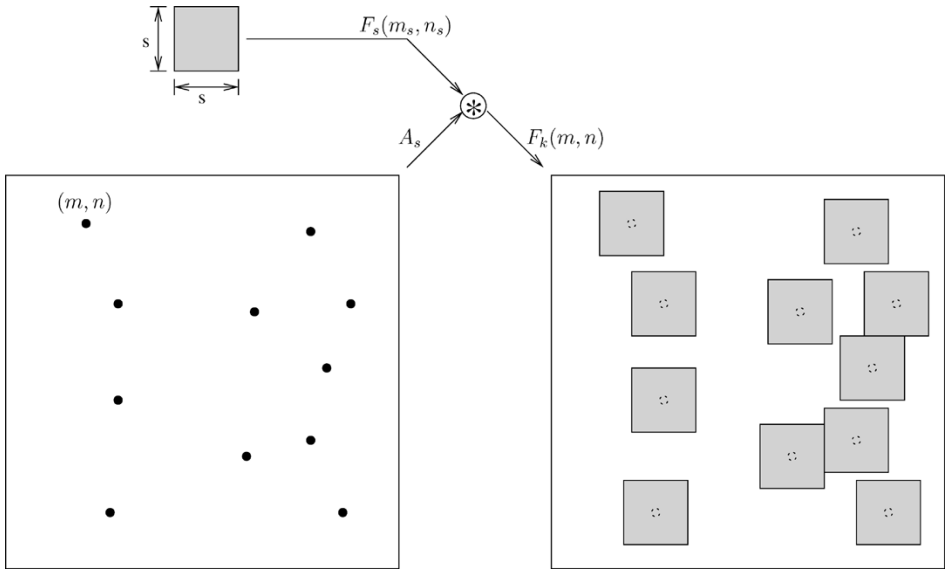

Fig. 1. Example of watermark footprint generation.

varying the alignment of these subframes in an image-dependent manner. First we introduce the watermark footprint, one of the tools used to describe the spatially localized nature of the proposed structure. Next we demonstrate how spatial localization and image-dependence respectively, enable us to achieve enhanced statistical invisibility.

A. Watermark Footprint

The watermark footprint describes the spatio-temporal spread of the watermark's energy. It can have any form, e.g., one obtained by block- or object-based segmentation of specific objects or regions of interest. However in a real-time video watermarking application the watermark footprint must possess two important qualities:

- invariance to small distortions introduced by attacks or other image transformations, i.e., since it describes the placement of the watermark, the footprint must be recoverable at the detector to facilitate detection;
- efficient representation that is compatible with watermark embedding and extraction algorithms, e.g., irregularly shaped objects are generally incompatible with existing techniques.

Most of the existing video watermark proposals embed the mark throughout the entire video sequence, i.e., they have global footprints. Global watermark footprints do possess the qualities described above, however many such watermarks have a critical weakness to multiple frame collusion. Therefore, we propose a watermark footprint that is comprised of a set of nonoverlapping basic patterns or *subframes*. All of the subframes are square in shape and have the same dimensions specified by an integer sidewidth s. The set of points covered by a subframe centered at (0, 0) is a neighborhood denoted $N_s(0,0)$, where

$$N_s(0,0) = \left\{(m^s, n^s) : m^s, n^s \in \left\{-\left\lfloor \frac{s}{2} \right\rfloor, \ldots, \left\lceil \frac{s}{2} \right\rceil - 1\right\}\right\}.$$

Next, we define a basic footprint pattern, which is a binary mask that selects the image pixels that lie within a subframe centered at (0, 0):

$$F_s(m,n) = \begin{cases} 1, & (m,n) \in N_s(0,0) \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we choose a finite set of L anchor points in the coordinate space of the frame, such that subframes of sidewidth s placed around each of the anchors do not overlap. Thus, the L anchor points can be represented as

$$\begin{aligned} A_s = \{(m_i, n_i) : |m_a - m_b| < s \Rightarrow |n_a - n_b| \geq s, \\ |n_a - n_b| < s \Rightarrow |m_a - m_b| \geq s, \\ \forall\, a, b, \in \{1, 2, \ldots, L\}, \text{ and } a \neq b\} \end{aligned}$$

where $i = 1, 2, \ldots, L$.

The footprint is then constructed by convolving the basic pattern with the set of anchor points, as illustrated in Fig. 1. From the perspective of the embedding and detection algorithms, each subframe can be processed just like an image. Therefore all of the ideas from image watermarking become immediately applicable, and we can make use of already existing theory in the research community.

Since the watermark patterns are embedded into the subframes, in order for efficient detection to be possible, the locations of the anchor points must not deviate significantly under image transformations. In other words, the robustness of the anchor points, and hence of the footprint, to watermarking attacks plays a critical role in detector synchronization. To enhance footprint robustness, we propose to select the anchor points A_s from a set of invariant features extracted from the image. Then, assuming that the attacked signal is perceptually equivalent to the watermarked signal, the positions of the anchor points should remain consistent with the visual content of the frames. Thus, the proposed scheme is characterized by its novel *content-based synchronization* approach, as opposed to the *structural synchronization* (i.e., based on the structure of the frame rather than the content) used in most current techniques.

B. Case for Spatial Localization

The *spatial localization* property means that our watermark is specifically placed only into certain spatio-temporal regions of the video by design. Its footprint does not cover all of the pixels in the video sequence. This is 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 our notion of a spatially localized watermark from another class of marks, also with localized footprints, known as *region of interest watermarks*. For instance, in [15] Su *et al.* propose an image 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 image watermark is presented by Brisbane *et al.* in [6]. 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 work (due to the additional growing step).

To motivate our proposed watermarking framework, we consider 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 white spread spectrum watermarks with global footprints, e.g., [3], [4], and [16] we see that

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

Either the same pseudo-noise (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 as proposed in Section III.A with the same basic zero-mean stationary white watermark pattern W_s embedded into each subframe for lower complexity. Taking two arbitrary video frames with overall watermarks W_a, and W_b, 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. Specifically,

$$\begin{aligned} &\rho(W_a, W_b) \\ &= \frac{\mathbf{E}W_a W_b}{\sqrt{var(W_a)var(W_b)}} \\ &= \frac{\sum_{l=1}^{L} \mathbf{E}_{(m^s,n^s)} W_a(m_l^a+m^s, n_l^a+n^s) W_b(m_l^a+m^s, n_l^a+n^s)}{\sum_{l=1}^{L} var(W_s)} \\ &= \frac{\sum_{l=1}^{L} \mathbf{E}W_s(m,n) W_s(m-[m_l^b-m_l^a], n-[n_l^b-n_l^a])}{L\sigma_W^2} \\ &= \frac{L'}{L} \end{aligned} \tag{1}$$

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 $m_l^b - m_l^a = n_l^b - n_l^a = 0$ which corresponds to the number of numerator terms in (1) that are nonzero, and $L = |A_s^a| = |A_s^b|$ is the total number of anchor points in each frame.

Based on the number of anchor points that are selected in each frame, the correlation of the overall watermark patterns W_a, and W_b can be adjusted. The resolution of these adjustments is limited to discrete steps governed by L, 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, i.e., through the judicious selection of the content-dependent anchor points, we can attempt to achieve greater statistical invisibility. This requirement brings us to the *image-dependent* part of the proposed framework.

C. Case for Image-Dependence

The correlation between two host video frames is 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. Thus, similar sets of anchor points are 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, the sets of anchor points extracted from host frames with low correlations should not share many common points.

Using this idea, we can draw a link between the concepts behind the current work and a recent publication on an optimal image collusion attack [17]. In contrast to the definition of linear collusion proposed in this work, the authors of [17] 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 combating 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 statement supports our goal of watermarking identical or highly similar video frames in the same manner.

D. Discussion

Our use of an irregularly tiled footprint structure for watermark embedding leads to some potential compromises that we discuss in this section. It is clear form Fig. 1 that the footprint structure limits the volume of host video into which the watermark is embedded hence limiting its capacity. This necessary trade-off in our formulation for collusion resistance may be acceptable for some applications given the overall volume of the host video still available for watermark embedding. In contrast, this approach of localizing the watermark may not be appropriate for audio or image watermarking applications in which the absolute host signal volume available for embedding is low.

The footprint structure, it seems, also has disadvantages in terms of robustness and detection time compared to globally spread watermarks as less of the watermark is embedded within each individual video frame. However, if we consider the work of Voloshynovskiy *et al.* [18] in which an attack is devised that can remove the watermark from all "flat" image components, we see that watermark robustness is not only dependent on the overall span of the watermark embedded per frame, but also of the visual characteristics of the host signal in which it is embedded. Thus, judicious selection of the footprint may be able to overcome some of the negative impact on robustness.

The reader should also note that the use of irregularly tiled subframes in our work can make the approach more susceptible to geometric distortions. For instance, in [19] the authors present, JAWS+, a method in which the symmetry of regularly tiled frames in JAWS is exploited in order to estimate scaling parameters to undo a class of geometric distortions. Such a technique for geometric recovery cannot be applied in our framework. To overcome geometric distortions in the proposed scenario, we would take a different approach in which the anchor points are selected by a feature extraction algorithm that has superior robustness to geometric distortions. This would be combined with the use of, for instance, a circularly symmetric or affine invariant watermark pattern. From experience with feature extraction and watermark generation alternatives, we believe that proper design of both stages requires significant investigation and increases algorithmic sophistication and complexity compared to [19].

IV. Novel Collusion-Resistant Video Watermark

We now overview the implementation details of a watermark that we have developed based on these concepts. The essential novelty of the work is twofold:

- 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 such as cropping and row/column deletion.
- 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 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.

A. Watermark Embedder

The SLIDE framework presents an overall paradigm for video watermarking; two core components have to be designed in the proposal of a practical mark. The first is the footprint generation algorithm, that produces a set of anchor points about which watermark subframes are then embedded. The second is the basic pattern generation step. Many ideas from the image watermarking literature can be used in the construction of this noise-like basic pattern. We have chosen the simplest one, a spatial domain spread spectrum pattern, as an illustrative example. Because of its subframe-oriented nature, there are clear

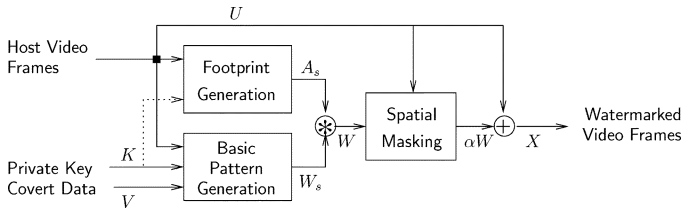

Fig. 2. Block diagram of proposed watermark embedder.

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.

After the basic watermark pattern has been generated, it is convolved with the set of extracted feature points to form a frame-sized watermark. Then spatial masking is applied to the watermark frame to modulate its strength locally, according to the properties of the video frame itself. Finally, the scaled watermark is embedded by addition to the host. The five main steps of the proposed embedding algorithm are illustrated graphically in Fig. 2. Each block is discussed in more detail in the ensuing sections.

1) Footprint Generation: In this work, the structure of the watermark footprint is completely specified by two parameters: the side-width s and a set of anchor points A_s. We consider odd side-widths so that the centers of the subframes fall into well-defined pixel locations. Because the selection of a feature as an anchor point excludes other features within a distance of s from subsequently being selected, the robustness of the selection algorithm increases as the number of excluded points decreases, i.e., as s decreases. However, the robustness of the watermark pattern itself increases as its spatial redundancy increases, i.e., as s increases. Using the value of $s = 128$ chosen for JAWS as a general guideline, tests of the feature extraction algorithm were conducted and a side-width of $s = 81$ is chosen to provide a good compromise between these two desirable properties.

The next important question to consider when generating the footprint is how to place the subframes in an image-dependent manner that can be reproduced at the detector, i.e., what types of functions of the image are advantageous for robustness and watermarkability. To address this issue, we turn our attention to image feature extraction algorithms used in computer vision and object recognition-related applications. Feature extraction, sometimes also referred to as *local feature extraction* in the literature, is a mechanism for locating parts of an image that possess some special properties. In our case, we will be interested in features that enable the identification of visually significant artifacts, e.g., those based on sharp intensity gradients like edges and corners. These are expected to be robust to distortions that are perceptually insignificant, and therefore be suitable for watermarking in the proposed collusion resistant framework.

We tested a number of feature extraction and related image processing algorithms, including image normalization [20], perceptually significant masking [21], corner detection [22], Sobel edge detection [22], and an approach based on a visual scale-interaction model [23]. Because either their complexities were too high for frame-based video watermarking, or the additional

processing required to produce a feature set usable for watermarking purposes was too great, we did not find that any of these could be directly applied in our implementation. Therefore, we have developed a feature selection procedure based on the following key ideas borrowed from the above algorithms:

- perceptual masking to reduce distortion arising from small amplitude noise while preserving important visual structures [21];
- selecting features by identifying local extrema of a nonlinear function of the image [23].

In the proposed approach, watermark transmission via picture components with small peak interpolation distortions is favored [24]. The procedure for selecting anchor points from each host frame $U(m,n)$[1] is as follows:

1. Apply a perceptual significance mask to remove perceptually insignificant components from the ensuing analysis, e.g., small amplitude noise signals **[21]**.
2. Compute the peak spatial *interpolation distortion* bound on the masked frame **[24]**

$$M(m,n) = \max_{|i-m|,|j-n|\leq 1} |U(i,j) - U(m,n)|.$$

3. Average this bound over $s \times s$ subframes

$$\overline{M}_s(m,n) = \frac{1}{s^2} \sum_{(m^s,n^s)\in N_s(0,0)} M(m+m^s, n+n^s).$$

where the summation is over the $s \times s$ neighborhood centered at (m,n). We constrain the averaging operation such that the potential subframes lie completely within the image space.

4. Extract all of the local minima of $\overline{M}_s$ and for each minimum point (m_i,n_i), compute its *strength*:

$$T_i = \min_{(m',n')\in N_3(m,n)} [\overline{M}_s(m',n') - \overline{M}_s(m_i,n_i)].$$

where $N_3(m,n)$ denotes the 3×3 neighborhood about (m,n).

5. Arrange the minima in order of decreasing strength, then choose the strongest set of nonoverlapping minima using a *greedy* approach, i.e., beginning from the top of the list, declare a minimum to be an anchor point if a subframe centered around it does not overlap with those surrounding any previously selected anchor points.

The robustness of the subframe selection algorithm is discussed in more detail in [24]; generally it was found that stronger minima are more robust, i.e., the visual correspondence between minima identified before and after perceptually invisible transformations is higher than that between maxima or other points. This notion of robustness comes from the field of image feature extraction [23], where invariant features are required for vision and pattern recognition. Finally, Fig. 3 illustrates the features extracted from four different frames of a test video.

It is easy to verify visually that similar sets of features are chosen from temporally adjacent frames with highly similar content, and also that dissimilar sets of features are chosen from frames with dissimilar content. To qualitatively justify why this might be the case, we make the following intuitive argument: Recall that the interpolation distortion bound can also be interpreted as the output of an omni-directional edge detector. Although in typical feature extraction procedures the quantity of interest is the maximum of these outputs, in our proposal they are first averaged before subsequent processing. Therefore the peaks become blurred, while the minima exhibit more robustness to the small differences that are present in visually similar video frames.

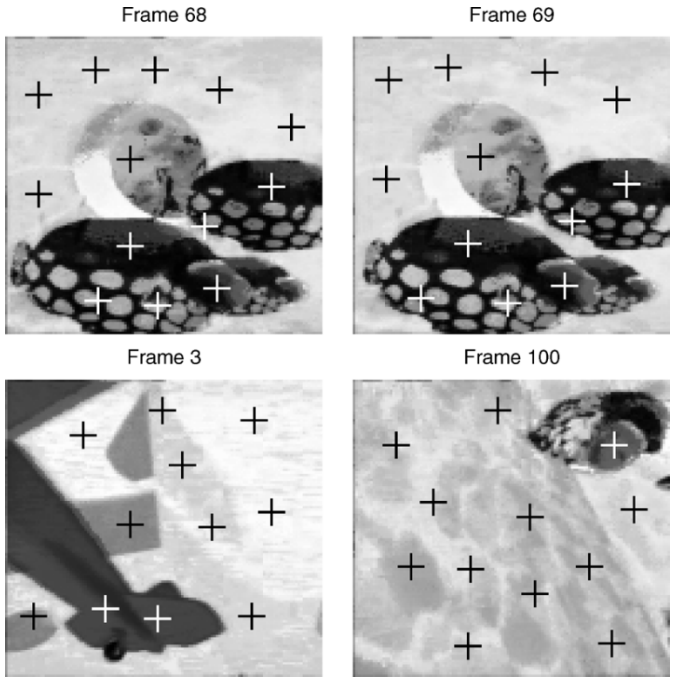

Fig. 3. Selected features for subframe watermarking, shown in four frames of fish_c2 video: Anchor points are marked by crosshairs in all frames. The top two frames are temporally adjacent and contain visually similar content. Therefore the extracted feature sets have a high correspondence. The bottom two frames contain highly dissimilar visual content, and we observe that the extracted feature sets also differ greatly.

Thus the proposed watermark footprint is generated, and is depicted for the Barbara image in Fig. 4.

2) Subframe Watermark Pattern Generation: The subframe watermark is a key-dependent real-valued Gaussian $s \times s$ pattern. Within each subframe, B raw data bits are embedded.[2] In this work, our goal is to investigate the SLIDE approach, therefore, we take the most straightforward technique to constructing the watermark pattern using spread spectrum principles which has been heavily studied in the literature. Here, each data bit in a subframe is encoded using a finite length real-valued Gaussian sequence that is pseudo randomly generated independent of the host video; the Gaussian sequence generation stage uses a seed derived from the secret key and the data bit. Each data bit, therefore, has a corresponding real-valued Gaussian sequence which is its "encoded" representation. The actual embedded process of the watermark into the host video subframe involves scaled

[1]The reader should note that since we consider feature point extraction in individual frames, we drop any reference to frame number in the video sequence.

[2]If error correction coding is applied to enhance robustness of the watermark, then $n = B$ is the block length of the code with data rate $k < B$. Thus, B is the "uncoded" data rate.

Fig. 4. Barbara with proposed watermark footprint overlaid $s = 81$.

addition into a given subset of pixels in the subframe, as we discuss in the next section. Each Gaussian sequence is embedded into a different subset of pixels in the subframe. Therefore, there is no overlap of the encoded watermark sequences for different bits.

The reader should note that almost any image watermark found in the literature can be adapted to form an appropriate subframe pattern. In addition, different patterns in distinct subframes or multiple orthogonal patterns in the same subframe are also possible. The selection of the method of pattern generation should depend on the attacks to which the method must be robust and the required capacity of the embedded watermark. Effective watermark pattern generation is possible to support enhanced robustness to a broader class of attacks.

To improve the watermark detector's performance in this work, we exploit the fact that a video frame is comprised of many irregularly spaced subframes. Therefore, the watermark is repeatedly available in the video sequence. Thus, if we estimate the accuracy of data detection in each subframe, we can then combine all the watermark repetitions in an optimal way for added robustness. As described in [24], we propose the use of Maximal Ratio Combining (MRC) to optimize the detector's performance. This requires employing a pilot sequence also referred to as a reference watermark in the literature [25], [26]. Given that our subframe has dimensions 81×81, there are 6561 pixels in which to embed both the data bits and pilot sequence. In our implementation, the proposed pilot is a sequence of i.i.d. Gaussian random variables of length 3281 produced using a random number generator and the watermark key as the seed. The pilot is pixel interleaved with the spread spectrum Gaussian encoded data sequences of combined length 3280 as illustrated in Fig. 5, to form the basic watermark pattern $W_s(m^s, n^s)$. This data signal may be encoded for error protection or embedded in its raw form. For the tests reported

R	D	R	D	⋯	D	R	D	R
D	R	D	R	⋯	R	D	R	D
R	D	R	D	⋯	D	R	D	R
D	R	D	R	⋯	R	D	R	D
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
D	R	D	R	⋯	R	D	R	D
R	D	R	D	⋯	D	R	D	R
D	R	D	R	⋯	R	D	R	D
R	D	R	D	⋯	D	R	D	R

Fig. 5. Interleaving pattern for the reference watermark (R) and embedded data (D) in a square subframe with odd side-width.

in Section V, a simple repetition code is used, i.e., given the desired raw data rate of B message bits per frame, the spread spectrum sequence for each bit is of length $\lfloor 3280/B \rfloor$. Further details can be found in [24].

3) Full-Frame Watermark Construction, Masking, and Embedding: The full-frame watermark is constructed by convolving the footprint and the basic watermark pattern.

$$W_k(m,n) = W_s(m,n) \star \left[\sum_{(m_l,n_l)\in A_s} \delta(m - m_l, n - n_l) \right]$$

where $\star$ denotes 2-D linear convolution, and $\delta(\cdot,\cdot)$ is the 2-D Dirac delta function.

Then local image-dependent scaling factors, derived from the Noise Visibility Function (NVF) proposed by Voloshynovskiy *et al.* [27], are applied to optimize robustness while maintaining imperceptibility. Specifically, this factor is given by

$$\alpha_k(m,n) = 1 - \frac{1}{1 + \theta\sigma_d^2(m,n)}$$

where $\sigma_d^2(i,j)$ is an unbiased estimate of the local variance computed over a square window of side width d centered at (m,n), and

$$\theta = \frac{D}{max_{(m,n)}\, \sigma_d^2(m,n)}$$

where $D \in [50, 100]$ is an image-dependent constant. For our purposes, we use $d = 3$ and take θ to be inversely proportional to the maximum local variance estimated over each subframe rather than the whole image. We also arbitrarily choose $D = 75$.

Finally, the watermark and host are combined:

$$X_k(m,n) = U_k(m,n) + \alpha_k(m,n) W_k(m,n).$$

B. 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 we did at the embedder in Section IV-A1. Observe that the detected footprint is denoted $\hat{A}_s$ to emphasize that it will

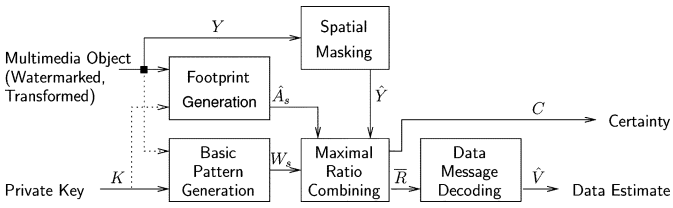

Fig. 6. Block diagram of proposed watermark detector.

not necessarily be identical to the footprint selected at the embedder. 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 Fig. 6.

To reduce the power of the host image component, a 3×3 Laplacian filter is applied before any subsequent processing. Then given the estimated local scaling factors, denoted $\hat{\alpha}$ again to indicate a possible deviation from the original factors α used at the embedder, an MRC detector is implemented to take full advantage of the spatial diversity inherent in the watermark [24]. Basically we compute the projections R_l of each subframe onto the reference watermark pattern generated by K. The square of this projection gives an estimate of the watermark signal power, and the power of the noise can be estimated by summing over the squared frame residuals, i.e., the part of the received frames that is orthogonal to the watermark. For each subframe, the ratio of these two quantities gives a signal-to-noise ratio (SNR), which is then used to weigh the data component of the watermark in the formation of a combined data signal. The certainty of the watermark's detection is obtained by looking at the average SNR of the subframes that are used in the combination, and the extracted data estimate is obtained by decoding the combined signal.

To improve performance, the contributions from subframes that have very low SNRs are rejected. Specifically, we define a threshold T, depending on the desired false positive probability. The probability of a false positive corresponds to the probability that the projection of an unmarked subframe onto the watermark pattern will be greater than T. Because by assumption the subframe and watermark are independent, and the watermark is zero-mean, we can invoke the Central Limit Theorem (CLT) and approximate this projection as Gaussian random variable, also with a mean of $\mu = 0$. The standard deviation σ of this random variable is dependent on the properties of the image, however it was experimentally determined (by analyzing standard test images) that this value is approximately $\sigma = 0.5$. Thus in order to achieve a false positive rate of $5.7(10^{-7})$, a detection threshold of at least 5σ or $T = 2.5$ should be used. Having chosen T, if $R_l > T$, we say that a watermark has been detected, otherwise we reject the subframe as corrupted, misaligned, or unmarked.

V. Simulation Results and Performance Evaluation

In this section, we present results that demonstrate the performance of the proposed scheme. The primary baseline used in these tests is our implementation of the JAWS watermarking system. This existing approach is chosen because of its current use and success in commercial applications, as well as the similarity between the embedding concepts employed in JAWS and in the proposed approach. For instance, both 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 centers are synchronized relative to visual features and hence they are irregularly located. We have also implemented the code division multiple access (CDMA) watermark and test the proposed algorithm against it for a collusion attack.

In the first three sections, we consider in detail the following attacks: fractional pixel translation (Section V-A), cropping, row and column removal (Section V-B), and JPEG compression (Section V-C). All of these, except for the first, are applied using the StirMark 3.1 watermark benchmarking software [28]. The purpose of these comparisons is to demonstrate that the performance of the proposed watermark is comparable to that of existing approaches under nongeometric attacks. These are also attacks that can occur in standard video processing and transmission operations.

To test the collusion-resistant properties of the proposed scheme, we consider a modified Type I collusion attack in Section V-D and a Type II attack in Section V-E. 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 second attack is conducted by averaging adjacent frames of a real video sequence after watermarking to obtain a mark-free copy of the host. The results demonstrate that the proposed schemes are indeed robust against Type I and II linear collusion.

The single frame tests reported in this chapter were all conducted using the Barbara standard image and the following general procedure: First the image is watermarked, then after watermarking it is subjected to the attack under consideration. Finally, the image is passed through the watermark detector and the number of errors in the extracted message is recorded. When detection of the watermark fails, half of the bits are deemed to be in error. Each test is repeated $N = 100$ times using randomly generated payload messages, and the overall bit error rate is obtained by averaging over the number of trials as well as the number of watermark payload bits:

$$\text{Bit error rate} = \frac{\text{Total number of errors in } N \text{ trials}}{N \cdot k}$$

where k is the size of the payload in bits. Also, a subframe side width $s = 81$ was found experimentally to give a good performance tradeoff between robustness and data rate. As specified in [3], our implementation of JAWS uses tile sizes of $M = 128$, and a detection threshold of $T = 5/M$. It also has a maximum payload size of $k = 8$ bits. We found that the simple detector presented in [3] resulted in decoding failures when more than two correlation peaks exceeded the threshold in magnitude. Therefore, results obtained using a threshold of $T = 15/M$ and

Fig. 7. Illustration of sample test images. Original Barbara (left) and Barbara watermarked using the proposed SLIDE watermark (right). $\mathrm{PSNR} = 38$ dB.

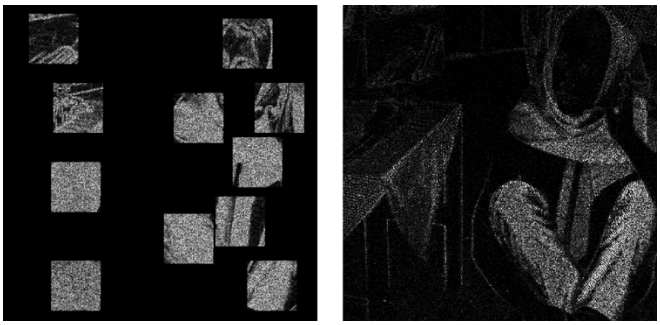

Fig. 8. Illustration of sample test watermarks and their footprints. Proposed SLIDE watermark (left) and JAWS watermark (right), both scaled by a constant factor of 20 to make them visible.

the same threshold with a *min-max peak selection* strategy are 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. If no correlation peaks exceed the threshold, a decoding failure still results.

Note that the bit error rate in JAWS is strongly dependent on the manner in which the payload messages are allocated to possible peak locations. In our implementation we used a simple sequential allocation strategy, resulting in an average bit error rate (BER) of 22.5% when there is a nearest neighbor error (i.e., one of the correlation peaks is shifted to a neighboring position at the detector). A random allocation strategy results in an average BER of 50%, and a grey coded allocation in 12.5%. Thus an optimized detector may perform better than our implementation, possibly achieving up to just over half the bit error rate shown on the plots. However, note that in many of the cases shown here we are actually seeing decoding failures rather than high bit error rates. Therefore, we assert that the lack of optimization does not play a significant role in the comparisons.

Finally, to ensure a fair comparison, the strengths of the watermarks are adjusted to a peak SNR (PSNR) of 38 dB after embedding and both implementations have comparable false positive rates below $1.0(10^{-6})$. In Fig. 7, the original Barbara image is depicted, along with the images watermarked using JAWS and the proposed watermark. In Fig. 8, the absolute values of the corresponding watermarks are shown, scaled by a constant factor of 20 for visibility. It is clear from Fig. 8 that through the straightforward implementation of our framework, SLIDE results in a watermark with a "blocker" appearance than its JAWS counterpart. Although no visual distortion is introduced by the

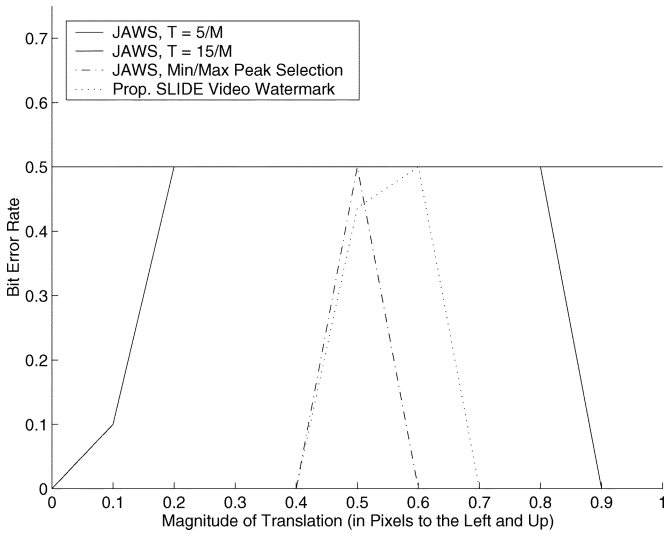

Fig. 9. Bit error versus diagonal fractional pixel translation attacks (using bilinear interpolation) for the proposed spatial domain algorithm and JAWS. The PSNR is fixed at 38 dB.

SLIDE mark, as observed in Fig. 7 (we attribute this, in part, to our use of the NVF-based scaling parameter), it is clear that both watermarks have different perceptual characteristics. We believe through observation of Fig. 8 (and results alike) that JAWS may have superior perceptual masking capability than our implementation of SLIDE. From the perspective of perceptual visibility of the watermark, as pointed out by one reviewer, the frame-to-frame differences in edges are relevant. Hence, use of edges as anchors for watermarks in video applications may not be appropriate. However, our implemented feature extraction algorithm does not identify strong edges per se. Hence, we do not believe that this is a serious concern in our scheme. Our experience from simulations did not demonstrate any strong trends of this sort. As we discuss in Section VI-A, future work involves identifying methods of feature extraction and watermark pattern generation to more readily guarantee imperceptibility and exploit temporal masking.

A. Fractional Pixel Translation

The fractional pixel translation test is an example of a transformation that can easily be applied to an image in an attempt to remove a watermark. It is also one for which the effects are bounded by the interpolation attack bound presented in Section IV-A1. The purpose of the test is to study the performance of JAWS compared to that of the proposed spatial domain approach under such an attack. Fig. 9 illustrates the results of the test; the translation attack is implemented using bilinear interpolation. We observed that as the image was translated by a fractional number of pixels, additional peaks appeared in the response of the JAWS detector, thus inducing more decoding failures and a higher bit error rate. The proposed spatial domain algorithm also encounters more bit errors, however the overall degradation is much more gradual. We believe that this improvement can be attributed to two factors:

- the energy of the proposed watermark is concentrated in subframes with low interpolation distortions, and
- the subframes are combined in an optimal SNR manner by the proposed watermark detector.

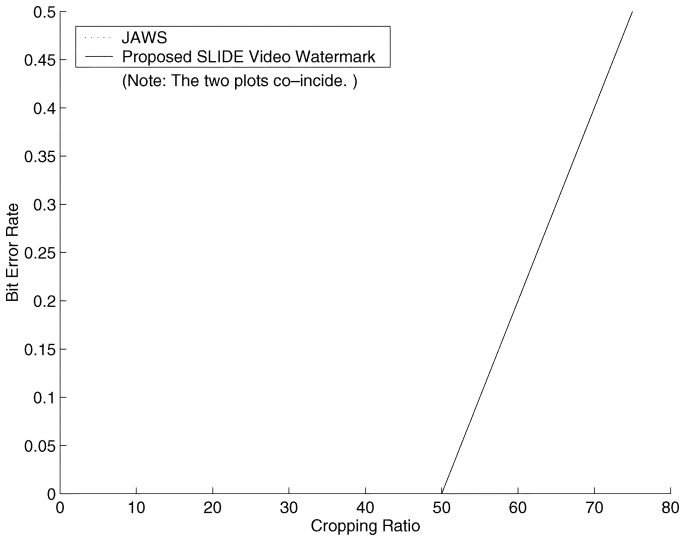

Fig. 10. Bit error versus cropping attacks (using StirMark 3.1) for the proposed algorithm and JAWS. The PSNR is fixed at 38 dB. Please note that both plots coincide in this diagram.

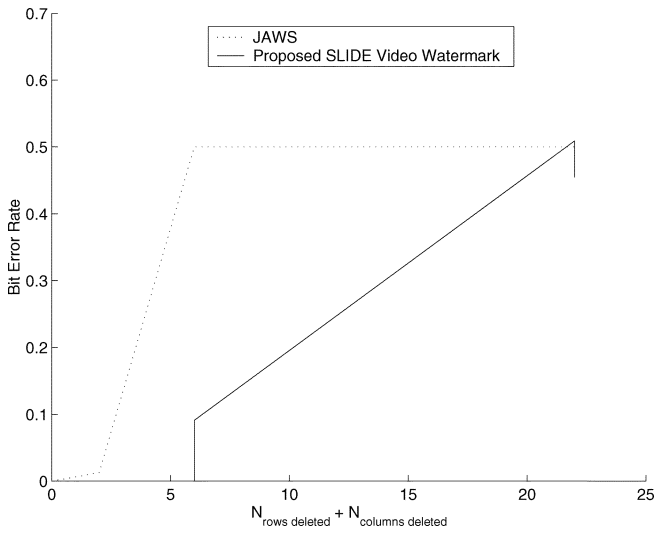

Fig. 11. Bit error versus row/column deletion attacks (using StirMark 3.1) for the proposed algorithm and JAWS. The PSNR is fixed at 38 dB. The y-axis shows the bit error rates achieved when a number of row/columns are removed from the image. The x-axis indicates the total number of rows and columns that have been removed. The specific tests in the suite include the removal of one row/one column, one row/five columns, five rows/one column, five rows/17 columns, and 17 rows/five columns.

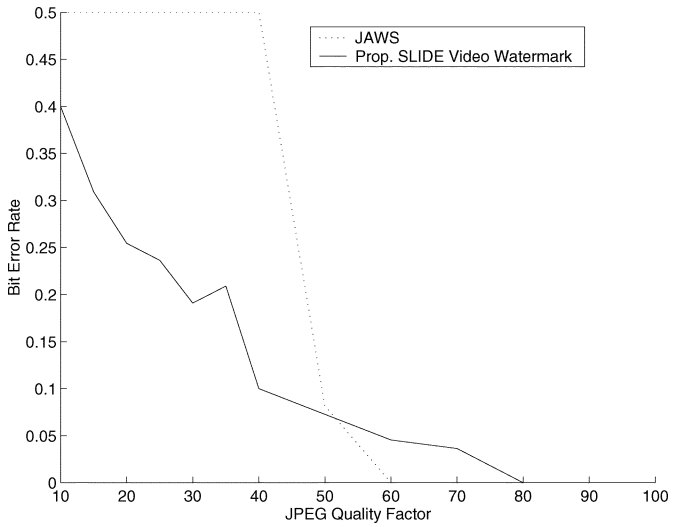

Fig. 12. Bit error versus JPEG compression attacks (using StirMark 3.1) for the proposed algorithm and JAWS. The PSNR is fixed at 38 dB.

B. Structural Attacks: Cropping, Row/Column Deletions

The two attacks that we consider in the category of structural manipulations are cropping and row/column deletion. In the case of cropping attacks, one detrimental effect that may be suffered by the watermark is mis-alignment of the footprint subframes. Given a received image that has been cropped, the footprint generated by the detector will be constrained to lie within the space of the cropped image. Thus subframes from the original footprint that include parts of the image that were cropped out, typically near the boundaries, will not be properly aligned. In addition, any mis-alignment along the boundaries may also propagate to interior subframes, due to the fact that they are selected in a nonoverlapping manner.

All of the remaining tests were conducted using StirMark 3.1. The test suite includes image cropping by 1, 2, 5, 10, 15, 20, 25, 50, and 75%. The no threshold version of JAWS is used in these tests, since as discussed earlier it represents the best case performance and therefore gives the best baseline against which to demonstrate improvements. We find that for both of the algorithms, cropping of up to 50% of the image does not have a critical effect on the error performance of the algorithms. However, when 75% of the image is cropped away, none of the watermarks are successfully detected. The plot in Fig. 10 displays the resulting bit error curves.

For row and column deletions, we observe that since both the proposed watermark and JAWS are based on directly added PN sequences, such deletions will result in desynchronization and significantly degraded detection. However, we also note that when a small number of row/columns are removed, many subframes still remain intact. Because the proposed algorithm does not rely on absolute spatial synchronization, these subframes can be located and properly processed as long as the footprint selection is reasonably well-preserved. We find experimentally that this is in fact the case. Therefore the diverse nature of the detector results in unaffected subframes with consequently higher SNR's playing a larger role in extracting the watermark payload. Thus improved performance is achieved by the proposed spatial domain watermark.

The results of these tests, illustrating the observations noted above, are presented in Fig. 11.

C. JPEG Compression: Quantization in the 8×8 DCT Domain

The StirMark test suite applies compression both at incidental levels, where perceptual quality is not severely degraded, as well as at levels that would be considered attacks, where significant quality degradation is encountered. For good quality full-color source images, the default JPEG quality factor setting is 75, and as a rule of thumb, at levels above 50 there should be little objectionable degradation [29]. We find in Fig. 12 that the proposed watermark is characterized by a soft decrease in detection performance, and a correspondingly soft increase in bit error rate, as the JPEG quality factor is reduced. The JAWS algorithm, on

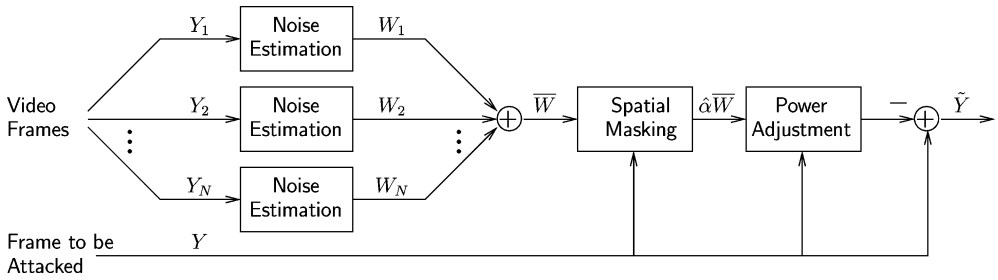

Fig. 13. Block diagram of modified Type I collusion attack.

the other hand, exhibits an almost binary nature, where the watermark is either perfectly decoded or not at all. This effect can be attributed to the observations made in Section V-A.

D. 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/s, and for instance in an action film, the scenes or shots may change dramatically every 0.5 s, thus making such sequences easy to construct [30]. 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. 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 Fig. 13.

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.* in [27]. Anisotropic diffusion is a popular image processing technique that can be used to remove noise from an image [31], [32]. 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). In the current implementation, a 3×3 window is used for computing the local gradients (as in [32]). The values of the coefficients are then selected experimentally, by choosing those that achieve the best estimates of a known PN pattern embedded in a number of test frames [24].

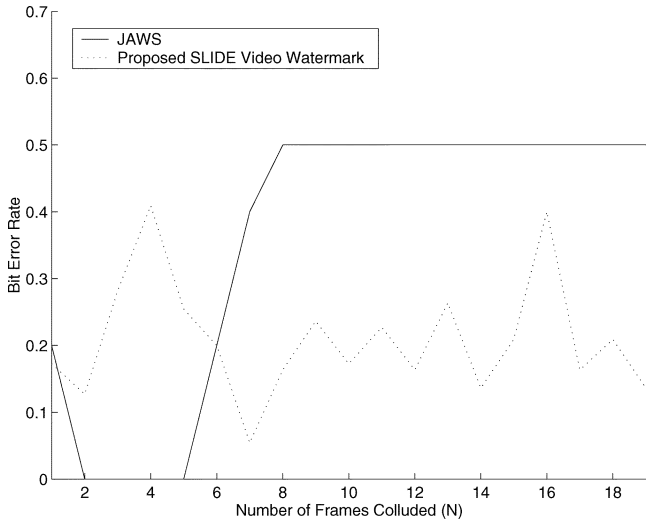

Fig. 14. Bit error versus number of frames used for collusion in modified Type I attack for the proposed algorithm and JAWS. The embedding and attack PSNRs are fixed at 38 dB.

The fish_c2 video sequence was used in our tests; every fifth frame was extracted to form the actual set of test frames. We watermarked this sequence using both JAWS and the proposed scheme. 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 Fig. 14. We can see that as the number of frames being combined increases, the performance of the JAWS system degrades. This behavior 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, using the proposed algorithm, the detector's performance does not severely degrade as more frames are combined. This behavior can be attributed to the time-varying nature of the watermark. As successive estimates are added to the combination, the overall estimate does not more closely resemble any of the actual watermark patterns.

To demonstrate this time-varying nature of the watermark, we refer the reader to Figs. 15 and 16. The plot in Fig. 15 shows the correlation coefficients between the first frame and a set of periodically spaced frames of the host video, the watermarked

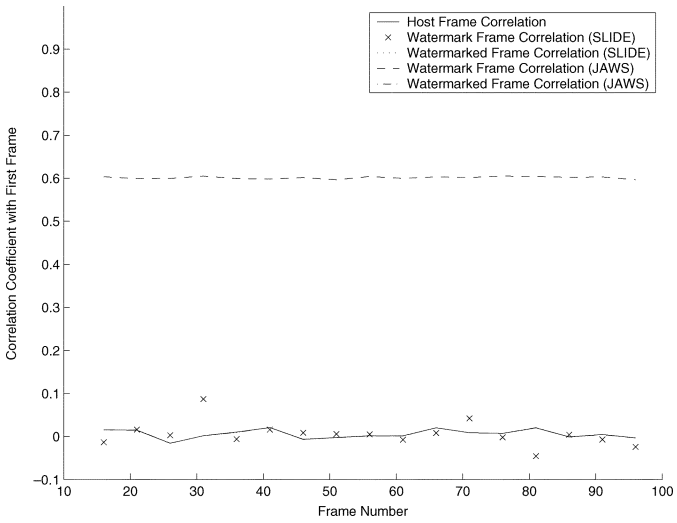

Fig. 15. Correlation coefficients between first frame and periodically spaced frames of the host video and watermark frames using JAWS and SLIDE. Please note that the solid and dotted lines coincide, and the dashed and dash-dotted lines coincide.

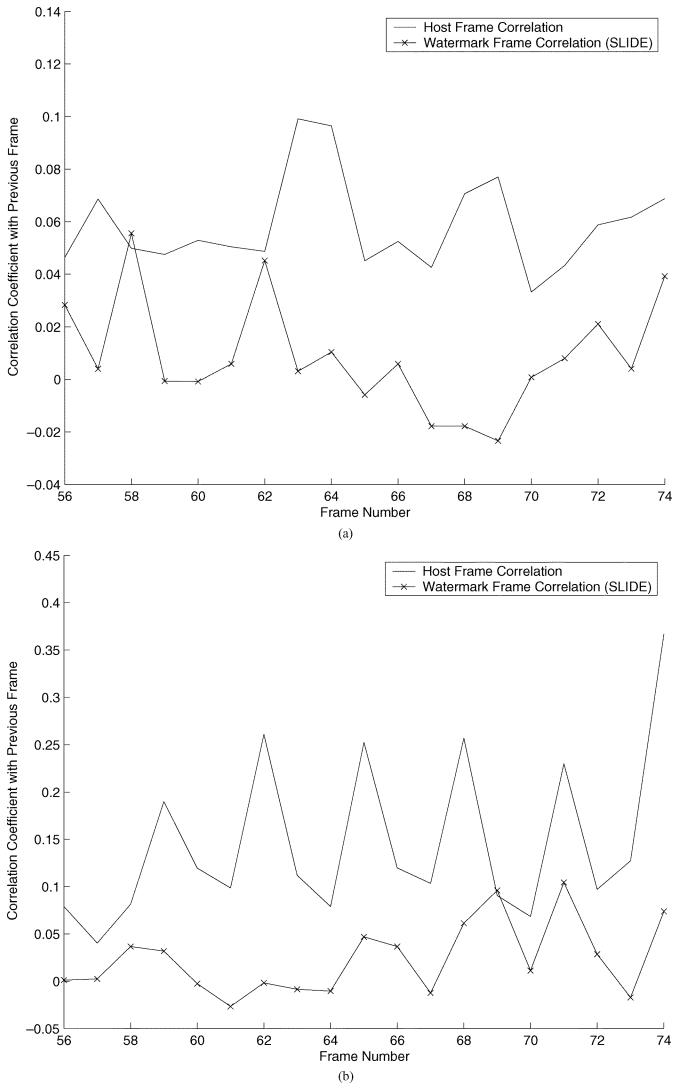

Fig. 16. Correlation coefficients between adjacent frames of the host video and watermark frames using SLIDE.

video and the watermark, using both JAWS and the proposed SLIDE algorithm. As expected, the correlations between the JAWS watermark frames are nearly constant. Those associated with the proposed approach vary in time, although not exactly as intended, according to variations in the host frame correlations. They are however much closer in magnitude to those of the host video than the JAWS coefficients.

To take a closer look at the correlation coefficients achieved by the proposed scheme, we also consider the two plots shown in Fig. 16. The plot in Fig. 16(a) depicts the correlation coefficients between the indexed (shown in the horizontal-axis) and previous frames of the fish_c2 video sequence. We can see that the overall trends and some features of the two curves are matched.

However, the watermark correlations are always quite close to 0. This effect is emphasized in the plot of Fig. 16(b), which shows the result of a similar experiment conducted with a different video sequence (a cartoon, wg_wt_3). The second video clip was chosen for this test since it is a relatively stationary sequence and thus has higher host video correlation coefficients. The main reason why the watermark correlations produced by the proposed algorithm are so small is because they are highly sensitive to overlap and mismatch of the selected subframes. We believe through analysis of our simulations that the deviations seen here are due to errors in the implementation of the anchor point selection mechanism.

E. 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 ten 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 proposed SLIDE and CDMA schemes (The reader should note that we have not included JAWS in the comparison because Type II is only applicable to cases in which different watermarks (i.e., independent patterns) are used in visually similar frames. Thus, JAWS is not susceptible to such an attack; the colluded frame will still contain the watermark pattern.). 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 schemes, the CDMA watermark is effectively removed by frame averaging, while the proposed watermark remains intact.

In Fig. 17, 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.

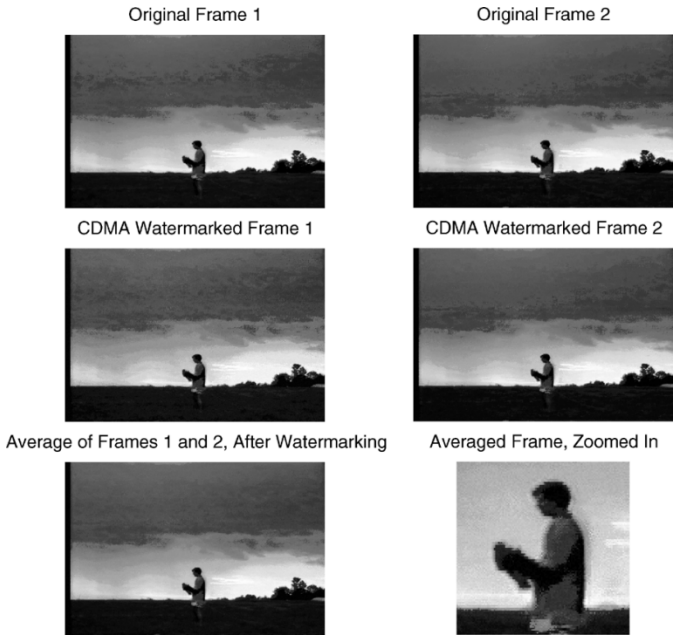

Fig. 17. Illustration of sample test frames. Original frames 1 and 2 from hawk3 video (top left and right), frames 1 and 2 watermarked using CDMA (middle left and right), averaged frame constructed from watermarked frames 1 and 2 (bottom left), and zoom in of averaged frame to show that the attack does not significantly damage visual quality although the watermark is removed (bottom right).

VI. Final Remarks

In this paper, we propose a simple spatially localized image-dependent video watermark based on a novel framework for collusion resistance. Continuing on our work in [1], we have focused on linear collusion attacks and presented some results based on a simple spread spectrum watermark pattern. These have provided us with a more concrete understanding of how collusion attacks work and how they may be practically counteracted. Two key ideas are emphasized: the watermark's energy is concentrated in subframes with desirable properties, and the locations of these subframes are *synchronized using visual content rather than structural markers*. The proposed watermarking scheme is distinguished by its ability to be embedded and extracted using frame-based algorithms, while resisting collusion.

In the development of the watermark, we introduce the notion of a watermark's *footprint*, the set of spatial coordinates over which its energy is spread. We are interested in localized footprints and consider the property of low average interpolation noise as a selector for good watermarking regions. Our experiments demonstrate that watermark footprints selected using this criteria have a reasonable robustness to geometric distortions. We also show that similar footprints are selected from video frames containing similar visual content, and dissimilar footprints from frames with differing visual content. Even so, this criteria has its weaknesses, that will be discussed in Section VI.A, and as a result we believe that low average interpolation noise will not ultimately be the best choice of footprint selectors. Lastly, we note that there is a tradeoff being made in terms of spatial redundancy when using a localized watermark. It is intuitively clear that some of the available capacity for information hiding is not being used. However, when working with video media, this loss can be compensated for by taking advantage of the large host bandwidth.

Another important contribution that we put forward is that of *content-based watermark synchronization*. Whereas many current watermarks use physical markers to achieve the spatial synchronization that their detectors require to determine the location or orientation of the hidden patterns, (one exception is the JAWS algorithm in which spatial synchronization is automatic due to the SPOMF detection in the frequency domain) the proposed scheme relies on image-dependent properties to perform this synchronization automatically. The underlying notion is that these image-dependent properties should be such that they cannot be modified without significantly affecting the visual quality or content of the frames. We argue that selecting subframes in a content-synchronized manner can enhance resistance to collusion. This approach also eliminates the need for absolute markers such as spatial start of frame tags, which can be affected by cropping, or temporal start of scene tags, which can be affected by frame averaging or dropping.

A. Potential Areas of Future Research

Many future research directions have been opened by this work.

- **Reducing computational complexity.** The speed and complexity of the approach is limited by the feature extraction step. To make the algorithm viable for use in real-time applications, it is necessary to improve the time performance of the footprint generation algorithm.
- **Working with small video frames.** As the dimension of the video frames decreases, we find that the robustness of the subframe selection algorithm drops because of the limited selection for potential subframe locations. The reduced amount of spatial redundancy available to the watermark also further degrades its performance.
- **Key-dependent subframe selection.** In the current implementation, the subframe locations are selected using a greedy algorithm, which chooses the global maximum of a function of the image at each step, as long as a subframe centered at that point would not overlap with any previously selected subframes. The danger of such an approach is that it can easily be duplicated by an informed attacker. A more clever solution would incorporate a key-based component in the selection of the anchor points.
- **Extracting robust features from high activity regions.** The current footprint selection favors regions with low average interpolation distortions. These correspond to regions with relatively low spatial gradients and hence low local activity levels. A recently published class of watermark attacks are based on de-noising the image, treating the estimated noise as the watermark signal, and remodulating the watermark to fool its detector [27]. This attack has been found to be particularly effective in low activity regions. In some preliminary tests, we found the robustness of features located in high activity regions to be lower than those in low activity regions. Therefore an avenue of

further study lies in devising feature extraction criteria that is both robust to distortion and also favors regions with higher activity levels.

- **Integration of the proposed watermark design philosophy with other strategies.** To implement our strategies for more general video watermarking attacks, we will investigate how best to integrate the concepts of spatial localization and image-dependence with existing tool-sets from the video watermarking community. Moreover, a methodology for design of the feature extraction and watermark frame generation stages to target robustness to specific attacks is a topic of further study.
- **Minimization of perceptual distortion of the watermark.** Given the distinct structure of the proposed method compared to methods such as JAWS, the SNR may not always be an accurate measure by which to evaluate the true perceptual distortion introduced by the watermark. Future work looks at relating the parameters of the scheme to actual perceptual tests in order to better characterize the obtrusiveness of the watermark and hence identify techniques to better mask it while retaining collusion resistance.

REFERENCES

[1] K. Su, D. Kundur, and D. Hatzinakos, "Statistical invisibility for collusion-resistant digital video watermarking," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 52–60, Feb. 2005.

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

[3] T. Kalker, G. Depovere, J. Haitsma, and M. Maes, "A video watermarking system for broadcast monitoring," *Proc. SPIE*, vol. 3657, pp. 103–112, Jan. 1999.

[4] B. G. Mobasseri, "Exploring CDMA for watermarking of digital video," *Proc. SPIE*, vol. 3657, pp. 96–102, Jan. 1999.

[5] L. Qiao and K. Nahrstedt, "Watermarking methods for MPEG encoded video: Toward resolving rightful ownership," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1998, pp. 276–285.

[6] G. Brisbane, R. Safavi-Naini, and P. Ogunbona, "Region-based watermarking for images," *Lecture Notes in Computer Science*, vol. 1729, pp. 425–435, Oct. 1999.

[7] W. Zhu, Z. Xiong, and Y.-Q. Zhang, "Multiresolution watermarking for images and video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 545–550, Jun. 1999.

[8] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[9] F. Deguillaume, G. Csurka, and T. Pun, "Countermeasures for unintentional and intentional video watermarking attacks," in *Proc. SPIE*, vol. 3971, Jan. 2000, pp. 346–357.

[10] S. Pereira and T. Pun, "Robust template matching for affine resistant image watermarks," *IEEE Trans. Image Process.*, vol. 9, no. 6, pp. 1123–1129, Jun. 2000.

[11] M. D. Swanson, B. Zhu, and A. T. Tewfik, "Multiresolution scene-based video watermarking using perceptual models," *IEEE J. Select Areas Commun..*, vol. 16, no. 4, pp. 540–550, May 1998.

[12] 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. Eur. Conf. on Multimedia Applications, Services and Techniques*, 1998, pp. 190–206.

[13] J. Fridrich, "Robust bit extraction from images," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems* , vol. 2, June 1999, pp. 536–540.

[14] ——, "Visual hash for oblivious watermarking," *Proc. SPIE, Security and Watermarking of Multimedia Contents III*, vol. 4314, pp. 286–294, Jan. 2000.

[15] P.-C. Su, H.-J. M. Wang, and C.-C. J. Kuo, "Digital image watermarking in regions of interest," in *Proc, IS&T Processing/Image Quality/Image Capture Systems (PICS)*, Apr. 1999.

[16] F. Hartung, P. Eisert, and B. Girod, "Digital watermarking of MPEG-4 facial animation parameters," *Comput. Graph.*, vol. 22, no. 4, pp. 425–435, Jul.-Aug. 1998.

[17] J. K. Su, J. J. Eggers, and B. Girod, "Capacity of digital watermarks subjected to an optimal collusion attack," in *Proc. Eur. Signal Processing Conf.*, 2000.

[18] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack modeling: Toward a second generation benchmark," *Signal Process., Special Issue on Information Theoretical Issues in Digital Watermarking*, vol. 81, no. 6, pp. 1177–1214, Jun. 2001.

[19] P. Termont, L. De Strycker, J. Vandewege, M. O. de Beeck, J. Haitsma, T. Kalker, M. Maes, and G. Depovere, "How to achieve robustness against scaling in a real-time digital watermarking system for broadcast monitoring," in *Proc. IEEE Int. Conf. on Image Processing*, vol. 1, 2000, pp. 407–410.

[20] S.-C. Pei and C.-N. Lin, "Image normalization for pattern recognition," *Image Vis. Comput.*, vol. 13, no. 10, pp. 711–723, Dec. 1995.

[21] J. Black and L. J. Karam, "Automatic detection and extraction of perceptually significant visual features," in *Rec. 31st Asilomar Conf. Signals, Systems, and Computers*, vol. 1, Nov. 1997, pp. 315–319.

[22] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Englewood Cliffs, NJ: Prentice-Hall PTR, 1998.

[23] B. S. Manjunath, C. Shekhar, and R. Chellappa, "A new approach to image feature detection with applications," *Pattern Recognit.*, vol. 29, no. 4, pp. 627–640, Apr. 1996.

[24] K. Su, "Digital Video Watermarking Principles for Robustness to Collusion and Interpolation Attacks," M. S., Univ. Toronto, Toronto, ON, Canada, 2001.

[25] D. Kundur and D. Hatzinakos, "Attack characterization for effective watermarking," in *Proc. IEEE Int. Conf. on Image Processing*, vol. 4, 1999, pp. 240–244.

[26] ——, "Diversity and attack characterization for improved robust watermarking," *IEEE Trans. Signal Processing*, vol. 49, pp. 2383–2396, Oct. 2001.

[27] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking," *Lecture Notes in Computer Science*, vol. 1768, pp. 212–236, Sep. 2000.

[28] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," *Lecture Notes in Computer Science*, vol. 1525, pp. 219–239, Apr. 1998.

[29] T. Lane. JPEG Image Compression FAQ, Part 1 of 2. [Online]. Available: http://www.faqs.org/faqs/jpeg-faq/part1/

[30] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 3–19, Jan. 2000.

[31] F. Torkamani-Azar and K. E. Tait, "Image recovery using the anisotropic diffusion equation," *IEEE Trans. Image Process.*, vol. 5, no. 11, pp. 1573–1578, Nov. 1996.

[32] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal.Mach. Intelligence*, vol. 12, no. 7, pp. 629–639, Jul. 1990.

Karen Su (S'98) received the B.A.Sc. degree in electrical engineering, honors athematics option with distinction, from the University of British Columbia, Vancouver, BC, Canada, in 1999, and the M.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2001. She is currently pursuing the Ph.D. degree in engineering at the University of Cambridge, Laboratory for Communication Engineering, Cambridge, U.K. Her research interests are in the areas of coding theory, wireless communications, digital watermarking, and digital signal processing.

Deepa Kundur (S'93–M'99–SM'03) was born in Toronto, ON, Canada. She received the B.A.Sc., M.A.Sc., and Ph.D. degrees, all in electrical and computer engineering, in 1993, 1995, and 1999, respectively, from the University of Toronto.

In January 2003, she joined the Electrical Engineering Department at Texas A&M University, College Station, where she is a member of the Wireless Communications Laboratory and an Assistant Professor. From September 1999 to December 2002, she was an Assistant Professor at the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto where she was Bell Canada Junior Chair-holder in Multimedia. Her research interests include multimedia and network security for digital rights management, video cryptography, data hiding and steganography, covert communications, and nonlinear and adaptive information processing algorithms.

Dr. Kundur has been on numerous technical program committees and has given tutorials at ICME 2003 and Globecom 2003 in the area of digital rights management. She is a Guest Editor for the PROCEEDINGS OF THE IEEE Special Issue on Enabling Technologies for Digital Rights Management. She was the recipient of the 2002 Gordon Slemon Teaching of Design Award and the 2002 Best Electrical Engineering Professor Award (Spring) presented by the ECE Club at the University of Toronto.

Dimitrios Hatzinakos (M'90–SM'98) received the Diploma degree from the University of Thessaloniki, Greece, in 1983, the M.A.Sc degree from the University of Ottawa, Ottawa, ON, Canada, in 1986 and the Ph.D. degree from Northeastern University, Boston, MA, in 1990, all in electrical engineering.

In September 1990, he joined the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, where now he holds the rank of Professor with tenure. He also serves as Chair of the Communications Group of the Department since July 1, 1999. His research interests are in the areas of digital communications and signal processing with applications to wireless communications, image processing and multimedia. He has organized and taught many short courses on modern signal processing frameworks and applications devoted to continuing engineering education and given numerous seminars in the area of blind signal deconvolution. He is author/co-author of more than 120 papers in technical journals and conference proceedings and he has contributed to six books in his areas of interest. His experience includes consulting through Electrical Engineering Consociates Ltd. and contracts with United Signals and Systems Inc., Burns and Fry Ltd., Pipetronix Ltd., Defense Research Establishment Ottawa (DREO), Vaytek, Inc., Nortel Networks, and Vivosonic, Inc.

Dr. Hatzinakos served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1998 through 2002 and was Guest Editor for the special issue on Signal Processing Technologies for Short Burst Wireless Communications for Elsevier's *Signal Processing*, in October 2000. He was a member of the IEEE Statistical Signal and Array Processing Technical Committee (SSAP) from 1992 through 1995 and Technical Program co-Chair of the 5th Workshop on Higher-Order Statistics in July 1997. He was co-organizer and Technical program co-Chair of the IEEE Toronto Centenial Workshop on Wireless Communications, held in Toronto in October 2003. He is a member of EURASIP, the Professional Engineers of Ontario (PEO), and the Technical Chamber of Greece.