# Multiresolution Scene-Based Video Watermarking Using Perceptual Models

Mitchell D. Swanson, *Member, IEEE*, Bin Zhu, *Member, IEEE*, and Ahmed H. Tewfik, *Fellow, IEEE*

***Abstract*— We present a watermarking procedure to embed copyright protection into digital video. Our watermarking procedure is scene-based and video dependent. It directly exploits spatial masking, frequency masking, and temporal properties to embed an invisible and robust watermark. The watermark consists of static and dynamic temporal components that are generated from a temporal wavelet transform of the video scenes. The resulting wavelet coefficient frames are modified by a perceptually shaped pseudorandom sequence representing the author. The noise-like watermark is statistically undetectable to thwart unauthorized removal. Furthermore, the author representation resolves the deadlock problem. The multiresolution watermark may be detected on single frames without knowledge of the location of the frames in the video scene. We demonstrate the robustness of the watermarking procedure to several video degradations and distortions.**

***Index Terms*— Copyright protection, data hiding, perceptual masking, video watermarking.**

## I. Introduction

DIGITAL media, e.g., images, audio, and video, are readily manipulated, reproduced, and distributed over information networks. These efficiencies lead to problems regarding copyright protection. As a result, creators and distributors of digital data are hesitant to provide access to their digital intellectual property. Technical solutions for copyright protection of multimedia data are actively being pursued.

Digital watermarking has been proposed as a means to identify the owner and distribution path of digital data. Watermarking is the process of encoding hidden copyright information into digital data by making small modifications to the data samples, e.g., pixels. Many watermark algorithms have been proposed. Some techniques modify spatial/temporal data samples (e.g., [1]–[5]), while others modify transform coefficients (e.g., [2], [6]–[10]). Unlike encryption, watermarking does not restrict access to the data. Once encrypted data is decrypted, intellectual property rights are no longer protected. A watermark is designed to *permanently* reside in the host data. When the ownership of data is in question, the information can be extracted to completely characterize the owner or distribution path.

In this paper, we present a novel multiresolution video watermarking scheme. The watermarking procedure explicitly exploits the human visual system (HVS) to guarantee that the embedded watermark is imperceptible. Similar to our image and audio watermarking procedures based on perceptual models [11], [12], the video watermark adapts to and is highly dependent on the video being watermarked. This guarantees an invisible and robust watermark.

Watermarking digital video introduces some issues that generally do not have a counterpart in images and audio. Due to large amounts of data and inherent redundancy between frames, video signals are highly susceptible to pirate attacks, including frame averaging, frame dropping, frame swapping, collusion, statistical analysis, etc. Many of these attacks may be accomplished with little or no damage to the video signal. However, the watermark may be adversely effected. Scenes must be embedded with a consistent and reliable watermark that survives such pirate attacks. Applying an identical watermark to each frame in the video leads to problems of maintaining statistical invisibility. Furthermore, such an approach is neccessarily video *independent*, as the watermark is fixed. Applying independent watermarks to each frame also is a problem. Regions in each video frame with little or no motion remain the same frame after frame. Motionless regions in successive video frames may be statistically compared or averaged to remove independent watermarks.

We employ a watermark that consists of fixed and varying components. The components are generated from a temporal wavelet transform representation of each video scene. A wavelet transform applied along the temporal axis of the video results in a *multiresolution temporal representation* of the video. In particular, the representation consists of temporal low-pass frames and high-pass frames. The low-pass frames consist of the static components in the video scene. The high-pass frames capture the motion components and changing nature of the video sequence. Our watermark is designed and embedded in each of these components. The watermarks embedded in the low-pass frames exist throughout the *entire* video scene. The watermarks embedded in the motion frames are highly localized in time and change rapidly from frame to frame. Thus, the watermark is a composite of *static* and *dynamic* components. The combined representation overcomes the aforementioned drawbacks associated with a fixed or independent watermarking procedure. Futhermore, averaging frames simply damages the dynamic watermark components. As shown in Section VIII, the static components survive such attacks and are easily recovered for copyright verification.

To generate a watermark, the visual masking properties of the wavelet coefficient frames are computed and used to filter (i.e., shape) a pseudorandom sequence that represents the author or distribution path of the video. Based on pseudo-

Manuscript received March 1997; revised July 1997. This work was supported by the Air Force Office of Scientific Research under Grant AF/F49620-94-1-0461.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: mswanson@ece.umn.edu; binzhu@ece.umn.edu; tewfik@ece.umn.edu).

Publisher Item Identifier S 0733-8716(98)01106-8.

random sequences, the noise-like watermark is statistically undetectable, thereby helping thwart pirate attacks. Furthermore, due to the combined static and dynamic watermark representation, the watermark is readily extracted from a *single frame* of the video *without knowledge of the location* of that particular frame in the video, even after printing and rescanning.

The work presented in this paper offers several major contributions to the field.

1) *A Perception-Based Video Watermarking Procedure:* Our watermark *adapts* to each individual video signal. In particular, the temporal and frequency distributions of the watermark are controlled by the masking characteristics of the host video signal. As a result, the strength of the watermark increases and decreases with host, e.g., higher amplitude in regions of the video with more textures, edges, and motion. This guarantees that the embedded watermark is invisible while having the maximum possible robustness.
2) *A Scene-Based Multiscale Watermark Representation:* Given one or more frames from a potentially pirated video, the watermark may be extracted from the frames without knowledge of the location of the frame being tested. This detection characteristic exists due to the combined static and dynamic representation of the watermark.
3) *An Author Representation Which Solves the Deadlock Problem:* The author or owner of the video is represented with a pseudorandom sequence created by a pseudorandom generator [13] and *two* keys. One key is *author* dependent, while the second key is *signal* dependent. The representation is able to resolve rightful ownership in the face of multiple ownership claims.
4) *A Dual Watermark:* The watermarking scheme introduced in this paper uses the original video signal to detect the presence of a watermark. The procedure can handle virtually *all* types of distortions, including cropping, temporal rescaling, frame dropping, etc., using a generalized likelihood ratio test. We integrate this procedure with a second watermark which does *not* require the original signal to address the deadlock problem.

In the next section, we introduce our author representation and dual watermarking scheme. Our frequency and spatial masking models are reviewed in Section III. The wavelet transform is reviewed in Section IV. Our watermarking design and detection algorithms are introduced in Sections V and VI. Finally, experimental results are presented. In Section VII, watermark statistics and fidelity results for two test videos are presented. In Section VIII, the robustness of our watermarking procedure is illustrated for an assortment of signal processing operations and distortions including colored noise, MPEG coding, multiple watermarks, frame dropping, and printing and scanning. We present our conclusion in Section IX.

## II. Author Representation and the Deadlock Problem

The main function of a video watermarking algorithm is to unambiguously establish and protect ownership of video data. Unfortunately, most current watermarking schemes are unable to resolve rightful ownership of digital data when multiple ownership claims are made, i.e., when a deadlock problem arises [14]. The inability to deal with deadlock is independent of how the watermark is inserted in the video data or how robust it is to various types of modifications.

Watermarking techniques that do *not* require the original (nonwatermarked) signal are the most vulnerable to ownership deadlocks. A pirate simply adds his or her watermark to the watermarked data. The data now has two watermarks. Current watermarking schemes are unable to establish who watermarked the data first.

Watermarking procedures that require the original data set for watermark detection also suffer from deadlocks. In such schemes, a party other than the owner may counterfeit a watermark by "subtracting off" a second watermark from the publicly available data and claim the result to be his or her original. This second watermark allows the pirate to claim copyright ownership because he or she can show that both the publicly available data and the original of the rightful owner contain a copy of their counterfeit watermark.

To understand how our procedure solves the deadlock problem, let us assume that two parties claim ownership of a video. To determine the rightful owner of the video, an arbitrator examines only the video in question, the originals of both parties and the key used by each party to generate their watermark.

We use a two-step approach to resolve deadlock: Dual watermarks and a video dependent watermarking scheme. Our dual watermark employs a *pair* of watermarks. One watermarking procedure requires the original data set for watermark detection. This paper provides a detailed description of that procedure and of its robustness. The second watermarking procedure does *not* require the original data set and hence is a simple data hiding procedure. Any watermarking technique which satisfies the restrictions outlined in [15] can be used to insert the second watermark. The technique must be secure to prevent unauthorized removal of the second watermark. For example, it may employ a secret key to embed the watermark. However, the second watermark need not be highly robust to editing of the video since, as we shall see below, it is meant to protect the video clip that the pirate claims to be his *original*. The robustness levels of most of the recent watermarking techniques that do not require the original for watermark detection are quite adequate. The arbitrator would expect the original to be of a high enough quality. This limits the operations that a pirate can apply to a video and still claim it to be his high-quality original. The watermark that requires the original video sequence for its detection is very robust as we show in this paper.

In case of deadlock, the arbitrator simply checks first for the watermark that requires the original for watermark detection. If the pirate is clever and has used the attack suggested in [14] and outlined above, the arbitrator would be unable to resolve the deadlock with this first test. The arbitrator simply then checks for the watermark that *does not* require the original video sequence in the video segments that each ownership contender claims to be his *original*. Since the original video sequence of a pirate is derived from the watermarked copy produced by the rightful owner, it

will contain the watermark of the rightful owner. On the other hand, the true original of the rightful owner will not contain the watermark of the pirate since the pirate has no access to that original and the watermark does not require subtraction of another data set for its detection.

Further protection against deadlock is provided by the technique that we use to select the pseudorandom sequence that represents the author. This technique is similar to an approach developed independently by [15]. Both techniques solve the shortcomings of the solution proposed in [14] for solving the deadlock problem.

Specifically, the author has two random keys $x_1$ and $x_2$ (i.e., seeds) from which a pseudorandom sequence $y$ can be generated using a suitable pseudorandom sequence generator [13]. Popular generators include RSA, Rabin, Blum/Micali, and Blum/Blum/Shub [16]. With the two proper keys, the watermark may be extracted. Without the two keys, the data hidden in the video is statistically undetectable and impossible to recover. Note that we do *not* use the classical maximal length pseudonoise sequence (i.e., $m$-sequence) generated by linear feedback shift registers to generate a watermark. Sequences generated by shift registers are cryptographically insecure: One can solve for the feedback pattern (i.e., the keys) given a small number of output bits $y$.

The noise-like sequence $y$, after some processing (cf. Section V), is the actual watermark hidden into the video stream. The key $x_1$ is *author* dependent. The key $x_2$ is *signal* dependent. The key $x_1$ is the secret key assigned to (or chosen by) the author. Key $x_2$ is *computed from the video signal* that the author wishes to watermark. It is computed from the video using a one-way hash function. In particular, the tolerable error levels supplied by the masking models (cf. Section III) are hashed to a key $x_2$. Any one of a number of well-known secure one-way hash functions may be used to compute $x_2$, including RSA, MD4, and SHA [13], [16]. For example, the Blum/Blum/Shub pseudorandom generator uses the one way function $y = g_n(x) = x^2 \bmod n$ where $n = pq$ for primes $p$ and $q$ so that $p = q = 3 \bmod 4$. It can be shown that generating $x$ or $y$ from partial knowledge of $y$ is *computationally infeasible* for the Blum/Blum/Shub generator.

The signal dependent key $x_2$ makes counterfeiting very difficult. The pirate can only provide key $x_1$ to the arbitrator. Key $x_2$ is automatically computed by the watermarking algorithm from the original signal. The pirate generates a counterfeit original by subtracting off a watermark. However, the watermark (partially generated from the signal dependent key) *depends* on the counterfeit original. Thus, the pirate must generate a watermark which creates a counterfeit original which, in turn, generates the watermark! As it is computationally infeasible to invert the one-way hash function, the pirate is unable to fabricate a counterfeit original that generates the desired watermark.

## III. Visual Masking

We use image masking models based on the HVS to ensure that the watermark embedded into each video frame is perceptually invisible and robust. Visual masking refers to a situation where a signal raises the visual threshold for other signals around it. Masking characteristics are used in high-quality, low-bit rate coding algorithms to further reduce bit rates [17]. The masking models reviewed here are based on image models. A detailed discussion of the models may be found in [18]. We are currently developing a watermarking algorithm that takes temporal masking (e.g., [19]) into account.

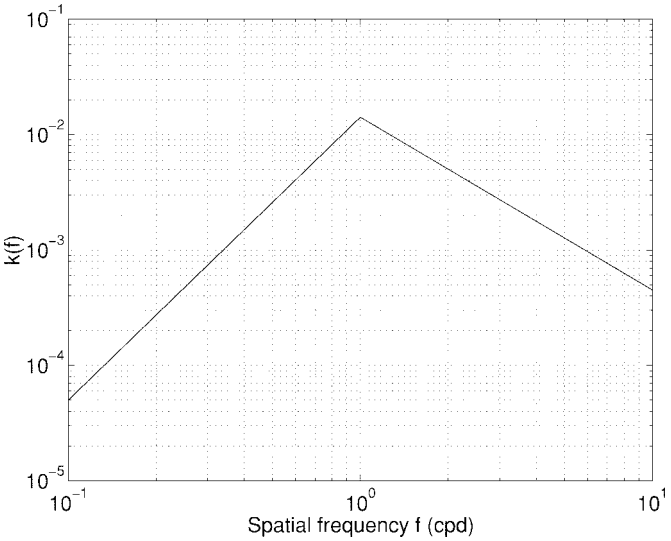



Fig. 1. The masking characteristic function $k(f)$.

### A. Frequency Masking

Our frequency masking model is based on the knowledge that a masking grating raises the visual threshold for signal gratings around the masking frequency [20]. The model we use [18], based on the discrete cosine transform (DCT), expresses the contrast threshold at frequency $f$ as a function of $f$, the masking frequency $f_m$, and the masking contrast $c_m$

$$c(f, f_m) = c_0(f) \cdot \max\{1, [k(f/f_m)c_m]^\alpha\} \tag{1}$$

where the detection threshold at frequency $f, c_0(f)$, and $\alpha = 0.62$ are determined by psychovisual tests [20]. The mask weighting function $k(f)$ is shown in Fig. 1.

To find the contrast threshold $c(f)$ at a frequency $f$ in an image, we first use the DCT to transform the image into the frequency domain and find the contrast at each frequency. Then, we use a summation rule of the form

$$c(f) = \left[\sum_{f_m} c(f, f_m)^2\right]^{1/2} \tag{2}$$

to sum up the masking effects from all the masking signals near $f$. If the contrast error at $f$ is less than $c(f)$, the model predicts that the error is invisible to human eyes.

### B. Spatial Masking

Our spatial masking model is based on the threshold vision model proposed by Girod [19]. The model accurately predicts the masking effects near edges and in uniform background. Assuming that the modifications to the image are small, the upper channel of Girod's model can be linearized [18] to

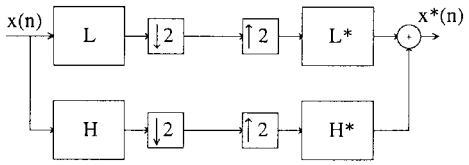



Fig. 2. Diagram of 2-band filter bank.

obtain the tolerable error level for each coefficient. This is a reasonable assumption for transparent watermarking.

Under certain simplifying assumptions [18], the tolerable error level for a pixel $p(x,y)$ can be obtained by first computing the contrast saturation at $(x,y)$

$$dc_{\text{sat}}(x,y) = dc_{\text{sat}} = \sqrt{\frac{T}{\sum_{x',y'} w_4(0,0,x',y')}} \tag{3}$$

where the weight $w_4(x,y,x',y')$ is a Gaussian centered at the point $(x,y)$. Parameter $T$ is a threshold based on a series of psychovisual tests. Once $dc_{\text{sat}}(x,y)$ is computed, the luminance on the retina, $dl_{\text{ret}}$, is obtained from the equation

$$dc_{\text{sat}}(x,y) = w_2(x,y) \cdot dl_{\text{ret}}(x,y). \tag{4}$$

From $dl_{\text{ret}}$, the tolerable error level $ds(x,y)$ for the pixel $p(x,y)$ is computed from

$$dl_{\text{ret}}(x,y) = w_1(x,y) \cdot ds(x,y). \tag{5}$$

The weights $w_1(x,y)$ and $w_2(x,y)$ are based on Girod's model. The masking model predicts that changes to pixel $p(x,y)$ less than $ds(x,y)$ introduce no perceptible distortion.

## IV. Temporal Wavelet Transform

A wavelet transform [21], [22] is a powerful tool employed to represent signals at multiple resolutions. The multiresolution nature of a wavelet decomposition provides signal specific information localized in time, space, or frequency that can be exploited for signal analysis and processing.

We employ the wavelet transform along the *temporal axis* of the video sequence. The wavelet transform is used to provide a compact *multiresolution temporal representation* of a video, leading to static and dynamic video components. A wavelet transform can be computed using a two-band perfect reconstruction filter bank as shown in Fig. 2. The video signal is simultaneously passed through low-pass $L$ and high-pass $H$ filters and then decimated by 2 to give *static* (no motion) and *dynamic* (motion) components of the original signal. The two decimated signals may be upsampled and passed through complementary filters and summed to reconstruct the original signal. Wavelet filters are widely available [22].

In Fig. 3, we show an example of the temporal wavelet transform. The top row consists of four consecutive frames from a sample "Football" video. The bottom row consists of the four temporal wavelet coefficient frames computed from the original "Football" sequence. The two temporal low-pass frames (bottom left) represent the static components of the "Football" frames. The detail, i.e., dynamic, components are represented by the two temporal high-pass frames (bottom right) in Fig. 3. Note that a filter bank may be cascaded with additional filter banks to provide further temporal resolutions

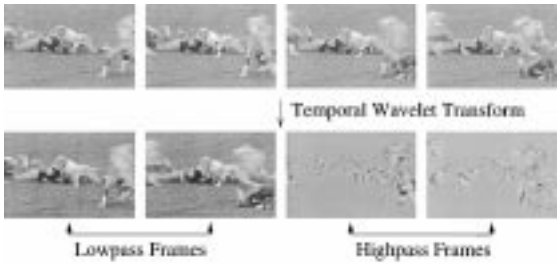



Fig. 3. Example of temporal wavelet transform.

of the input signal. The output of the cascaded filter banks consists of multiple temporal resolutions of the input.

## V. Watermark Design

The first step in our watermarking algorithm consists of breaking the video sequence into *scenes* [23]. Recall from the Section I that segmentation into scenes allows the watermarking procedure to take into account temporal redundancy. Visually similar regions in the video sequence, e.g., frames from the same scene, must be embedded with a consistent watermark. To address assorted pirate attacks on the watermark, we perform a temporal wavelet transform on the video scenes (cf. Section IV). The multiresolution nature of the wavelet transform allows the watermark to exist across *multiple temporal scales*, resolving the above-mentioned pirate attacks. For example, the embedded watermark in the lowest frequency (DC) wavelet frame exists in *all frames* in the scene.

We denote indexed temporal variables by capital letters with subscripts, e.g., the $i$th frame $F_i$ in a video scene. Frames are ordered sequentially according to time. The tilde representation is used to denote a wavelet representation, e.g., $\tilde{F}_i$ is the $i$th wavelet coefficient frame. Without loss of generality, wavelet frames are ordered from lowest frequency to highest frequency, i.e., $\tilde{F}_0$ is a DC frame. Finally, primed capital letters, e.g., $F'_i$, denote the DCT representation of an indexed variable.

In Fig. 4, we show our video-watermarking procedure. Consider a scene of $k$ frames from the video sequence. Let each frame from the scene be of size $n \times m$. The video may be grayscale (8 b/pixel) or color (24 b/pixel). Let $F_i$ denote the frames in the scene, where $i = 0, \cdots, k-1$. Initially, we compute the wavelet transform of the $k$ frames $F_i$ to obtain $k$ *wavelet coefficient frames* $\tilde{F}_i, i = 0, \cdots, k-1$. The watermark is constructed and added to the video using the following steps:

1) segment each wavelet frame $\tilde{F}_i$ into $8 \times 8$ blocks $\tilde{B}_{ij}, i = 0, 1, \cdots, \lfloor n/8 \rfloor$ and $j = 0, 1, \cdots, \lfloor m/8 \rfloor$;
2) for each block $\tilde{B}_{ij}$:
   a) compute the DCT $\tilde{B}'_{ij}$ of the frame block $\tilde{B}_{ij}$;
   b) compute the frequency mask $M'_{ij}$ (cf. Section III-A) of the DCT block $\tilde{B}'_{ij}$;
   c) use the mask $M'_{ij}$ to weight the noise-like author $Y'_{ij}$ (cf. Section II) for that frame block, creating the frequency-shaped author signature $P'_{ij} = M'_{ij} Y_{ij}$;
   d) create the wavelet coefficient watermark block $\tilde{W}_{ij}$ by computing the inverse DCT of $P'_{ij}$ and locally

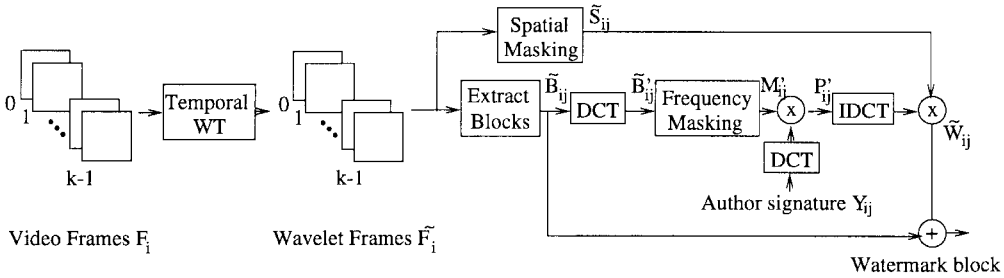



Fig. 4. Diagram of video watermarking procedure.

increase the watermark to the maximum tolerable error level provided by the spatial mask $\tilde{S}_{ij}$;

e) add the watermark $\tilde{W}_{ij}$ to the block $\tilde{B}_{ij}$, creating the watermarked block;

3) repeat for each wavelet coefficient frame $\tilde{F}_i$.

The watermark for each wavelet coefficient frame is the block concatenation of all the $8 \times 8$ watermark blocks $\tilde{W}_{ij}$ for that frame. The wavelet coefficient frames with the embedded watermarks are then converted back to the temporal domain using the inverse wavelet transform. As the watermark is designed and embedded in the wavelet domain, the individual watermarks for each wavelet coefficient frame are spread out to varying levels of support in the temporal domain. For example, watermarks embedded in highpass wavelet frames are localized temporally. Conversely, watermarks embedded in lowpass wavelet frames are generally located throughout the scene in the temporal domain.

## VI. Watermark Detection

The watermark is designed to be easily extracted by the owner, even when signal processing operations are applied to the host video. As the embedded watermark is noise-like, a pirate has insufficient knowledge to directly remove the watermark. Therefore, any destruction attempts are done blindly. Unlike other users, the owner has a copy of the original video and the noise-like author signature which was embedded into the video. Typically, the owner is presented with one or more video frames which he or she wishes to prove ownership rights. Two methods have been developed to extract the potential watermark from a test video or test video frame. Both employ hypothesis testing [24]. One test employs index knowledge during detection, i.e., we know the placement of the test video frame(s) relative to the original video. The second detection method does *not* require knowledge of the location of the test frame(s). This is extremely useful in a video setting, where thousands of frames may be similar, and we are uncertain where the test frames reside.

*Detection I—Watermark Detection with Index Knowledge:* When the location of the test frame is known, a straightforward hypothesis test may applied. For each frame in the test video $R_k$, we perform a hypothesis test

$$\begin{aligned} \mathbf{H}_0&: X_k = R_k - F_k = N_k \quad \text{(No watermark)} \\ \mathbf{H}_1&: X_k = R_k - F_k = W_k^* + N_k \quad \text{(Watermark)} \end{aligned} \tag{6}$$

where $F_k$ is the original frame, $W_k^*$ is the (potentially modified) watermark recovered from the frame, and $N_k$ is noise. The hypothesis decision is obtained by computing the scalar similarity between each extracted signal $X_k$ and original watermark $W_k$

$$S_k = \mathrm{sim}_k(X_k, W_k) = \frac{X_k \cdot W_k}{W_k \cdot W_k}. \tag{7}$$

The overall similarity between the extracted and original watermark is computed as the mean of $S_k$ for all $k$: $S = \mathrm{mean}(S_k)$. The overall similarity is compared with a threshold to determine whether the test video is watermarked. As shown in Section VIII, our experimental threshold is chosen around 0.1, i.e., a similarity value $\geq 0.1$ indicates the presence of the owner's copyright. In such a case, the video is deemed the property of the author, and a copyright claim is valid. A similarity value $<0.1$ indicates the absence of a watermark.

When the length (in terms of frames) of the test video is the same as the length of the original video, we perform the hypothesis test in the *wavelet domain*. A temporal wavelet transform of the test video is computed to obtain its wavelet coefficient frames $\tilde{R}_k$. Substituting wavelet transform values in (6)

$$\begin{aligned} \mathbf{H}_0&: \tilde{X}_k = \tilde{R}_k - \tilde{F}_k = N_k \quad \text{(No watermark)} \\ \mathbf{H}_1&: \tilde{X}_k = \tilde{R}_k - \tilde{F}_k = \tilde{W}_k^* + N_k \quad \text{(Watermark)} \end{aligned} \tag{8}$$

where $\tilde{F}_k$ are the wavelet coefficient frames from the original video, $\tilde{W}_k^*$ is the potentially modified watermarks from each frame, and $N_k$ is noise. This test is performed for each wavelet frame to obtain $\tilde{X}_k$ for all $k$. Similarity values are computed as before: $S_k = \mathrm{sim}_k(\tilde{X}_k, \tilde{W}_k)$.

Using the original video signal to detect the presence of a watermark, we can handle virtually *all* types of distortions, including cropping, rotation, rescaling, etc., by employing a generalized likelihood ratio test [24]. We have also developed a second detection scheme that is capable of recovering a watermark after many distortions *without* a generalized likelihood ratio test. The procedure is fast and simple, particularly when confronted with the large amount of data associated with video.

*Detection II—Watermark Detection Without Index Knowledge:* In many cases, we may have no knowledge of the indexes of the test frames. Pirate tampering may lead to many types of derived videos which are often difficult to process. For example, a pirate may steal *one* frame from a video. A pirate may also create a video that is *not* the same length as the original video. Temporal cropping, frame dropping, and frame interpolation are all examples. A pirate may also swap the order of the frames. Most of the better watermarking schemes currently available use different watermarks for different images. As such, they generally require knowledge of *which*

frame was stolen. If they are unable to ascertain which frame was stolen, they are unable to determine which watermark was used.

Our second method can extract the watermark *without* knowledge of where a frame belongs in the video sequence. No information regarding cropping, frame order, interpolated frames, etc., is required! As a result, no searching and correlation computations are required to locate the test frame index. The hypothesis test is formed by removing the *low temporal wavelet frame* from the test frame and computing the similarity with the watermark for the low temporal wavelet frame. The hypothesis test is formed as

$$\begin{aligned} \mathbf{H}_0\text{: } X_k &= R_k - \tilde{F}_0 = N_k \quad \text{(No watermark)} \\ \mathbf{H}_1\text{: } X_k &= R_k - \tilde{F}_0 = \tilde{W}_k^* + N_k \quad \text{(Watermark)} \end{aligned} \tag{9}$$

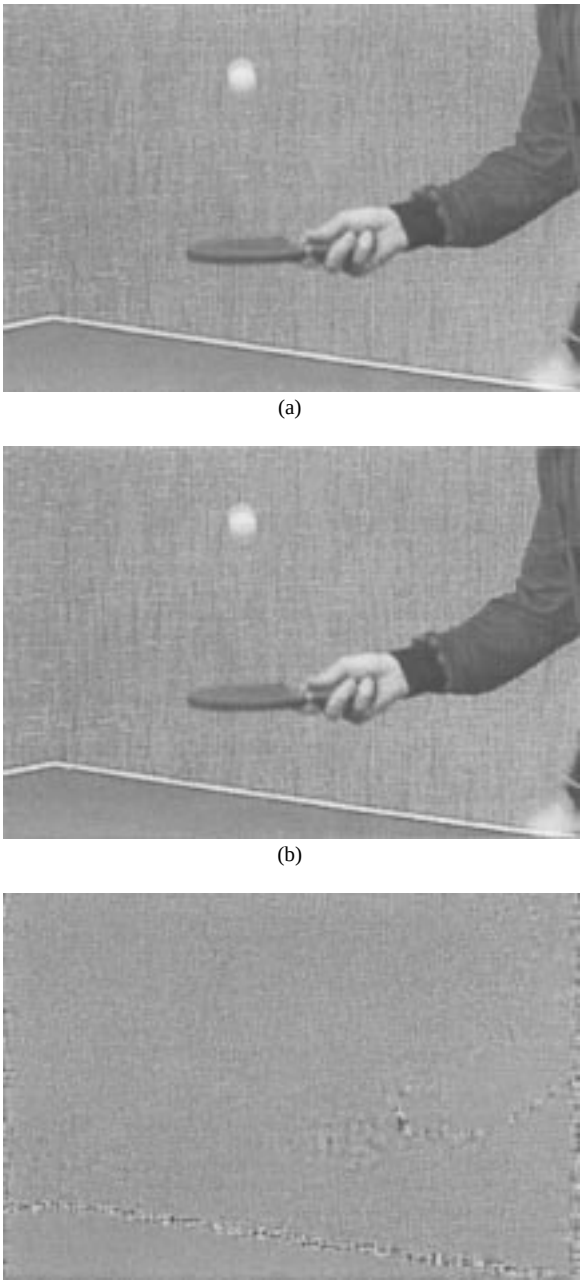
where $R_k$ is the test frame in the *spatial* domain and $\tilde{F}_0$ is the lowest temporal *wavelet* frame. The hypothesis decision is made by computing the scalar similarity between each extracted signal $X_k$ and original watermark for the low temporal wavelet frame $\tilde{W}_0$: $\mathrm{sim}_k(X_k, \tilde{W}_0)$. This simple yet powerful approach exploits the wavelet property of varying temporal support.

## VII. Visual Results

We illustrate the invisibility and robustness of our watermarking scheme on two grayscale (8 bpp) videos: "Pingpong" and "Football." Each frame is of size 240 × 352. An original frame from each video is shown in Figs. 5(a) and 6(a). The corresponding watermarked frame for each is shown in Figs. 5(b) and 6(b). In both cases, the watermarked frame appears visually identical to the original. In Figs. 5(c) and 6(c), the watermark for each frame, scaled to graylevels for display, are shown. Although the watermarks are computed on the *wavelet* frames, we display them in the *spatial* domain for visual convenience. The watermark for each frame is the same size as the host frame, i.e., 240 × 352. For each frame, the watermark values corresponding to smoother background regions are generally smaller than watermark values near motion and edge regions. This is to be expected, as motion and edge regions have more favorable masking characteristics.

Some statistical properties for each of the watermarks are shown in Table I. The values are computed for the frames presented in Figs. 5 and 6, which are representative of the watermarks for the other frames in each of the videos. The maximum and minimum values are in terms of the watermark values over the 240 × 352 watermark. Peak signal-to-noise ratio (PSNR), a common image quality metric, is defined as $20\log_{10}(255/\sqrt{SNR})$. The signal-to-noise ratio (SNR) is computed between the original and watermarked frame.

To determine the quality of the watermarked videos, we performed a series of informal visual tests. For each test video, we displayed the original to the viewer. Then two randomly selected videos "A" and "B" were sequentially displayed to the viewer. The ordered pair was randomly selected as (original, watermarked) or (watermarked, original). The viewer was asked to select the video "A" or "B" that was visually more pleasing. This test was performed ten times for each video. A total of ten viewers (not including the authors) took part in the blind test. The results of the test are displayed in Table II. As predicted by the visual-masking models, the original and watermarked videos appeared visually similar and each was preferred approximately 50% of the time. We conclude that the watermark causes no degradations to the host video.



(a) (b) (c)

Fig. 5. Frame from Pingpong video (a) original, (b) watermarked, and (c) watermark.

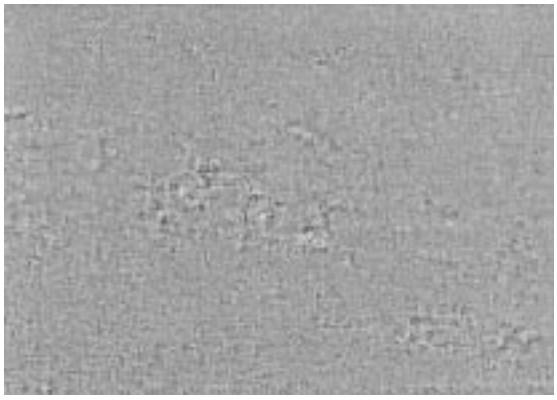
## VIII. Robustness Results

To be effective, the watermark must be robust to incidental and intentional signal distortions incurred by the host video. Clearly, any lossy signal operations performed on the host video effect the embedded watermark.

The robustness of our watermarking approach is measured by the ability to *detect a watermark when one is present in*

(a)

(b)



(c)

Fig. 6. Frame from Football video (a) original, (b) watermarked, and (c) watermark.

*the video*. Robustness is further based on the ability of the algorithm *to reject a video when a watermark is not present*. For a given distortion, the overall performance may be ascertained by the relative difference between the similarity when a watermark is present (hypothesis $\boldsymbol{H}_1$) and the similarity when a watermark is not present (hypothesis $\boldsymbol{H}_0$). In each robustness experiment, similarity results were obtained for both hypotheses. In particular, the degradation was applied to the video when a watermark was present. It was also applied to the video when a watermark was *not* present. The similarity was computed between the original watermark and the recovered signal (which may or may not have a watermark). A large similarity indicates the presence of a watermark ($\boldsymbol{H}_1$), while a low similarity suggests the lack of a watermark ($\boldsymbol{H}_0$). As shown in our tests, no overlap between the hypotheses occurred during the degradations and distortions. This indicates a high probability of detection and a low probability of false alarm.

TABLE I
STATISTICAL PROPERTIES OF THE VIDEO WATERMARK

| Video | Maximum | Minimum | Variance | PSNR (dB) |
|---|---|---|---|---|
| Pingpong | 42 | -44 | 11.20 | 37.64 |
| Football | 43 | -47 | 14.44 | 36.54 |

TABLE II
BLIND TESTING OF WATERMARKED VIDEOS

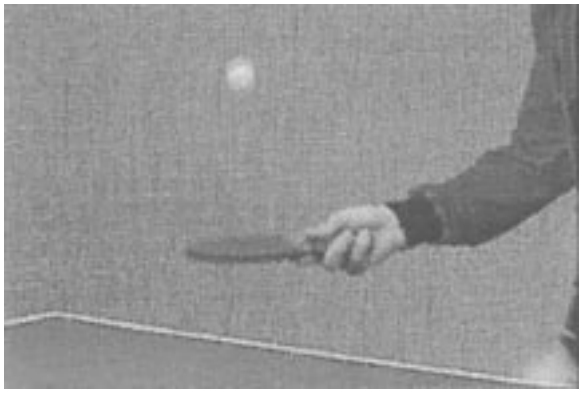
| Video | Preferred original to watermarked |
|---|---|
| Pingpong | 48.5 % |
| Football | 50.5 % |

We use the first 32 frames from each video for our tests. Both detection approaches were performed during each experiment. Specifically, we performed detection when the entire test sequence was available and the indexes known (8). We also performed detection on a frame-by-frame basis without knowledge of the frame index (9). In this case, we assume that the index of the frame is *unknown*, so we do not know the location of the frame in the video.

### A. Colored Noise

To model perceptual coding techniques, we corrupted the watermark with *worst case colored noise that follows the visual masks.* Colored noise was generated by shaping (i.e., multiplying) white noise with the frequency and spatial masks for the video. As the colored noise is generated in the same fashion as the watermark, it acts like another *interfering* watermark. We generated colored noise and added it to the video with and without the watermark. The variance of the noise for each test sequence was chosen nine times greater than the watermark embedded in the video. For example, the average variance of the watermark over all frames from the Football sequence is 14.0. The colored-noise sequence was constructed with a variance of approximately 126.0 ($\mathrm{PSNR} = 27.1$ dB). A noisy frame from each of the watermarked videos is shown in Fig. 7(a) and (b). The noisy frames correspond to those shown in Figs. 5 and 6.

For each video, this testing process was repeated 100 times with a new noise sequence for each run. In the first test, we use all of the frames in the video for detection (Detection I). The similarity values for each video sequence with and without the watermark are shown in Table III. The maximum, mean, and minimum similarity values are computed over all 100 noise runs. It is important to note that the minimum similarity values with watermark are much larger than the maximum similarity values without watermark. An overlap between the two indicates possible errors in detection. In this case, for example, the minimum similarity value of the Pingpong sequence with watermark is 0.91, which is much larger than the maximum value of 0.03 without watermark. As a result, one may readily decide whether a watermark exists in the video. Selecting a decision threshold $T$ somewhere in the



(a)

(b)

Fig. 7. Frame from videos with colored noise (PSNR = 25.1 dB) (a) Pingpong and (b) Football.

TABLE III
SIMILARITY RESULTS FOR PING-PONG AND FOOTBALL WITH COLORED NOISE

| Video | PSNR (dB) | With watermark | | | No watermark | | |
|---|---|---|---|---|---|---|---|
| | | Max | Mean | Min | Max | Mean | Min |
| Ping-Pong | 27.8 | 1.00 | 0.96 | 0.91 | 0.03 | 0.00 | -0.02 |
| Football | 27.1 | 1.00 | 0.97 | 0.93 | 0.04 | 0.00 | -0.03 |

range of approximately $0.1 \leq T \leq 0.9$ guarantees a correct hypothesis decision for these test videos in colored noise.

We also performed testing on a frame-by-frame basis without knowledge of the frame index (Detection II). Detection was performed by removing the lowpass temporal frame $\tilde{F}_0$ from the test frame, and correlating the result with the watermark $\tilde{W}_0$ corresponding to $\tilde{F}_0$. The similarity values obtained during testing indicate easy discrimination between the two hypotheses as shown in Fig. 8. The upper similarity values in each plot corresponds to each frame with a watermark. The lower similarity curve correspond to each frame without a watermark. The error bars around each similarity value indicate the *maximum* and *minimum* similarity values over the 100 runs. The x-axis corresponds to frame number and runs from 0 to 31. Observe that the upper value is widely separated from the lower value for each frame. An error-free hypothesis decision is easy to obtain without knowledge of the position of the frame in the video scene.

In all of the following distortion experiments, we add colored noise to each video *prior* to distortion (e.g., coding, printing, and scanning, etc.). The colored noise is used

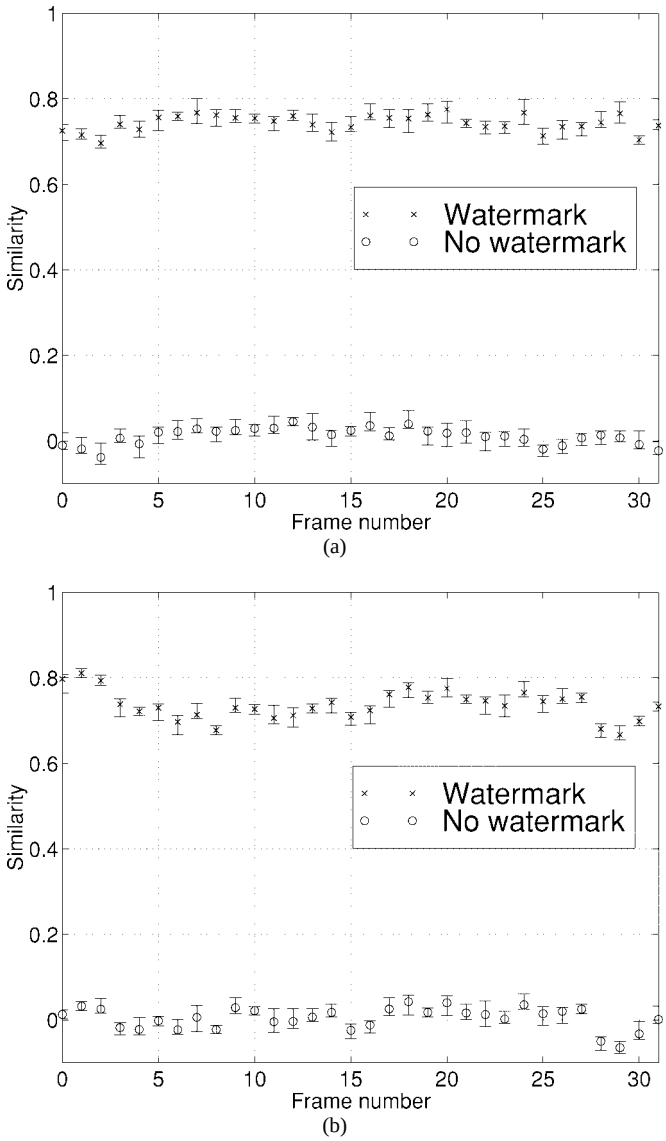


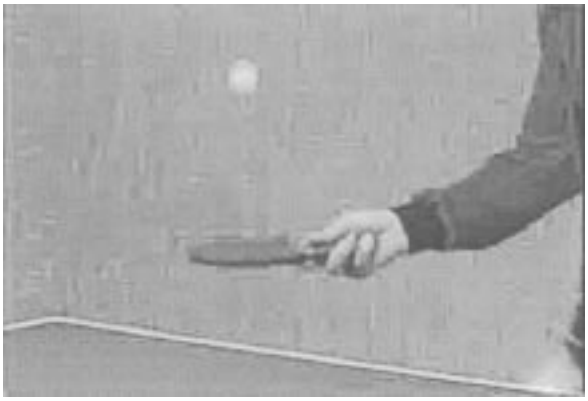

Fig. 8. Similarity values versus frame number in colored noise (a) Pingpong and (b) Football. The error bars around each similarity value indicate the maximum and minimum similarity values over the 100 runs.

to simulate additional attacks on the watermark, including masking-based coders and *other watermarks*. The strength of the colored noise is approximately the same as that of the watermark and is not visible.

### B. Coding

In most applications involving storage and transmission of digital video, a lossy coding operation is performed on the video to reduce bit rates and increase efficiency. We tested the ability of the watermark to survive MPEG coding [25] at very low quality. In our experiment, we set the MPEG tables at the coarsest possible quantization levels to maximize compression.

A watermarked Pingpong video frame coded at 0.08 bpp is shown in Fig. 9(a). The corresponding compression ratio (CR) is 100:1. The original (noncoded) frame is shown in Fig. 9(b). Note that a large amount of distortion is present in the frame. Using the same quantizatiion tables, a frame from the Football video at 0.18bpp (CR 44:1) is shown in Fig. 9(b). Note that the



(a)

(b)

Fig. 9. MPEG coded frame from (a) Pingpong (0.08 b/pixel, CR 100:1) and (b) Football (0.18 b/pixel, CR 44:1).

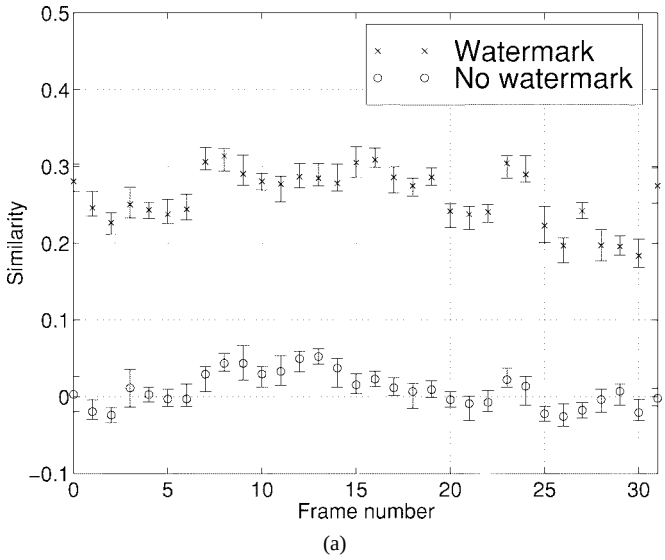
TABLE IV
SIMILARITY RESULTS FOR PING-PONG AND FOOTBALL AFTER MPEG CODING

| Video | CR | PSNR (dB) | With watermark | | | No watermark | | |
|---|---|---|---|---|---|---|---|---|
| | | | Max | Mean | Min | Max | Mean | Min |
| Ping-Pong | 100:1 | 26.8 | 0.41 | 0.35 | 0.28 | 0.06 | 0.00 | -0.08 |
| Football | 44:1 | 24.4 | 0.37 | 0.32 | 0.27 | 0.07 | 0.01 | -0.05 |

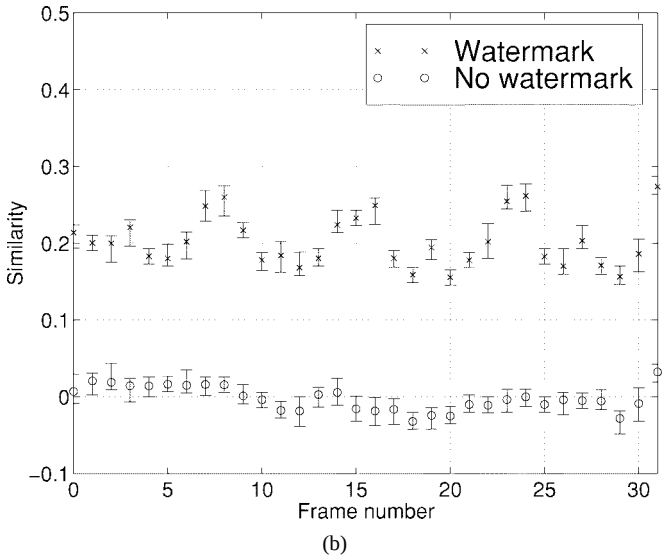
two videos used the same quantization tables. However, the football sequence has more motion present than the Pingpong sequence. As a result, it requires additional bits/pixel to encode the video.

To simulate additional attacks on the watermark, we added colored noise to each video *prior* to MPEG coding. Each video was tested 100 times, with a different colored-noise sequence used during each run. In the first test, we use all of the frames in the video for detection (Detection I). The maximum, mean, and minimum similarity values for each video sequence with and without the watermark are shown in Table IV. Again, observe that the "Min" similarity values with watermark are much larger than the "Max" similarity values without watermark. Even at very low coding quality, the similarity values are widely separated, allowing the existence of a watermark to be easily ascertained.

We also performed detection on single frames from the video (Detection II) without index knowledge. The plots are shown in Fig. 10(a) and (b). The error bars indicate no overlap between the two similarity curves. Even at very low-bit rate, the presence of a watermark is easily observed.




(a)




(b)

Fig. 10. Similarity values versus frame number after MPEG coding (a) Pingpong and (b) Football. The error bars around each similarity value indicate the maximum and minimum similarity values over the 100 runs.

### C. Multiple Watermarks

We also tested the ability to detect watermarks in the presence of other watermarks. This distortion seems likely to occur, as watermarks may be embedded sequentially to track legitimate multimedia distribution. Furthermore, a pirate may use additional watermarks to attack a valid watermark. We embedded three *consecutive* watermarks into each test video, i.e., one after another. All three use the original (nonwatermarked) video as their original during detection. We then added colored noise to the videos and MPEG coded the result. The Pingpong sequence was coded at 0.28 bpp (CR 29:1, PSNR 27.45 dB). Using the same MPEG parameters, the Football sequence was coded at 0.51 bpp (CR 16:1, PSNR 25.43 dB). The test was performed 100 times by generating a new colored-noise sequence each time. The curves for detecting the three watermarks without index knowledge are shown in Fig. 11(a) and (b). The presence of the three watermarks is easily determined.

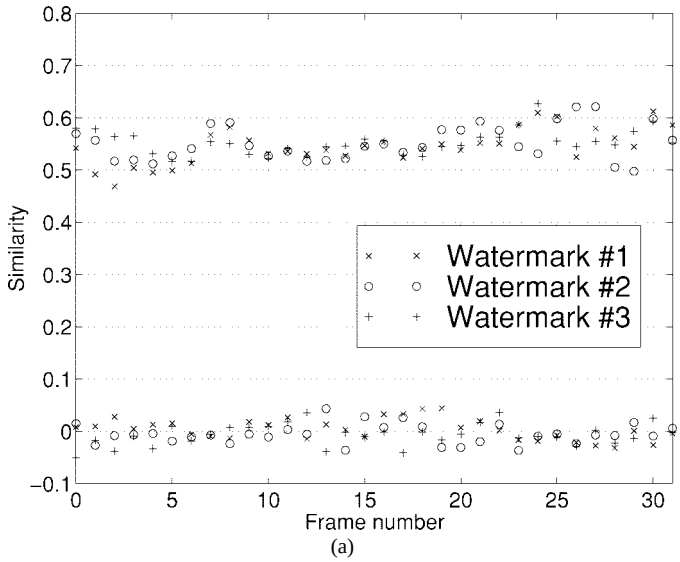



Fig. 11. Similarity values for 3 watermarks after MPEG coding (a) Pingpong and (b) Football.

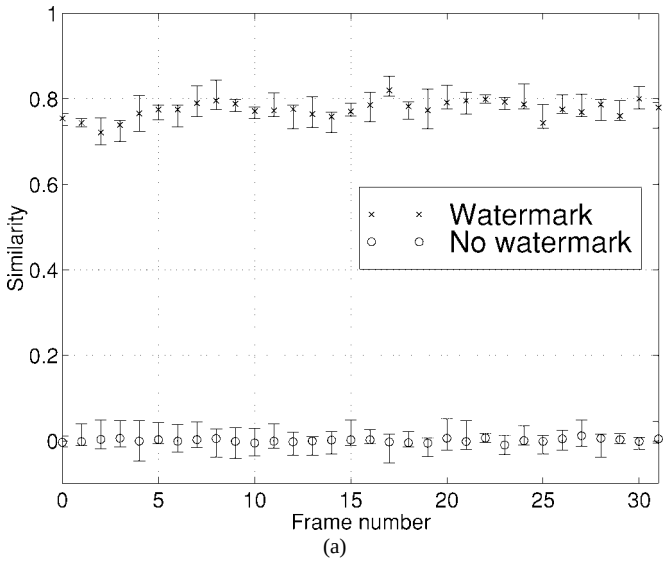



Fig. 12. Similarity values after frame dropping and averaging (a) Pingpong and (b) Football. The error bars around each similarity value indicate the maximum and minimum similarity values over the 100 runs.

### *D. Frame Averaging*

Some of the distortions of particular interest in video watermarking are those associated with temporal processing, e.g., temporal cropping, frame dropping, and frame interpolation are all examples. As we have shown, temporal cropping is handled with our Detection II approach that does not require information regarding frame indexes. To test frame dropping and interpolation, we dropped the odd index frames, i.e., $1, 3, \cdots$, from the test sequences. The missing frames were replaced with the average of the two neighboring frames, $F_{2n+1} = (F_{2n} + F_{2n+2})/2$. Again, we applied Detection II. The resulting detection curves are shown in Fig. 12(a) and (b). The curves with and without watermark are widely separated.

### *E. Printing and Scanning*

An important copyright issue is that of protecting individual video frames from being duplicated in print, e.g., magazines, technical documents, etc. For this test, we created a hardcopy of the original and watermarked frames shown in Figs. 5 and 6 and used a flatbed scanner to re-digitize them. The similarity results obtained from printing and scanning are shown in Table V. Detection was performed *without* knowledge of frame location (i.e., Detection II). The similarity values indicate easy discrimination between watermarked and nonwatermarked printed frames even without knowledge of frame position.

TABLE V
SIMILARITY RESULTS AFTER PRINTING AND SCANNING

| Video Frame | Similarity | |
|---|---|---|
| | With watermark | Without watermark |
| Pingpong | 0.734 | 0.011 |
| Football | 0.611 | 0.052 |

## IX. CONCLUSION

We presented a watermarking procedure to embed copyright protection into digital video by directly modifying the video

samples. The watermarking technique directly exploits the masking phenomena of the HVS to guarantee that the embedded watermark is imperceptible. The owner of the digital video piece is represented by a pseudorandom sequence defined in terms of two secret keys. One key is the owner's personal identification. The other key is calculated directly from the original video signal. The signal dependent watermarking procedure shapes the noise-like author representation according to the masking effects of the host signal. The embedded watermark is perceptually and statistically undetectable. Furthermore, the wavelet-based watermark exists at multiple scales in the video. We illustrated the robustness of the watermarking procedure to several video degradations, including colored noise, MPEG coding, multiple watermarks, frame dropping, and printing and scanning. The watermark was readily detected with and without index knowledge in all of these distortions.

## REFERENCES


[1] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc. 1994 IEEE Int. Conf. Image Processing*, vol. II, 1994, pp. 86–90.

[2] W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding." MIT Media Lab, Cambridge, MA, Tech. Rep., 1994.

[3] R. Wolfgang and E. Delp, "A watermark for digital images," in *Proc. 1996 Int. Conf. Image Proc.*, vol. III, pp. 219–222.

[4] I. Pitas and T. Kaskalis, "Applying signatures on digital images," in *Proc. 1995 IEEE Nonlinear Signal Processing Workshop*, 1995, pp. 460–463.

[5] K. Tanaka, Y. Nakamura, and K. Matsui, "Embedding secret information into a dithered multilevel image," in *Proc. 1990 IEEE Military Commun. Conf.*, 1990, pp. 216–220.

[6] E. Koch and J. Zhao, "Toward robust and hidden image copyright labeling," in *Proc. 1995 IEEE Nonlinear Signal Processing Workshop*, 1995, pp. 452–455.

[7] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," NEC Res. Inst., Tech. Rep. 95–10, 1995.

[8] F. Boland, J. O. Ruanaidh, and C. Dautzenberg, "Watermarking digital images for copyright protection," in *Proc. IEE Int. Conf. Image Proc. Applicat.*, 1995, pp. 321–326.

[9] J. F. Tilki and A. A. Beex, "Encoding a hidden digital signature onto an audio signal using psychoacoustic masking," in *Proc. 1996 7th Int. Conf. Signal Proc. Applicat. Technol.*, 1996, pp. 476–480.

[10] F. Hartung and B. Girod, "Digital watermarking of raw and compressed video," in *Proc. SPIE Dig. Comp. Tech. and Systems for Video Commun.*, vol. 2952, Oct. 1996, pp. 205–213.

[11] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Transparent robust image watermarking," in *Proc. 1996 Int. Conf. Image Processing*, vol. III, pp. 211–214.

[12] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *Proc. 1996 IEEE Int. Conf. Multimedia Comp. and Systems*, 1996, pp. 473–480.

[13] R. Rivest, "Cryptography," in *Handbook of Theoretical Computer Science*, J. van Leeuwen, Ed., vol. 1. Cambridge, MA: MIT Press, 1990, pp. 717–755.

[14] S. Craver, N. Memon, B. L. Yeo, and M. Yeung, "Can invisible watermarks resolve rightful ownerships?" IBM Research, Tech. Rep. RC 20509, *IBM CyberJournal*, July 1996.

[15] S. Craver, N. Memon, B. L. Yeo, and M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications," IBM Res. Tech. Rep. RC 20755, *IBM CyberJournal*, Mar. 1997.

[16] S. Goldwasser and M. Bellare, "Lecture notes on cryptography," Preprint, July 1996.

[17] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.

[18] B. Zhu, A. Tewfik, and O. Gerek, "Low bit rate near-transparent image coding," in *Proc. SPIE Int. Conf. Wavelet Apps. for Dual Use*, vol. 2491, pp. 173–184, 1995.

[19] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals," in *Proc. SPIE Human Vision, Visual Processing, and Digital Display*, vol. 1077, 1989, pp. 178–187.

[20] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, 1980.

[21] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, vol. 8, pp. 14–38, Oct. 1991.

[22] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[23] J. Nam and A. H. Tewfik, "Combined audio and visual streams analysis for video sequence segmentation," in *Proc. 1997 Int. Conf. Acoustics, Speech and Signal Processing*, pp. 2665–2668.

[24] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. 1. New York: Wiley, 1968.

[25] D. J. Le Gall, "MPEG: A video compression standard for multimedia applications," *Commun. ACM*, vol. 34, pp. 47–58, Apr. 1991.


**Mitchell D. Swanson** (M'93) was born in Minneapolis, MN, in 1969. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, in 1992, 1995, and 1997, respectively.

He has previously worked at Honeywell, Inc., Coon Rapids, MN, and Medtronic, Inc., Fridley, MN, and is now with the Department of Electrical and Computer Engineering, University of Minnesota. His research interests include multiscale signal processing, image and video coding for interactive retrieval, data hiding, and digital watermarking.

**Bin Zhu** (M'97) received the B.S. degree in physics from University of Science and Technology of China in 1986. He received the the M.S. and Ph.D. degrees in electrical engineering from University of Minnesota, Minneapolis, in 1993 and 1997, respectively.

His research interests include multimedia compression and processing, wavelet transform, and multiscale signal processing.

**Ahmed H. Tewfik** (S'81–M'82–SM'92–F'96) was born in Cairo, Egypt, on October 21, 1960. He received the B.Sc. degree from Cairo University, Cairo, Egypt, in 1982 and the M.Sc., E.E. and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1984, 1985, and 1987, respectively.

In 1987 he worked at Alphatech, Inc., Burlington, MA, and is currently the E. F. Johnson Professor of Electronic Communications with the Department of Electrical Engineering at the University of Minnesota, Minneapolis. He has served as a consultant to MTS Systems, Inc., Eden Prairie, MN, and is a regular consultant to Rosemount, Inc., Eden Prairie, MN. His current research interests are in signal processing for multimedia (in particular, watermarking, data hiding, and content-based retrieval), low power multimedia communications, adaptive search, and data acquisition strategies for World Wide Web applications, radar and dental/medical imaging, monitoring of machinery using acoustic emissions, and industrial measurements.

Dr. Tewfik is a Distinguished Lecturer of the IEEE Signal Processing Society for the period July 1997 to July 1999. He was a Principal Lecturer at the 1995 IEEE EMBS Summer School. He was awarded the E. F. Johnson Professorship of Electronic Communications in 1993. He received a Taylor Faculty Development Award from the Taylor Foundation in 1992 and an NSF Research Initiation Award in 1990. He gave a plenary lecture at the 1994 IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP'94) and an invited tutorial on wavelets at the 1994 IEEE Workshop on Time-Frequency and Time-Scale Analysis. He was selected to be the first Editor-in-Chief of the IEEE SIGNAL PROCESSING LETTERS in 1993. He is a past Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and was a Guest Editor of two special issues on wavelets and their applications in the same TRANSACTIONS.