A CONTENT-DEPENDENT SPATIALLY LOCALIZED VIDEO WATERMARK FOR RESISTANCE TO COLLUSION AND INTERPOLATION ATTACKS

Karen Su, Deepa Kundur, and Dimitrios Hatzinakos

University of Toronto
Edward S. Rogers Sr. Department of Electrical and Computer Engineering
10 King's College Road, Toronto, Ontario, Canada M5S 3G4
{karen,deepa,dimitris}@comm.utoronto.ca

ABSTRACT

This paper presents a novel video watermarking algorithm based on two key ideas: *statistical invisibility* and *content-synchronized placement*. We argue that statistical invisibility is essential to protect video watermarks from statistical collusion, and present a natural way to induce this property using a *content-dependent spatially localized* watermarking framework. We introduce the notion of a *watermark footprint*, the spatial locations over which its energy is spread. By defining localized footprints with regular structures, e.g., a set of subframes within each frame, current image watermarking techniques can immediately be applied at the subframe-level. We address the issue of reduced spatial redundancy by proposing an attack model based on bilinear interpolation, and embedding the watermark into regions with lower expected distortions. Results are presented to demonstrate the effectiveness of the algorithm.

1. INTRODUCTION

A digital watermark is a data message embedded into a digital signal such as an image, audio, or video stream. The main requirements are that the watermarked media must be perceptually equivalent to the original, and the watermark should be robust to a variety of spatial distortions. In addition to these basic criteria, video watermarks must support blind detection (i.e., detection without access to the original), be robust to temporal distortions such as frame averaging and swapping, and also resist multiple frame statistical analysis/estimation attacks (i.e., collusion).

Many existing video watermarking techniques are based on the idea of spreading the watermark energy globally over all of the pixels in each of the frames [1], [2], [3], [4]. Such schemes are essentially *content-independent* since the watermark structure does not vary according to the visual content of the host signal. They enjoy lower computational complexities at the expense of local control over the placement and embedding strength of the watermark.

Some approaches have also been proposed where the watermark energy is *spatially localized*, i.e., its footprint does not cover the entire pixel space of each frame [2], [5], [6]. In these algorithms, the video frames are partitioned into 8×8 blocks, and a subset of these is selected for watermarking, based on properties that make them more suitable candidates for data hiding. However, because the properties considered in the first two cases are MPEG-2 coding structures rather than visual components, they are still content-independent. In the third scheme, the watermarked blocks are chosen based on their visual characteristics, but they are constrained to lie in a regular tiling pattern.

Based on the following reasoning, we believe that a mark that is both content-dependent and spatially localized has distinct advantages for video watermarking purposes:

1. For reduced complexity compared to 3D transform techniques [7], and detection from any single isolated video frame, a *frame-by-frame* strategy is desirable.
2. In previous work [7], it was proposed to protect watermarks from statistical collusion by marking visually similar frames using similar patterns, and vice versa. To extend this idea, we write the marked video in the form $X = U + \alpha W$ and define *statistical invisibility* as $\rho(X_i, X_j) = \rho(U_i, U_j)$ for all frame indices i, j, where U is the host, W the watermark, and $\rho(A, B)$ the correlation coefficient between A and B.
3. It can then be shown that in order to achieve statistical invisibility, $\rho(U_i, U_j) = A \cdot \rho(W_i, W_j)$, i.e., *the statistical correlation of the watermark must be designed to match that of the host video frames*.
4. To resist collusion, we propose a content-dependent spatially localized watermarking framework. Each full-frame watermark is comprised of a number of smaller basic patterns or *subframes*, whose placement depends on the visual contents of the host frame.

Thanks to the Natural Sciences and Engineering Research Council (NSERC), Nortel Institute for Telecommunications, and Communications and Information Technology Ontario (CITO) for funding this work.

By applying content-dependent selection criteria, we can adjust $\rho(W_i, W_j)$ to meet the statistical invisibility condition. After determining the watermark placement, two basic patterns are considered, one based on 2D direct sequence Spread Spectrum (SS) concepts, similar to that used in JAWS [4], and another based on embedding peaks in the DFT domain [8]. No matter which pattern is used, the overall watermark possesses a unique feature: Using low complexity frame-by-frame processing, the illusion of a 3D structure is attained through content-dependence and the 3D nature of the video itself. Note that in contrast to other watermarks that vary from frame to frame, e.g., CDMA [3], frame swapping attacks are ineffective against the proposed algorithm. Since the locations of the subframes are dependent on visual properties instead of on structural properties of the video, temporal and spatial synchronization are not necessary.

2. STRATEGIC WATERMARK PLACEMENT

As illustrated in Figure 1, basic geometric attacks such as rotation and fractional pixel translation can be represented as the combination of two operations: a change of sampling grid and a re-sampling by interpolation. In this work, we consider how much the intensity of each pixel can be distorted due to such attacks; we also determine which characteristics of the image affect the magnitudes of these distortions, and devise a content-dependent strategy for placing the watermark in low distortion regions.

Given an image $U(i, j)$, the nearest neighbourhood of pixel (i, j) is

$$N_{(i,j)} = \{U(u, v)\colon |u - i| \leq 1, |v - j| \leq 1, (u, v) \neq (i, j)\}$$

Suppose that the sampling grid is modified by an attack, such that the intensity $U(i, j)$ is replaced by an interpolated value $\hat{U}(i, j)$. Assuming bilinear interpolation within the nearest neighbourhood of (i, j), we see that $|\hat{U}(i, j) - U(i, j)|$ will be less than the greatest magnitude difference between $U(i, j)$ and any of its neighbours. We define the *peak nearest neighbourhood interpolation noise* as

$$M(i, j) = \mathsf{max}_{I \in N(i,j)}(|I - U(i, j)|),$$

And the distortion noise at (i, j) is bounded by

$$|\hat{U}(i, j) - U(i, j)| \leq M(i, j).$$

Using the watermark attack model presented above, it is clear that $M(i, j)$ will be largest when the spatial gradients at (i, j) are large. This observation is supported by the HVS spatial edge masking property. Since $M(i, j)$ is a bound on the distortion that each visual component of the picture can suffer, we see that although regions with larger bounds can absorb more watermark energy, they are also subject to stronger attacks. In regions where $M(i, j)$ is small, the amount of distortion available to attackers is reduced. This hypothesis is supported by results reported in [6], where a watermark localized in 8×8 MPEG-2 blocks is found to enjoy the best balance between imperceptibility and robustness when embedded into blocks with a noise contrast (e.g., regions with relatively small spatial gradients). Therefore, in the proposed approach, watermark transmission via the picture components that will be the least distorted by interpolation noise is favoured.

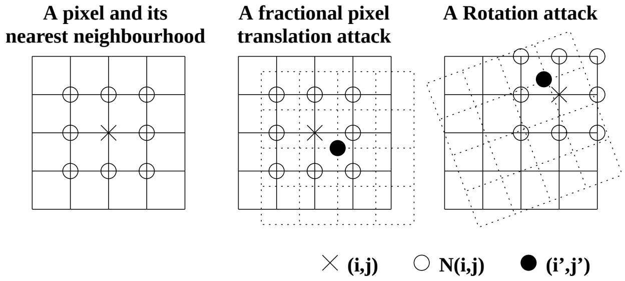
A pixel and its nearest neighbourhood
A fractional pixel translation attack
A Rotation attack
× (i,j)
○ N(i,j)
● (i',j')

Fig. 1. An illustration of fractional pixel translation and rotation as nearest neighbourhood interpolation attacks.

3. EMBEDDING ALGORITHM

The embedding algorithm comprises five main steps as illustrated in Figure 2:

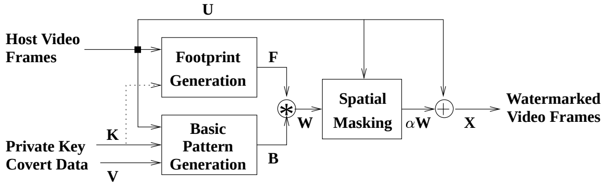

Fig. 2. Block diagram of proposed watermark embeddor.

1. *Footprint generation*: Compute the picture-dependent distortion bound, $M(i, j)$, and its average over all square windows of side width d, $\overline{M}_d(i, j) = M(i, j) \star Rect(\frac{i}{d}, \frac{j}{d})$. The minima of $\overline{M}_d$ correspond to the centers of $d \times d$ subframes with low average interpolation distortions; a non-overlapping union of these forms the watermark footprint (see Figure 3).
2. *Basic watermark pattern generation*: The watermark pattern is a key-dependent noise-like $d \times d$ pattern. As discussed in Section 1, we will consider spatial domain SS watermarking for low complexity and peak embedding in the DFT domain for higher robustness. Both watermarks have interleaved reference components to aid in detection [9].
3. *Full-frame watermark construction*: Convolve the footprint and the basic pattern to form the watermark.
4. *Spatial masking*: Apply local image-dependent scaling factors derived from the Noise Visibility Function

(NVF) proposed by Voloshynovskiy *et al.* [10] to optimize robustness while maintaining imperceptibility.

5. *Data embedding*: The watermark is added to the host in the spatial domain. For the SS case, this strategy is similar to that proposed in JAWS, however in this case the subframes are irregularly tiled, offering better resilience to statistical estimation.

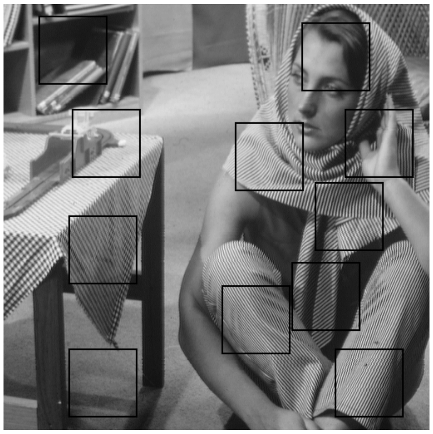

Fig. 3. A sample watermark footprint for barbara, d = 81.

4. DETECTION AND EXTRACTION ALGORITHMS

The SS watermark is detected and extracted as follows:

1. *Footprint generation*: Determine the watermark footprint (as in Section 3).
2. *High-pass filtering*: To reduce the power of the image component, a 3×3 Laplacian filter is applied before any subsequent processing.
3. *Subframe-level detection*: Compute the projection D_i of each subframe F_i onto the reference component of the 2D SS watermark pattern W_{ref}. A threshold T is defined, depending on the desired false positive probability; if $D_i > T$, we say that a watermark is detected and store F_i for use in Step 4, otherwise we reject it as corrupted, misaligned, or unmarked.
4. *Maximal ratio combining*: The remaining subframes F_i are combined, weighted by their SNRs, i.e., reference watermark to image power ratios, thus exploiting the watermark's diversity.

$$F = \sum_i \frac{D_i^2 \sum_{m,n} W_{ref}(m,n)^2}{\sum_{m,n}[F_i(m,n) - D_i W_{ref}(m,n)]^2} F_i$$

5. *Data extraction*: The encoded data message is extracted from F by de-interleaving, and decoded to recover the message data bits.

5. EXPERIMENTAL RESULTS

For these tests, a subframe side width $d = 81$ is used. This value gives a good performance tradeoff between robustness and data rate. Figure 4 illustrates how the proposed algorithm performs in comparison to JAWS as an image (barbara) is translated diagonally by a fractional number of pixels; the translation attack is implemented using bilinear interpolation. Our implementation of JAWS uses tile sizes of $M = 128$, and a detection threshold of $T = \frac{15}{M}$. The *no detection threshold* version of JAWS is shown only as an illustration; since it considers only the maximum and minimum correlation coefficients, it has an impractical false positive rate of 1. 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 algorithm also encounters more bit errors, however the decrease is much more gradual. We believe that this enhancement can be attributed to the fact that the energy of the spatially localized watermark is concentrated in low interpolation distortion regions.

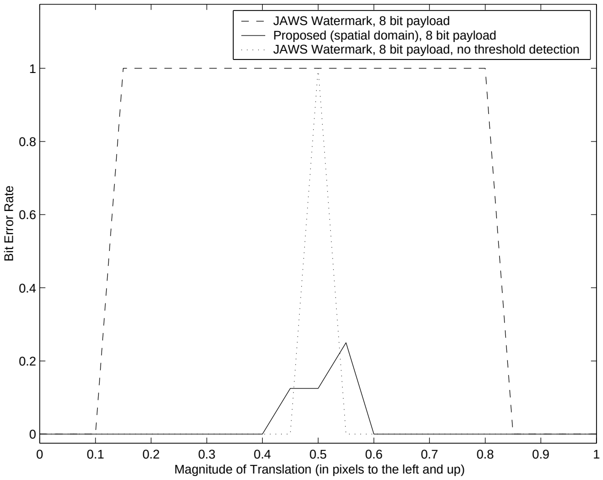

Fig. 4. Bit error vs. diagonal fractional pixel translation attacks (implemented using bilinear interpolation) for the proposed spatial domain algorithm and JAWS. The PSNR was fixed at 38dB.

Figures 5 and 6 show the original and watermarked images respectively. Table 1 summarizes the performance of the proposed algorithm (DFT domain) against StirMark 3.1 [11]. The algorithm performs well for a large range of frame-as-image attacks, while exhibiting collusion resistance properties that are crucial for video watermarks.

6. CONCLUDING REMARKS

We propose a video watermark in a novel content-dependent spatially localized framework, where the watermark energy is concentrated in subframes with desirable properties, and

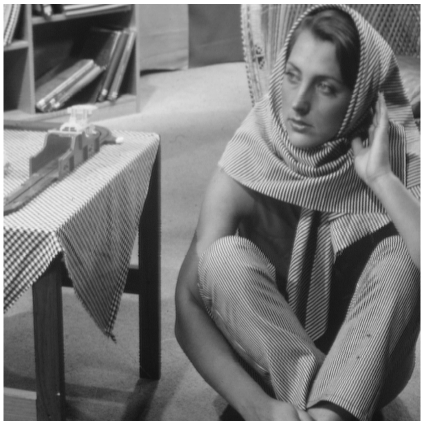

Fig. 5. Original barbara.

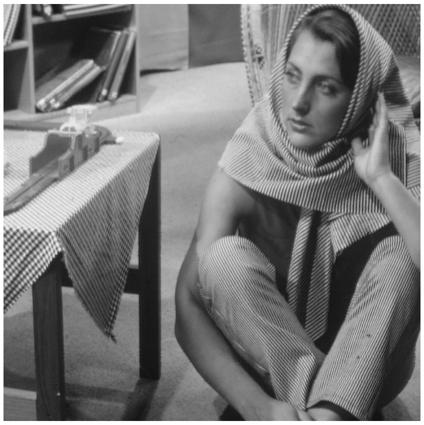

Fig. 6. Watermarked (spatial domain mark) barbara.

the subframe locations are *synchronized using visual content rather than structural markers*. We consider the property of low average interpolation noise, and demonstrate that watermark footprints selected using this criteria have a high robustness to geometric distortions. The method is distinguished by its ability to be embedded and extracted using frame-based algorithms, while resisting collusion. Future directions include improving performance against scaling and aspect ratio changes, enhancing the subframe selection algorithm, e.g., incorporating key-dependence to improve secrecy, and applying state-of-the-art image watermarking ideas at the subframe level, e.g., turbo codes.

Test set	**Averaged score**
Signal enhancement	100
Compression (JPEG QF≥40)	99
Cropping (up to 50%)	100
Rotation (small-angle, no scale)	100
Flip	100

Table 1. Summary of performance of proposed DFT domain watermark against StirMark 3.1.

7. REFERENCES

[1] Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon, “Secure spread spectrum watermarking for multimedia,” *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1673–1687, December 1997.

[2] Frank Hartung and Bernd Girod, “Watermarking of uncompressed and compressed video,” *Signal Processing*, vol. 66, no. 3, pp. 283–301, May 1998.

[3] Bijan G. Mobasseri, “Exploring CDMA for watermarking of digital video,” in *Proceedings of the SPIE*, January 1999, vol. 3657, pp. 96–102.

[4] Ton Kalker, Geert Depovere, Jaap Haitsma, and Maurice Maes, “A video watermarking system for broadcast monitoring,” in *Proceedings of the SPIE*, January 1999, vol. 3657, pp. 103–112.

[5] Gerrit C. Langelaar, Reginald L. Lagendijk, and Jan Biemond, “Real-time labeling of MPEG-2 compressed video,” *J. of Visual Comm. and Image Representation*, vol. 9, no. 4, pp. 256–270, December 1998.

[6] V. Darmstaedter, J.-F. Delaigle, D. Nicholson, and B. Macq, “A block based watermarking technique for MPEG2 signals: Optimization and validation on real digital TV distribution links,” in *Proc. Euro. Conf. on Multimedia Applications, Services and Techniques*, 1998, pp. 190–206.

[7] Mitchell D. Swanson, Bin Zhu, and Ahmed T. Tewfik, “Multiresolution scene-based video watermarking using perceptual models,” *IEEE J. on Sel. Areas in Comm.*, vol. 16, no. 4, pp. 540–550, May 1998.

[8] Shelby Pereira and Thierry Pun, “Robust template matching for affine resistant image watermarks,” *IEEE Trans. on Image Processing*, vol. 9, no. 6, pp. 1123–1129, June 2000.

[9] D. Kundur and D. Hatzinakos, “Attack characterization for effective watermarking,” in *Proc. Int’l Conf. on Image Processing*, 1999, vol. 4, pp. 240–244.

[10] S. Voloshynovskiy, A. Herrigel, N. B., and T. Pun, “A stochastic approach to content adaptive digital image watermarking,” *Lecture Notes in Computer Science*, vol. 1768, pp. 212–236, September 2000.

[11] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn, “Attacks on copyright marking systems,” *Lecture Notes in Computer Science*, vol. 1525, pp. 219–239, April 1998.