Image Authentication and Tampering Localization using Distributed Source Coding

Yao-Chung Lin, David Varodayan, and Bernd Girod
Information Systems Laboratory, Stanford University, Stanford, CA 94305
{yclin79, varodayan, bgirod}@stanford.edu

***Abstract*— Media authentication is important in content delivery via untrusted intermediaries, such as peer-to-peer (P2P) file sharing. Many differently encoded versions of a media file might exist. Our previous work applied distributed source coding to distinguish the legitimate diversity of encoded images from tampering. An authentication decoder was supplied with a Slepian-Wolf encoded lossy version of the image as authentication data. Distributed source coding provided the desired robustness against legitimate encoding variations, while detecting illegitimate modification.**

We augment the decoder to localize tampering in an image already deemed to be inauthentic. The localization decoder requires only incremental localization data beyond the authentication data since we use rate-adaptive distributed source codes. Both decoders perform joint bitplane decoding, rather than conditional bitplane decoding. Our results demonstrate that tampered image blocks can be identified with high probability using authentication plus localization data of only a few hundred bytes for a 512x512 image.

I. Introduction

Media authentication is important in content delivery via untrusted intermediaries, such as peer-to-peer (P2P) file sharing or P2P multicast streaming. In these applications, many differently encoded versions of the original file might exist. Moreover, transcoding and bitstream truncation at intermediate nodes might give rise to further diversity. But intermediaries might also tamper with the media for many reasons, such as interfering with the distribution of a particular file, piggybacking unauthentic content, or generally discrediting a distribution system. In previous work [1], we applied distributed source coding to image authentication to distinguish the diversity of legitimate encodings from malicious manipulation.

Past approaches fall into two groups: watermarks and media hashes. A "fragile" watermark can be embedded into the host signal waveform without perceptual distortion [2] [3]. Users can confirm the authenticity by extracting the watermark from the received content. The system design should ensure that the watermark survives lossy compression, but that it "breaks" as a result of a malicious manipulation. Unfortunately, watermarking authentication is not backward compatible with previously encoded contents; unmarked contents cannot be authenticated later. Embedded watermarks might also increase the bit-rate required when compressing a media file.

This work has been supported, in part, by a gift from NXP Semiconductors to the Stanford Center for Integrated Systems and, in part, by the Max Planck Center for Visual Computing and Communication.

Media hashing [4] [5] achieves authentication of previously encoded media (as well as localization of tampering) by using an authentication server to supply authentication data to the user. Media hashes are inspired by cryptographic digital signatures [6], but unlike cryptographic hash functions, media hash functions offer proof of perceptual integrity. Using a cryptographic hash, a single bit difference leads to an entirely different hash value. If two media signals are perceptually indistinguishable, they should have identical hash values. A common approach of media hashing is extracting the features which have perceptual importance and should survive compression. The authentication data are generated by compressing the features or generating their hash values. The user checks the authenticity of the received content by comparing the features or their hash values to the authentication data.

We review our image authentication system based on distributed source coding [1] in Section II. Compared to conventional media hashing, our scheme also exploits knowledge of the variation among legitimate images. In Section III, we augment the authentication decoder into a localization decoder that localizes tampering in images already deemed to be inauthentic by the authentication decoder. Simulation results in Section IV show that tampered pixels are identified with high probability.

II. Background

Fig. 1 is the block diagram for both image authentication of [1] and tampering localization of this paper. We first review the authentication system. We denote the source image as x. The user receives the image-to-be-authenticated y as the output of a two-state lossy channel that models legitimate and illegitimate modifications. The left-hand side of Fig. 1 shows that the authentication data consist of a Slepian-Wolf encoded lossy version of x and a digital signature of that version. The authentication decoder, in the right-hand side of Fig. 1, knows the statistics of the worst permissible legitimate channel and can correctly decode the authentication data only with the help of an authentic image y as side information.

We model the image-to-be-authenticated y by way of the space-varying two-state lossy channel in Fig. 2. The legitimate state of the channel performs lossy JPEG2000 or JPEG compression and reconstruction with peak signal-to-noise ratio (PSNR) of 30dB or better. The illegitimate state additionally includes malicious tampering.

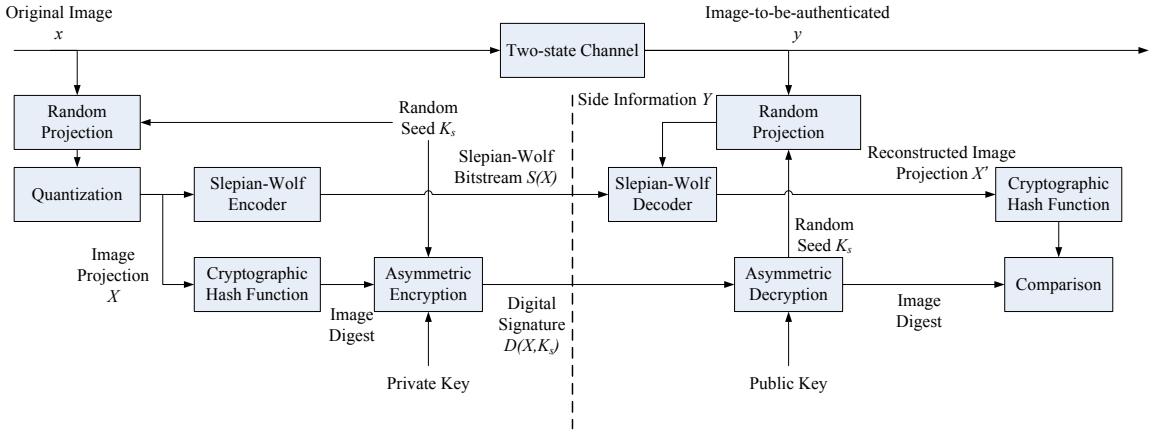

Fig. 1. Image authentication and tampering localization systems based on distributed source coding

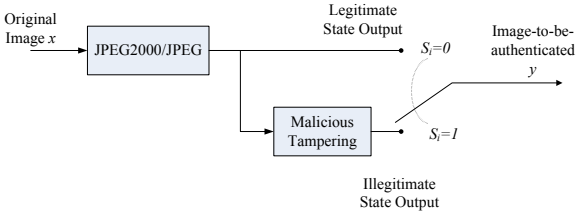

Fig. 2. Space-varying two-state lossy channel

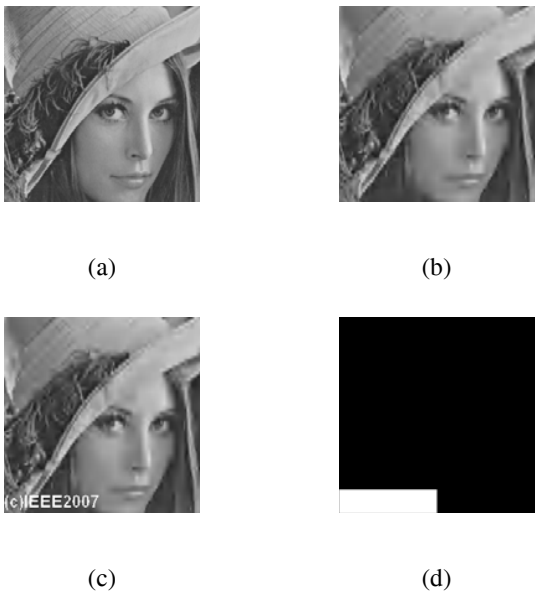

Fig. 3. Fragment of "Lena" image (a) x original, (b) y if $\sum_i S_i = 0$, (c) y with $\sum_i S_i > 0$, (d) channel states S_i associated with the 16x16 blocks of output (c).

Fig. 3 demonstrates this channel. The source image x is "Lena" at 8-bit 512x512 resolution. In the legitimate state, the channel output is JPEG2000 compression and reconstruction at (the worst permissible) 30dB PSNR. In the illegitimate state, a text banner is overlaid on the reconstructed image. The channel state variable S_i is defined per nonoverlapping 16x16 blocks of image y. If any pixel in block B_i is part of the banner text, $S_i = 1$; otherwise, $S_i = 0$.

In our authentication system shown in Fig. 1, a pseudo-random projection (based on a randomly drawn seed K_s) is applied to the original image x and the projection coefficients are quantized to yield X. The authentication data comprise two parts, both derived from X. The Slepian-Wolf bitstream $S(X)$ is the output of a Slepian-Wolf encoder based on rate-adaptive low-density parity-check (LDPC) codes [7]. The much smaller digital signature $D(X, K_s)$ consists of the seed K_s and a cryptographic hash value of X signed with a private key.

In our system, the authentication data are generated by a server upon request. Each response uses a different random seed K_s, which is provided to the decoder as part of the authentication data. This prevents an attack which simply confines the tampering to the nullspace of the projection. Based on the random seed, for each 16x16 nonoverlapping block of pixels B_i, we generate a 16x16 pseudorandom matrix P_i by drawing its elements independently from a Gaussian distribution $\mathcal{N}(1, \sigma_z^2)$ and normalizing so that $||P_i||_2 = 1$. We choose $\sigma_z = 0.2$ empirically. The inner product $\langle B_i, P_i \rangle$ is quantized into an element of X.

The rate of the Slepian-Wolf bitstream $S(X)$ determines how statistically similar the image-to-be-authenticated must be to the original to be declared authentic. If the conditional entropy $H(X|Y)$ exceeds the bit-rate R in bits per pixels, X can no longer be decoded correctly [8]. Therefore, the rate of $S(X)$ should be chosen to distinguish between the different joint statistics induced in the images by the legitimate and illegitimate channel states. At the encoder, we select a Slepian-Wolf bit-rate just sufficient to authenticate both legitimate 30dB JPEG2000 and JPEG reconstructed versions of x.

At the receiver, the user seeks to authenticate the image y with authentication data $S(X)$ and $D(X, K_s)$. It first projects y to Y in the same way as during authentication data

generation. A Slepian-Wolf decoder reconstructs X' from the Slepian-Wolf bitstream $S(X)$ using Y as side information. Decoding is via LDPC belief propagation [9] initialized according to the statistics of the legitimate channel state at the worst permissible quality for the given original image. Finally, the image digest of X' is computed and compared to the image digest, decrypted from the digital signature $D(X, K_s)$ using a public key. If these two image digests are identical, the receiver recognizes image y as authentic.

With this system, we demonstrated false positive rates close to zero for authentication data size less than 40 bytes [1].

III. Tampering Localization

A. Problem

The authentication problem discussed above is a decision on the sum of channel states over all blocks in an image; whether $\sum_i S_i = 0$ or $\sum_i S_i > 0$. In the case that the image is inauthentic ($\sum_i S_i > 0$), the tampering localization problem can be formulated as deciding on S_i for each block, given the Slepian-Wolf bitstream $S(X)$ and the digital signature $D(X)$.

B. Authentication and Localization Data Generation

The localization decoder requires more information than the authentication decoder. Fortunately, since we use rate-adaptive LDPC codes [7] for Slepian-Wolf coding, the localization decoder re-uses the authentication data. Incremental localization data is sent through the Slepian-Wolf bitstream $S(X)$.

In our previous paper [1], a separate Slepian-Wolf bitstream was used for each bitplane of X. At the authentication decoder, the bitplanes were decoded conditionally with previously decoded ones used as additional side information [10]. But the localization decoder requires all bitplanes together to estimate the channel state S_i per block. Hence we adopt joint bitplane coding [11], wherein a single Slepian-Wolf bitstream is used for all transmitted bitplanes. In order to enable rate-adaptivity for the overall authentication/localization system, we use joint bitplane coding for the authentication system as well.

C. Localization Decoder

The localization decoder applies the sum-product algorithm [12] on the factor graph in Fig. 4 to estimate each channel state likelihood $P(S_i = 1)$. Decoding is initialized with the syndrome nodes values $S(X)$ and the side information Y.

In terms of the factor graph, the joint probability of the bits of the image projection X and the channel states given the syndrome values and the side information can be factored as follows. The factor at each syndrome node is an indicator function of the satisfaction of that syndrome constraint. The factor connected to each state node $f_s^i(S_i) = P(S_i)$. The factor $f_b^i(X_i, S_i) = P(X_i|Y_i; S_i)$. When $S_i = 0$, $f_b^i(X_i, 0)$ is proportional to the integral of a Gaussian distribution with mean Y_i and a fixed variance σ^2 over the quantization interval of X_i. When $S_i = 1$, $f_b^i(X_i, 1)$ is uniform.

The iterations of belief propagation terminate when the hard decisions on bits of X satisfy the constraint imposed by the syndrome $S(X)$. Finally, each block B_i of y is declared to be tampered if $P(S_i = 1) > \alpha$, a fixed decision threshold.

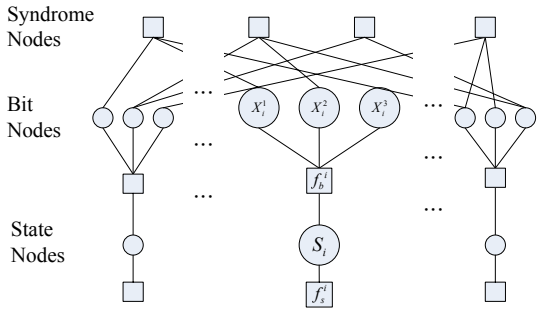

Fig. 4. Factor graph for the localization decoder

IV. Simulation results

In practice, the localization decoder would only run if the authentication decoder deems an image to be inauthentic, so we test the tampering localization system only with maliciously tampered images. We use test images "Barbara", "Lena", "Mandrill", and "Peppers" at 512x512 resolution in 8-bit gray resolution. The space-varying two-state channel in Fig. 2 applies JPEG2000 or JPEG compression and reconstruction at several qualities above 30dB. The malicious tampering consists of the overlaying of up to five text banners of different sizes at random locations in the image. The text banner sizes are 198x29, 29x254, 119x16, 16x131 and 127x121 pixels. The text color is white or black, depending on which is more visible. This avoids generating trivial attacks, such as overlaying white text on a white area.

We first compare the minimum authentication data rate required by the authentication decoder and the minimum authentication plus localization data rate required by the localization decoder. Fig. 5 shows the Slepian-Wolf bitstream components $S(X)$ of these rates (in bits per pixel of the original image x) for "Lena" with X quantized to 3 bitplanes. All five text banners are placed for malicious tampering, because greater tampering makes 'disauthentication' easier and localization more difficult. The placement is random for 100 trials, leading to tampering of 12% to 17% of the nonoverlapping 16x16 blocks of the original image x. To localize tampering in JPEG2000 or JPEG reconstructions above 30dB, Fig. 5 indicates that the required authentication plus localization rate is roughly 2.5 times the required authentication rate. The incremental localization rate (the gap between the rates) discovers not only the location of the tampering but also the magnitude of the tampering.

Next we investigate the worst-case authentication plus localization rate necessary for localizing tampering in JPEG2000 or JPEG reconstructions above 30dB. We randomly place one to five text banners over 2000 trials and run the localization decoder with decision threshold $\alpha = 0.5$. Fig. 6 is a scatter plot of authentication plus localization rates versus percentage of affected blocks, for "Peppers" with the 32x32 image projection X quantized to 3 bitplanes. The scatter plots for the other test images are similar. We select the highest rate observed as the Slepian-Wolf rate for authentication plus localization,

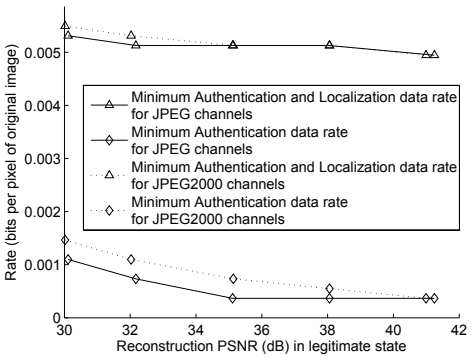

Fig. 5. Minimum authentication and authentication plus localization data rates for decoding Slepian-Wolf bitstream $S(X)$ for "Lena"

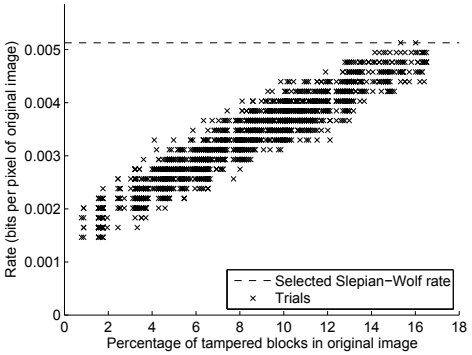

Fig. 6. Scatter plot of authentication plus localization rates versus percentage of affected samples for "Peppers"

which gives a bitstream size of 174 bytes for X quantized to 3 bitplanes (less than 6.3% of the compressed image size). For X quantized to 4 bitplanes, the bitstream size is 232 bytes (less than 8.4% of the compressed image size).

Using these Slepian-Wolf bitstream sizes, we measure various failure rates. The blockwise falsely deemed tampered rate is the proportion of untampered blocks (that is, with $S_i = 0$) that were mistaken for tampered blocks. Conversely, the blockwise falsely deemed untampered rate is the proportion of tampered blocks (that is, with $S_i = 1$) that were mistaken for untampered blocks. We also consider the pixelwise falsely deemed untampered rate. Fig. 7 shows these failure rates for X quantized to 3 and 4 bitplanes as the decision threshold α varies. The curves for X quantized to 4 bitplanes indicate that our choice of authentication plus localization rate can zero the blockwise falsely deemed tampered rate, while keeping the blockwise falsely deemed untampered rate near 20%. Although the latter result seems weak, we note that most of the blocks falsely deemed untampered have only a few pixels tampered. This explains why the corresponding pixelwise falsely deemed untampered rate is an order of magnitude less, roughly 2%. This performance is acceptable to localize banners consisting of hundreds of pixels.

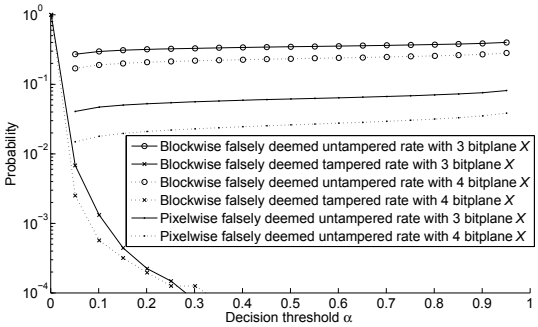

Fig. 7. Localization decoder failure rates versus α for X quantized to different numbers of bitplanes

V. Conclusions

We augment our earlier image authentication system using distributed source coding to perform tampering localization in images already deemed to be inauthentic. Since both systems use rate-adaptive distributed source codes, the localization decoder only requires incremental localization rate beyond the authentication rate. We demonstrate that an authentication plus localization Slepian-Wolf bitstream of 232 bytes (less than 8.4% of the compressed image size) is sufficient to identify tampered pixels with 98% confidence, while correctly classifying untampered blocks. In future work, we will consider other forms of legitimate and illegitimate editing.

References

[1] Y.-C. Lin, D. Varodayan, and B. Girod, "Image authentication based on distributed source coding," in *IEEE Int. Conf. on Image Processing*, San Antonio, TX, Sep. 2007, submitted.

[2] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *IEEE Int. Conf. on Acoust., Speech, Signal Processing*, Salt Lake City, UT, May 2001.

[3] R. B. Wolfgang and E. J. Delp, "A watermark for digital images," in *IEEE Int. Conf. on Image Processing*, Lausanne, Switzerland, Sep. 1996.

[4] C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing JPEG compression from malicious manipulation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 2, pp. 153–168, 2001.

[5] C.-S. Liu and H.-Y. M. Liao, "Structural digital signature for image authentication: an incidental distortion resistant scheme," *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 161–173, 2003.

[6] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory*, vol. 22, no. 6, pp. 644–654, 1976.

[7] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive codes for distributed source coding," *EURASIP Signal Processing J.*, vol. 86, no. 11, pp. 3123–3130, 2006.

[8] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[9] A. Liveris, Z. Xiong, and C. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Commun. Lett.*, vol. 6, no. 10, pp. 440–442, 2002.

[10] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *SPIE Visual Communications and Image Processing Conf.*, San Jose, CA, Jan. 2004.

[11] D. Varodayan, A. Mavlankar, M Flierl, and B. Girod, "Distributed grayscale stero image coding with unsupervised learning of disparity," in *IEEE Data Compression Conf.*, Snowbird, UT, March 2007.

[12] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 10, pp. 498–519, 2001.