SPATIAL MODELS FOR LOCALIZATION OF IMAGE TAMPERING USING DISTRIBUTED SOURCE CODES

Yao-Chung Lin, David Varodayan, and Bernd Girod

Information System Laboratory, Stanford University, Stanford, CA 94305
{yao-chung.lin, 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 not only to distinguish the legitimate diversity of encoded images from tampering but also localize the tampered regions in an image already deemed to be inauthentic. An authentication decoder was supplied with a Slepian-Wolf encoded image projection as authentication data. A localization decoder required only incremental localization data beyond the authentication data since we use rate-adaptive distributed source codes.

We extend the localization decoder with 1D and 2D spatial models to exploit the contiguity of the tampered regions. Our results show that the spatial decoders save 10% to 17% of authentication plus localization data size and offer greater confidence in tampering localization.

Index Terms— image authentication, image tampering localization, distributed source coding

1. 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, we applied distributed source coding to image authentication to distinguish the diversity of legitimate encodings from malicious manipulation [1] and demonstrated that the same framework can localize tampering in images deemed to be inauthentic [2]. In this paper, we improve the performance of tampering localization by exploiting the contiguity of tampered regions using one and two dimensional spatial models.

Past media authentication approaches fall into two groups: watermarks and media hashes. A "fragile" watermark can be embedded into the host signal waveform without perceptual distortion [3][4]. Users can confirm the authenticity by extracting the watermark from the received content. The watermark should survive lossy compression, but should "break" 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.

Media hashing [5][6] 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 [7], 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 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.

Section 2 reviews our image authentication system using distributed source codes [1], an extension of media hashing. It has similarities with secure biometric authentication [8][9] and the semi-fragile watermarking scheme in [10]. We extend our previous tampering localization work [2] by applying spatial models to exploit spatial contiguity of tampered regions in Section 3, just as [11] and [12] used spatial models to reduce distributed compression rate. Simulation results in Section 4 show that this can reduce the localization data size and system failure rates.

2. BACKGROUND

Fig. 1 is the block diagram for the image authentication scheme of [1] and the tampering localization extensions of [2] and the current paper. What differs in each case are the models of the source and channel at the Slepian-Wolf decoder.

We denote the source image as x. 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. 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, $S_i = 0$.

We now review the authentication system. The left-hand side of Fig. 1 shows that a pseudorandom projection (based on a randomly drawn seed K_s) is applied to the original image x and the projection

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.

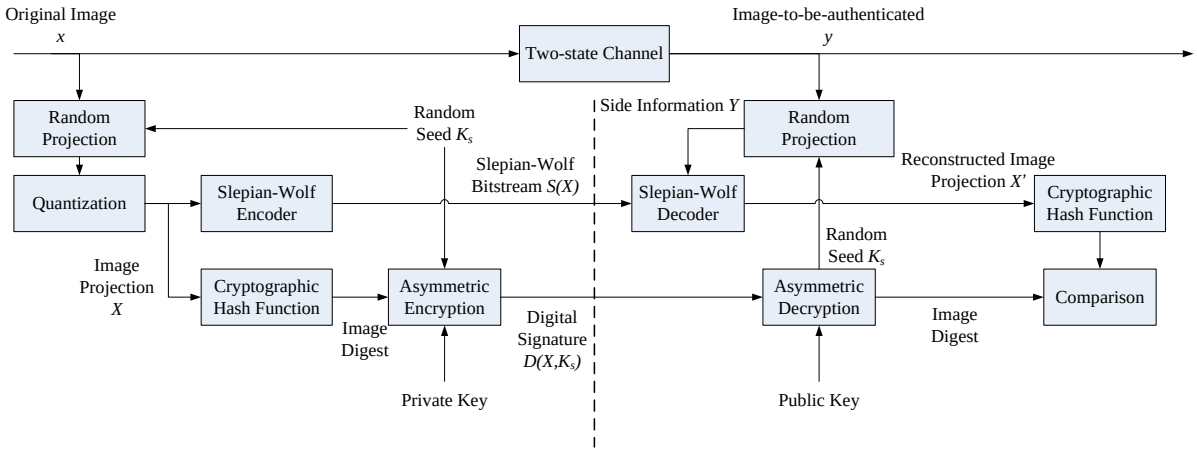


Fig. 1. Image authentication and tampering localization systems based on distributed source coding

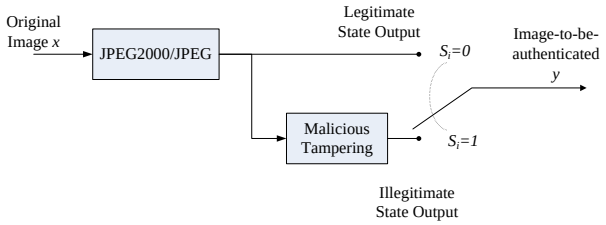
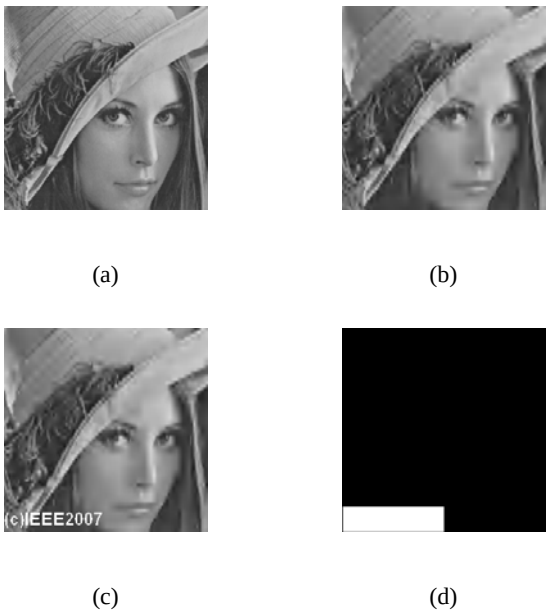


Fig. 2. Space-varying two-state lossy channel



(a) (b) (c) (d)

Fig. 3. Portion 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).

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 [13]. 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. 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 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 authentication decoder, in the right-hand side of Fig. 1, 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 [14] initialized according to the known 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.

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 [15]. 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.

3. SPATIAL MODELS FOR TAMPERING LOCALIZATION

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 $D(X)$.

The localization decoder requires more information than the authentication decoder. Fortunately, since we use rate-adaptive LDPC codes [13] for Slepian-Wolf coding, the localization decoder re-uses the authentication data. Incremental localization data is sent through the Slepian-Wolf bitstream $S(X)$. The amount of incremental localization data required depends on the accuracy of the tampering model for the channel states. We consider three models, shown in Fig. 4. The first is applied in [2] and assumes the channel states are independent. The second takes rows of channel states to be 1D Markov chains and the third models the channel states according to a 2D Markov random field.

These spatial models are concatenated with a symbol model from [16] and the LDPC decoding graph from [14] to produce the factor graph in Fig. 5. The estimation of the channel state likelihoods $P(S_i = 1)$ proceeds via iterations of the sum-product algorithm [17].

In particular, the spatial models receive messages $u^i_{f_b \to s}$ from factor nodes f^i_b, and reply with messages $u^i_{s \to f_b}$. The decoder for the 1D Markov model in Fig. 4(b), being parameterized by probability $f_t(S_i, S_{i-1}) = P(S_i|S_{i-1})$, achieves this with one iteration of the Baum-Welch algorithm (without re-estimation) [18]. The 2D Markov random field in Fig. 4(c) is parameterized by probability $f_t(S_i, S_{i-1}, S_{i-w}, S_{i-w-1}) = P(S_i|S_{i-1}, S_{i-w}, S_{i-w-1})$, and so employs a modified Baum-Welch iteration similar to that of [11]. The forward and backward recursions are

$$\alpha^i(s) = \sum_{s_{i-1}, s_{i-w}, s_{i-w-1}} \Big(P(s|s_{i-1}, s_{i-1}, s_{i-w-1}) \prod_{j \in \{i-1, i-w, i-w-1\}} \alpha^j(s_j) u^j_{f_b \to s}(s_j) \Big)$$

$$\beta^i(s) = \sum_{s_{i+1}, s_{i+w}, s_{i+w+1}} \Big(P(s_{i+w+1}|s_{i+1}, s_{i+w}, s) \prod_{j \in \{i+1, i+w, i+w+1\}} \beta^j(s_j) u^j_{f_b \to s}(s_j) \Big).$$

The resulting message $u^i_{s \to f_b}$ is given by

$$u^i_{s \to f_b}(s) \propto \alpha^i(s) \beta^i(s),$$

which is normalized such that $u^i_{s \to f_b}(0) + u^i_{s \to f_b}(1) = 1$.

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 $P(S_i = 1) > T$, a fixed decision threshold.

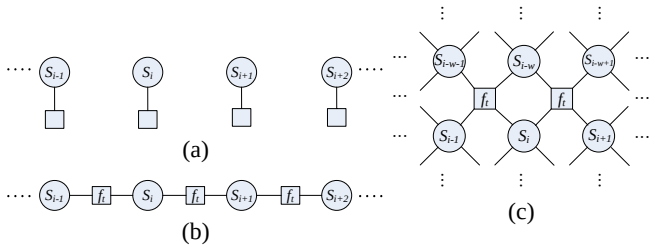


Fig. 4. Factor graphs of spatial models for the channel states (a) independent, (b) 1D Markov chain, (c) 2D Markov random field.

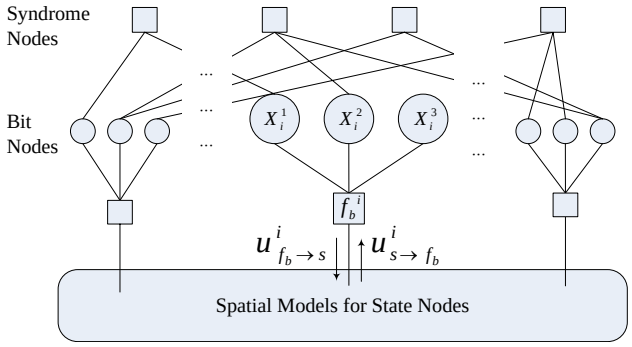


Fig. 5. Factor graph for the localization decoder

4. SIMULATION RESULTS

In practice, the localization decoder only runs 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 "Lena" 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 qualities above 30dB. The malicious tampering consists of the overlaying of up to five text banners of different sizes (198x29, 29x254, 119x16, 16x131, 75x75) at random locations in the image. The text color is white or black, depending on which is more visible to avoid generating trivial attacks.

We compare the minimum authentication data rate required by the authentication decoder to the average minimum authentication plus localization data rates of the localization decoders with various spatial models. Fig. 6 shows the Slepian-Wolf bitstream portion $S(X)$ of these rates (in bits per pixel of the original image x) for "Lena" with X quantized to 4 bitplanes. All five text banners are placed for malicious tampering, since greater tampering makes 'disauthentication' easier and localization more difficult. The placement is random for 100 trials, leading to tampering of 11% to 15% of the 16x16 blocks of the original image x. The decision threshold T is set to 0.5. The required authentication plus localization rate is roughly twice the authentication rate, but the 1D and 2D spatial decoders reduce the average authentication plus localization rate. In the worst case over 2000 trials, the largest bitstream sizes are 232, 208 and 192 bytes for the independent, 1D and 2D spatial models, respectively; savings of 10% and 17% for the 1D and 2D spatial models.

Using these Slepian-Wolf bitstream sizes, we measure various failure rates. The blockwise falsely deemed tampered rate is the proportion of untampered blocks mistaken for tampered blocks; and vice versa for the blockwise falsely deemed untampered rate. Fig. 7 shows failure rates for X quantized to 4 bitplanes as T varies, indicating that the blockwise falsely deemed tampered rates can be set to zero, while keeping the blockwise falsely deemed untampered rates near 10%. The corresponding pixelwise falsely deemed untampered rates are near 1%, since most of the blocks falsely deemed untampered have only a few pixels tampered, and improve along with the spatial model.

5. CONCLUSIONS

We compare localization of image tampering with various spatial models under the framework of [1] and [2]. The 1D and 2D spatial

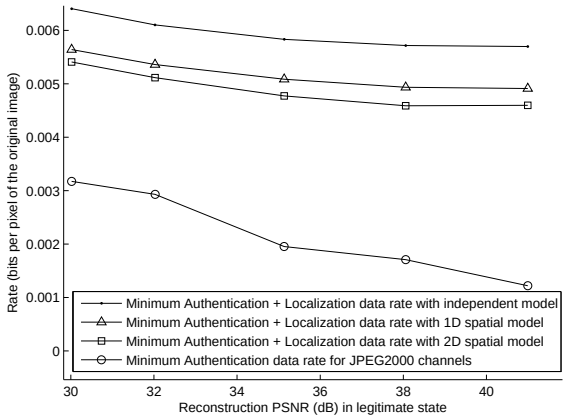


Fig. 6. Minimum authentication and authentication plus localization data rates for decoding Slepian-Wolf bitstream $S(X)$ for "Lena"

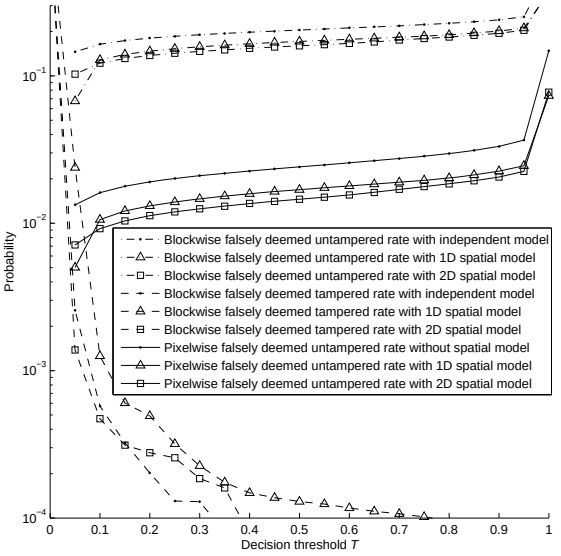


Fig. 7. Localization decoder failure rates versus T

decoders exploit the contiguity of the tampered regions and reduce the authentication and localization data sizes by 10% to 17% compared to an independent model, while offering improved confidence up to 99% in classifying tampered pixels. Future work should analyze the trade-off between sensitivity and robustness and compare the proposed scheme with other image authentication systems.

6. REFERENCES

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

[2] Y.-C. Lin, D. Varodayan, and B. Girod, "Image authentication and tampering localization using distributed source coding," in *IEEE Multimedia Signal Processing Workshop*, Crete, Greece, Oct. 2007.

[3] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, May 2001.

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

[5] C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing JPEG compression from malicious manipulation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 2, pp. 153–168, Feb. 2001.

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

[7] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. IT-22, no. 6, pp. 644–654, Jan. 1976.

[8] E. Martinian, S. Yekhanin, and J. S. Yedidia, "Secure biometrics via syndromes," in *Allerton Conference on Communications, Control and Computing*, Monticello, IL, Sep. 2005.

[9] S. C. Draper, A. Khisti, E. Martinian, A. Vetro, and J. S. Yedidia, "Using distributed source coding to secure fingerprint biometric," in *IEEE International Conference on Acoustics, Speech, and Singal Processing*, Honolulu, HI, April, 2007.

[10] Q. Sun, S.-F. Chang, M. Kurato, and M. Suto, "A new semi-fragile image authentication framework combining ECC and PKI infrastructure," in *IEEE International Symposium on Circuits and Systems*, Phoenix, AZ, May 2002.

[11] D. Varodayan, A. Aaron, and B. Girod, "Exploiting spatial correlation in pixel-domain distributed image compression," in *Picture Coding Symposium*, Beijing, China, April 2006.

[12] D. Schonberg, S. C. Draper, and K. Ramchandran, "On compression of encrypted images," in *IEEE International Conference on Image Processing*, Atlanta, GA, Oct. 2006.

[13] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive codes for distributed source coding," *EURASIP Signal Processing Journal, Special Section on Distributed Source Coding*, vol. 86, no. 11, pp. 3123–3130, Nov. 2006.

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

[15] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.

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

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

[18] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, Oct. 1970.