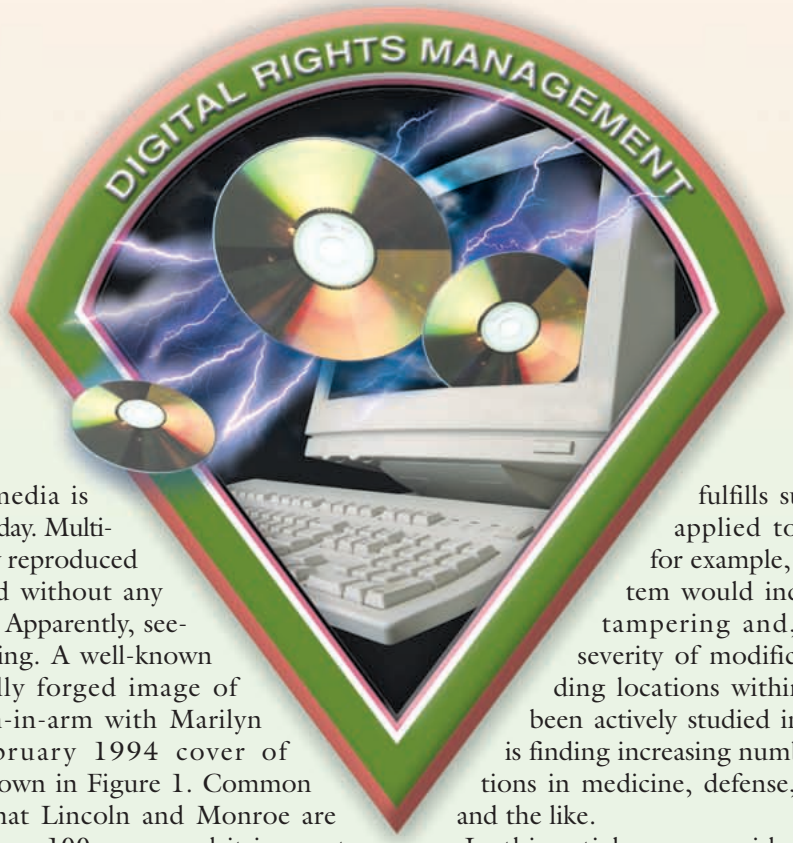
When Seeing Isn't Believing

Bin B. Zhu, Mitchell D. Swanson, and Ahmed H. Tewfik

Current multimedia authentication technologies and their applications.



LOGO COMPOSITE: ©1995 PHOTODISC, INC., ©DIGITAL STOCK, ©COREL

Digital multimedia is ubiquitous today. Multimedia is easily reproduced and modified without any trace of manipulations. Apparently, seeing is no longer believing. A well-known example is the digitally forged image of Abraham Lincoln arm-in-arm with Marilyn Monroe on the February 1994 cover of *Scientific American*, shown in Figure 1. Common knowledge indicates that Lincoln and Monroe are separated in birth by over 100 years, and it is easy to conclude that the image is forged. However, the image appears visually genuine.

In most cases, a human will not be able to judge whether a multimedia signal is authentic by perceptual inspection. Ideally, the integrity and authenticity of multimedia data is ascertained by a system without access to information external to the challenged multimedia data itself, e.g., common-knowledge, original multimedia signals. Multimedia authentication (MA) fulfills such a purpose. When applied to the Lincoln image, for example, an authentication system would indicate the subsequent tampering and, in some cases, the severity of modification and corresponding locations within the image. MA has been actively studied in the past decade and is finding increasing numbers of critical applications in medicine, defense, commerce, industry, and the like.

In this article, we provide a compact yet fairly comprehensive introduction of MA to the general signal processing audience. MA inherits many characteristics from a generic data authentication such as integrity verification, origination verification, nonrepudiation, and security, which are well discussed in [1]. However, MA has a few unique features that render generic data authentication algorithms well studied in cryptography inadequate or undesirable. Unlike other data, a multimedia signal can be represented equivalently in differ-

1053-5888/04/$20.00©2004IEEE

ent forms or formats, e.g., an image represented in the Joint Photographic Experts Group (JPEG) format that is subsequently converted to the graphic interchange format (GIF) format carries exactly the same visual information. MA seeks to authenticate the multimedia content instead of its specific binary representation.

For more comprehensive coverage of the subject, interested readers are referred to [2].

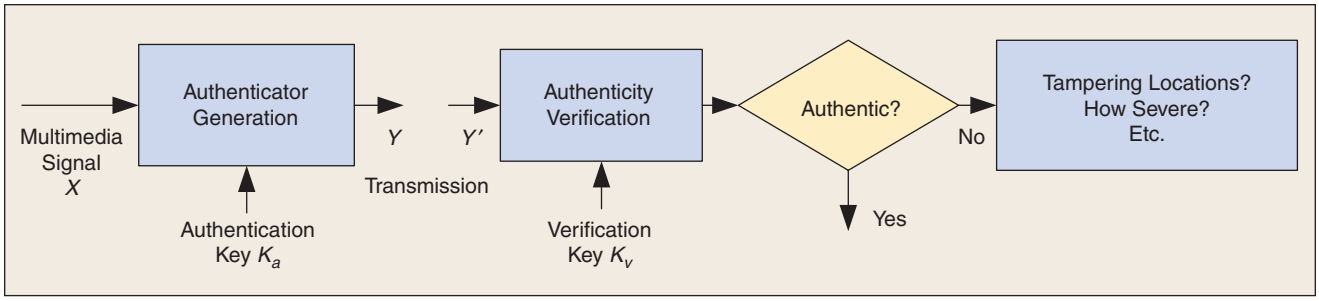
MA—A Brief Background and Applications

Multimedia signals typically contain a great amount of data. Many applications allow or even require certain processing, such as near-transparent compression, to be applied to multimedia without affecting its authenticity due to high redundancy and perceptual irrelevancy present in the signal. MA should be able to discriminate malicious manipulations from admissible manipulations. Other desirable features for MA include localization of tampered regions and indication of tamper severity and characteristics so that the untampered portion can still be used and the altered content can be analyzed to determine if the semantic meaning is preserved and if the alteration is recoverable. Another desirable feature is to determine authenticity of a received segment of a signal, especially for an audiovisual signal, which typically has a long play time. In applications where the multimedia authenticator is generated and verified by different parties, it is desirable that knowledge for verification cannot deduce the secret to generate the authenticator. It is worth noting that some of these requirements are mutually competitive. A reasonable compromise is always necessary in the design of an authentication system.

A general MA system is depicted in Figure 2. A multimedia signal X along with an authentication key K_a is input into an authenticator generation system to generate an authenticator T, which is then either tagged or embedded to the signal X and output to the result Υ. In the verification stage, a received segment or whole signal Υ' is input into a verification system along with the verification key K_v, where the tagged or embedded authenticator T is extracted and compared with the authenticator calculated from the received signal to determine if Υ' is authentic. In the watermarking case, a verification system may extract the watermark from Υ' and compare with some a priori knowledge to make a decision. Some authentication systems may also give more information such as tamper locations and/or severity, etc., when a received signal is determined not to be authentic. In a practical system, the verification key or the a priori knowledge used for verification should be content agnostic, i.e., independent of the multimedia content, either the original or the challenged. This requirement rules out the possibility of using the original signal at the verification stage. Some of watermarking-based authentication algorithms proposed in the literature do not meet this content agnostic requirement and do not apply within the scope of this article.

▲ *1. Digitally forged image on the February 1994 cover of* Scientific American. *Courtesy of Jack Harris, who manipulated the image.*

MA can be classified according to integrity criteria into hard (or complete) and soft authentication. Hard authentication rejects any modifications to multimedia content. The only manipulation accepted by the hard authentication is lossless compression or format conversion that preserves visual pixel values or audio samples. This is similar to the classical authentication except that those lossless operations are also rejected by the classical authentication. Soft authentication passes certain content modifying, called incidental or admissible



▲ *2. Flowchart of a general multimedia authentication system.*

The major challenge in content-based authentication is to define a computable feature vector that can capture the major content characteristics from a human perspective.

manipulations and rejects all the rest, called malicious manipulations. Soft authentication can be further divided into quality-based authentication, which rejects any manipulations that lower the perceptual quality below an acceptable level, and content-based authentication, which rejects any manipulations that change the semantic meaning of the content. Classification of acceptable and unacceptable manipulations depends on a specific application. Soft authentication typically measures distortion in some metric between a feature vector from the received signal and the corresponding vector from the original signal and compares with a preset threshold to make a decision on challenged signal's authenticity. There is typically no sharp boundary between authentic and inauthentic signals. In many applications, it is often difficult to distinguish distortion caused by an incidental manipulation from that caused by a malicious manipulation. This intrinsic fuzziness makes the soft authentication design challenging and, likely, ad hoc in most cases. Many soft authentication systems give a confidence (or a degree) of authentication instead of binary outputs.

We have already described the difference between MA and generic data authentication and general requirements for MA. We will characterize in detail in the following sections several prominent MA algorithms proposed in the literature. In the section that follows, algorithms for hard authentication are described, along with their vulnerabilities and fixes. Algorithms for soft authentication are described in the next section. This section includes algorithms for quality-based authentication and content-based authentication. We conclude the article with existing issues and future research directions. Before we move to the next section, we characterize several common attacks.

In designing any practical authentication system, threat models and attacks need to be considered. While it is impossible or impractical to design a system to resist all forms of possible attacks, good authentication systems should be designed to survive common operations designed to reduce their effectiveness. In addition to common attacks discussed in detail in [1] for classical authentication, some operations may be designed to exploit specific features in MA. Some of these common attacks include the following:

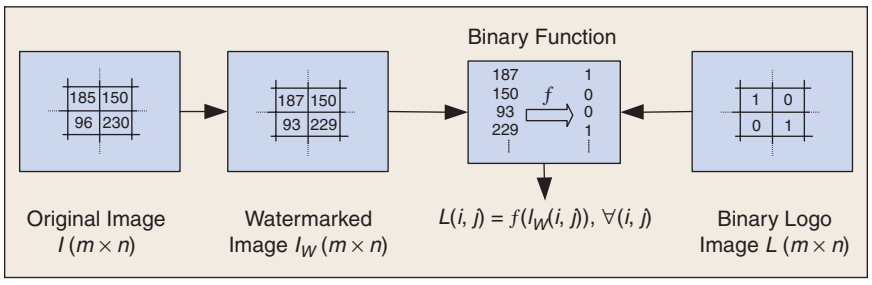
▲ *Undetected modification*—High redundancy and strong correlation in multimedia may be exploited to (maliciously) modify an authenticated media without being detected. Ill-defined distinction between incidental and malicious manipulations for soft authentication aggravates the problem. Tamper localization may enable an attacker to mount a successful attack by swapping components within the same signal or among different signals.

▲ *Authenticator transfer*—The same high redundancy and strong correlation of multimedia may also be exploited to forge a valid authenticator for an arbitrary media signal from available authenticated signals. A famous mark transfer attack is the vector quantization attack proposed by Holliman and Memon [3] to watermark-based block-wise authentication algorithms.

▲ *Information leakage*—Large amounts of data in a multimedia signal or structures in the underlying secret information may be exploited to deduce the secret information or key, or to dramatically reduce the search space. Once the key is deduced, the attacker can then forge a valid authenticator to an arbitrary signal. The authenticity verifier may also be exploited by an attacker to achieve the same goal.

Hard Authentication

Hard authentication rejects any modification to a multimedia signal. Most proposed algorithms are based on fragile watermarking so the authenticator is embedded into the signal to be authenticated to simplify bookkeeping and maintenance of authenticators. In fragile watermarking, the inserted watermark is so weak that any manipulation to the multimedia content disturbs its integrity. Tampered parts of the multimedia signal may be located by checking the presence and integrity of the local fragile watermark. In this section, three major hard authentication schemes will be described, with image authentication as an example. These technologies can be applied to other media types with minor modifications. For example, the frame index should be used in generating a video authenticator to avoid a frame-reordering attack.



▲ *3. Yeung-Mintzer fragile watermarking scheme [4] for grayscale images.*

Single Pixel/Sample Authentication

A simple approach referred as the Yeung-Mintzer scheme [4], which enables single pixel authentication, is shown in Figure 3 for grayscale

images. The watermarked image I_w is generated by disturbing the original image I to enforce the relationship $L(i, j) = f(I_w(i, j)), \forall(i, j)$, where L is a secret logo and f is a secret binary function. A simple way to generate the binary function f is to flip a coin for each possible pixel value. Illustrative pixel and logo values are also shown in the figure. Error diffusion is employed to reduce perceptual artifacts from the disturbance. Tampered pixels are found by examining visually or against the original logo the resulting binary image obtained by applying f to the challenged image. This scheme can be easily extended to color images and other multimedia signals.

This approach can locate a tampered pixel but only half of modified pixels on average can be detected since each pixel is individually mapped to a binary value. The scheme's security depends critically on the secrecy of the underlying logo. Once the logo is known or its structure is exploited, the search space for the secret mapping function f is dramatically reduced [5], and the vector quantization attack described in the next section can be successfully mounted. Even if the logo is not exploited, if the same logo and mapping function f are reused in watermarking images, only two watermarked grayscale images are needed on average to recover 90% of f by simply solving the equations of f at all pixels [6]. Once f is known, the secret logo can be readily derived. Making the mapping function dependent on pixel positions and a unique image index or on processed neighbor pixels can dramatically enhance the security [5], [7]. Figure 4 shows the scheme proposed in [7], which replaces f with the parity of the ciphertext after encryption is applied to a block including both the current pixel and previously processed neighbor pixels. Pixels are modified in rows to enforce the desired relationship. Such a scheme removes the above-mentioned vulnerabilities. Detection sensitivity to a single modified pixel is also increased at the cost of reduced accuracy to localize a tampered pixel, thanks to the pixel being used in watermarking both the current and following neighbor pixels.

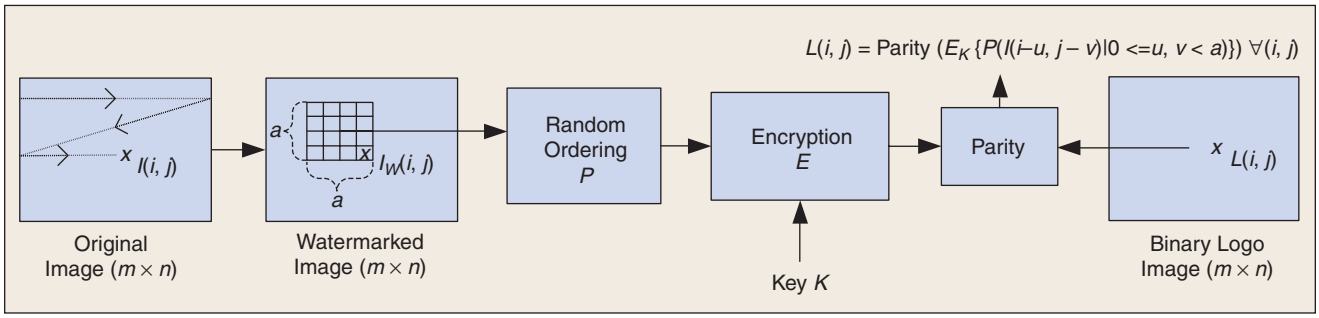
If the sequence to modify pixels is known and the verifier indicates tampered pixels, these schemes suffer an oracle attack [8]: for an arbitrary image, each pixel is modified until the output from the verifier indicates the current pixel has not been tampered with. The same sequence is applied until the whole image is processed. This attack needs on average trials of twice the number of pixels in the image.

Multimedia authentication should be able to discriminate malicious manipulations from admissible manipulations.

Block Authentication

Another approach is to partition a multimedia signal into two disjointed parts: a signature part and an embedding part. The signature part captures all the significant information of the signal. An authenticator such as a message authentication code (MAC) or a digital signature is generated from the content of the signature part and is then embedded into the embedding part. One of the first fragile watermarking techniques was to insert key-dependent check sums of the seven most significant bits into the least significant bits (LSBs) of pseudo-randomly selected pixels [9]. Figure 5 shows a well-studied block-based fragile image watermarking technique referred as Wong scheme [10], where the hash value from a block with LSB zeroed out is XORed with the corresponding block of the binary logo image, encrypted, and inserted to the LSB of the block. To verify an image, the LSBs are extracted, decripted, and XORed with the hash value calculated from the challenged image in the same way as shown in Figure 5. If the result is the original binary logo, then the image is authentic. Any tamper to a block will generate a very different binary output for the block due to the property of the hash function.

If two blocks have identical logos, they can be swapped without detection. This can be avoided by using random logos or including a block index to the input to the hash function in Figure 5. This swapping is still possible among blocks of different images if they are authenticated with the same key and identical logos. A vector quantization attack can also be mounted to



▲ *4. Enhanced Yueng-Mintzer scheme uses a mapping function that depends on processed neighbor pixels [7].*

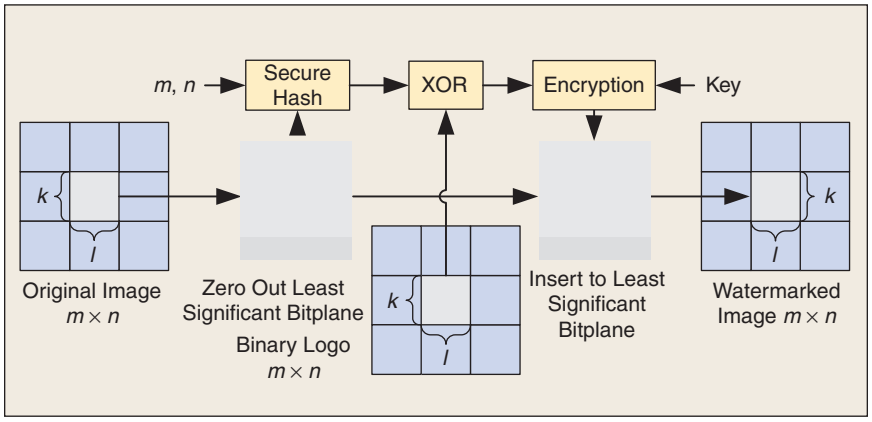
authenticate an arbitrary image [3], [6]: an arbitrary block is approximated by the closest authentic block from a collage of authentic blocks authenticated with the same key and the desired logo. The quality of the forgery depends on the number of authentic blocks available. A remedy is to include a unique image ID in input to the hash function. An elegant solution is proposed in [8] where the logo for each block contains both block index and image index plus redundancy to check logo validity. Another solution is to introduce dependency on neighbor pixels [3], [7], [11] such as the scheme shown in Figure 4 or to build a hierarchical tree structure where a leaf block embeds its own signature as well as part of signature bits from its ancestors [12].

Lossless Watermarking for Authentication

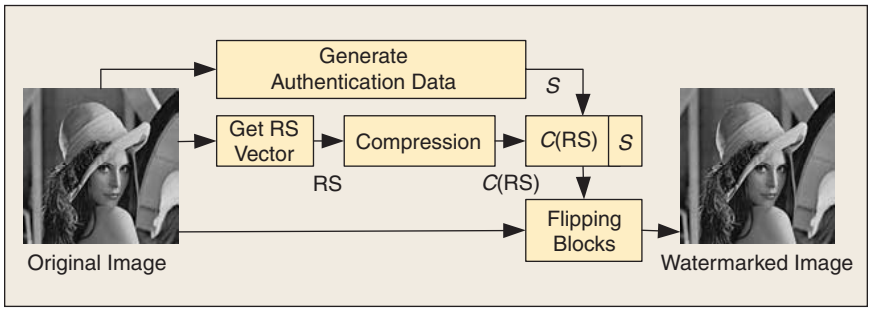
All the previously described watermarking-based schemes introduce small and irreversible distortion to signals to be authenticated. It is often desirable to design an authentication system that incurs no distortion to underlying signals like the classical data authentication yet the authenticator is still embedded into the signal for easy storage and maintenance of authenticators. This can be achieved with recently proposed lossless watermarking schemes [13]–[16].

One approach [13], [14] is to use a spatially additive, signal-independent robust watermarking to embed signal authentication data using a reversible modular addition. The watermark has to be robust enough to survive the reversible addition in the watermarking process so that for an unmodified watermarked signal the authentication data can be correctly recovered and the original watermark can be subsequently regenerated and removed from the watermarked signal to recover the original signal. The amount of authentication data is typically constrained by the limited embedding capacity of the underlying robust watermark. Another approach [14]–[16] is to losslessly compress some perceptually insignificant signal component, such as the LSB bit plane, for an image so the original component can be replaced by its compressed version appended with the authentication data for the signal. This scheme works only for signals whose components can be compressed to leave enough bits for authentication data.

A lossless watermarking scheme proposed in [15] is shown in Figure 6 for grayscale images where a binary sequence called "RS vector" is obtained from blocks partitioned from the original image, losslessly compressed and appended with the image's authentication data, and then embedded into the image by adjusting the image to generate an RS vector that matches the resulting vector. To generate an RS vector, a discrimination function f such as variation is defined to measure the smoothness for a block and another invertible flipping function F to permute pixel values such that $F(F(x)) = x, \forall x$. A block B is classified into either R-type, S-type, or U-type, depending on if $f(F(B))$ is greater than, less than, or equal to $f(B)$. The RS vector for an image is generated by scanning image blocks in a certain order and assigning 1 to R-type blocks and 0 to S-type blocks. Only R-type and S-type blocks are used for embedding data; U-type blocks are ignored. If the type of a block does not match the bit to be embedded, F is applied to the block to flip the block's type from R-type to S-type, or vice versa. Authenticity is verified by extracting the embedded authentication data and the original RS vector from the RS vector calculated from the challenged image, adjusting the challenged image as necessary by flipping mismatched blocks to match the extracted original RS vector and comparing the extracted authentication data with that calculated from the resulting image. The challenged image is verified to be authentic and the resulting image is the original image if no difference is found.



▲ 5. Block-based fragile watermarking scheme for image authentication [10].



▲ 6. RS lossless watermarking for authentication [15].

These proposed lossless watermarking schemes can authenticate an image as a whole. They cannot indicate tamper locations once modification occurs. The perceptual quality of the watermarked signal needs to be considered to avoid annoying watermarking artifacts even though watermarking can be reversed for authentic signals. Some signals may not be able to be authenticated by a lossless watermarking scheme if there is not enough available space to embed the authentication data.

Soft Authentication

In many applications, a multimedia signal may be processed after its generation. For example, a video clip may be transcoded to match the targeted devices. If hard authentication is used in these applications, any intermediate stage that performs legitimate processing on the multimedia signal will have to first verify authenticity of the signal to be processed and then authenticate the processed signal. This means that both authenticator generation and verification secrets have to be shared with these intermediate stages. Because of high correlation and perceptual irrelevancy in multimedia signals, some modifications, such as high-quality lossy compression, do not generate detectable perceptual distortion to human end-users. Multimedia signals that are modified yet retain their original perceptual quality or semantic meaning are desired to be considered as authentic in many applications. Two types of MA algorithms are described in this section. In the following section, we describe authentication algorithms that accept only manipulations that preserve the perceptual quality. In the section after that, we discuss authentication algorithms that accept only those manipulations that preserve the multimedia's semantic meaning.

Quality-Based Authentication

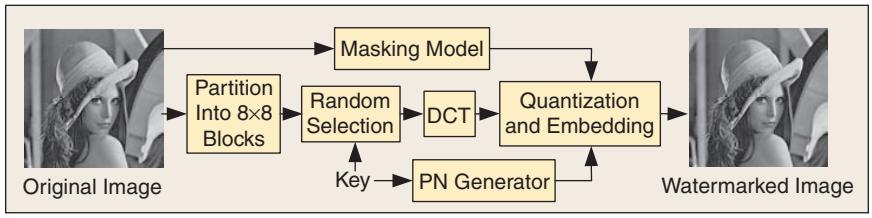
One of the first quality-based authentication algorithms uses a quantization-based watermarking scheme to embed a pseudorandom pattern into a signal to check integrity and measure distortion [17], [18]. Figure 7 shows a variant of the scheme, where a set of image blocks are pseudorandomly selected to embed data for distortion measurement. Each selected block n is first transformed into the discrete cosine transform (DCT) domain, and then each frequency bin $F^n_{i,j}$ is modified as $F^n_{i,j} \rightarrow F^{n,W}_{i,j} = M^n_{i,j}\{\lfloor F^n_{i,j}/M^n_{i,j} \rfloor + r^n_{i,j} \cdot \text{sign}(F^n_{i,j})\}$, where $r^n_{i,j}$ is a key-based random number in the interval $(0, 1)$, $M^n_{i,j}$ is the masking value for the frequency bin, $\lfloor \cdot \rfloor$ rounds towards 0, and sign(x) is 1 or -1, depending on if x is nonnegative or negative. These blocks are then transformed back to the spatial domain to get the watermarked image. To find distortion for a challenged image, each embedded DCT coefficient $F^{n,T}_{i,j}$ and its corresponding masking value $M^{n,T}_{i,j}$ are calculated from the challenged image, and the distortion at the frequency bin is estimated as

$$\hat{e}^n_{i,j} = F^{n,T}_{i,j} - M^{n,T}_{i,j} \cdot \left\{ r^n_{i,j} \cdot \text{sign}\left(F^{n,T}_{i,j}\right) + \left\lfloor F^{n,T}_{i,j} / M^{n,T}_{i,j} - \left(r^n_{i,j} - 0.5\right) \cdot \text{sign}\left(F^{n,T}_{i,j}\right) \right\rfloor \right\}.$$

It is shown that the estimation is rather accurate if distortion to the frequency bin is up to half of $M^n_{i,j}$ [17], [18]. Local or global distortion can be found by a weighted sum of distortion to each frequency bin and then compared with a preset threshold to check authenticity and find out tamper locations. The type of modification can also be estimated by examining the distortion pattern at different frequency bins and locations.

The above-estimated distortion for an undistorted watermarked image is small but not the desired 0. This small error is caused by a small difference in masking values calculated from the original and watermarked images and by the integer-rounding error introduced when transforming back to the spatial domain. The first factor can be removed by replacing masking values with a signal-independent quantization vector such as a JPEG quantization table. The latter can be removed by embedding in the spatial domain. An elegant solution is proposed in [19] where possible modifications to a wavelet coefficient f_l at a resolution l after Haar transform are $\pm 2^l \Delta$, where Δ is a secret quantization parameter, so that any resulting spatial domain modifications during watermarking are integers, and the

Content-based authentication passes multimedia as authentic when the semantic meaning of the signal remains unchanged, i.e., the content does not change.



▲ *7. A variant of quantization-based watermarking for authentication proposed in [17] and [18].*

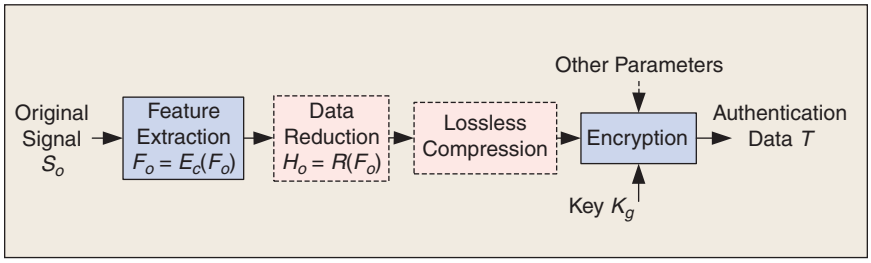
Quality-based authentication still needs work to improve robustness to incidental changes yet remain sensitive to malicious modifications.

distortion caused by rounding pixel values to integers is avoided. In [19], a set of wavelet coefficients $\{f_l(i)\}$ are randomly selected based on a secret key, and the equation $Q_{\Delta,l}(f_l(i)) = w(i)\,\text{XOR}\,q\text{key}(i)$ is enforced for each selected coefficient $f_l(i)$, possibly modified by subtracting $2^l\Delta \cdot \text{sign}(f_l(i))$ from it if necessary, where $Q_{\Delta,l}(f_l(i))$ equals 0 or 1 depending on if $\lfloor f_l(i)/(2^l\Delta)\rfloor$ is even or odd, $w(i)$ is the corresponding watermark bit, and $q\text{key}(i)$ is the corresponding bit from a bit sequence generated from the image and a secret key. To check authenticity, the selected wavelet coefficients $\{f'_l(i)\}$ are used to extract the embedded watermark: $w'(i) = Q_{\Delta,l}(f'_l(i))\,\text{XOR}\,q\text{key}(i)$. The percentage of correctly recovered watermark bits is used to estimate tamper extent. Note that if a wavelet coefficient is close to boundary of regions that map to different binary values, a minor change may make it cross the boundary to result in a wrong extracted watermark bit.

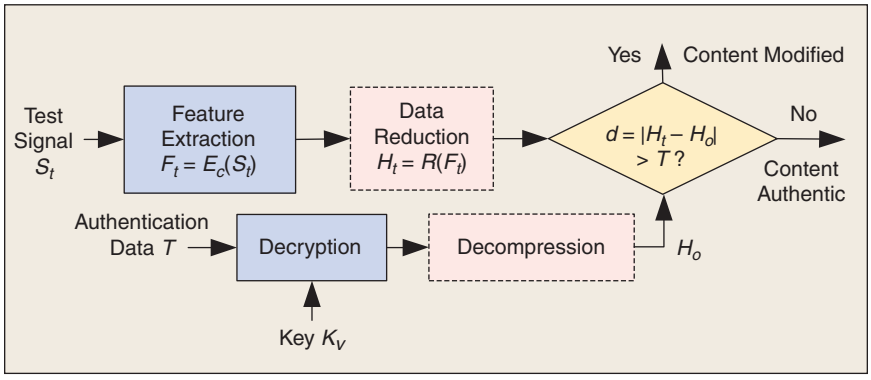
These two algorithms cannot detect modifications that are multiples of watermarking quantization steps, which may be exploited to pass as authentic a signal with large modification. They both use some collective measures to differentiate malicious from incidental manipulations. An alternative approach is to define a threshold for maximum allowed distortion to each pixel as proposed in [20], where the resulting image after each pixel is quantized by $2k+1$ is treated as the "original" image in applying the Yeung–Mintzer scheme for authentication. The scheme allows changes to a pixel within distortion $\pm k$ and rejects any changes outside this bound. Similarly, an image can be partitioned and vector quantized, and the resulting image is authenticated with the Wong scheme. A drawback for these alternative solutions is that the watermarked signal may be distorted to the lowest bound of authenticity, which results in large perceptual distortion. A recent article [21] tried to address this issue by choosing and remembering one of two quantization steps for each coefficient to be quantized. Note that a misaligned or cropped signal may be determined as inauthentic without a possibly extensive and time-consuming search to properly align the challenged signal.

Content-Based Authentication

Content-based authentication passes multimedia as authentic when the semantic meaning of the signal remains unchanged, i.e., the content does not change. Media content is represented by a feature vector extracted from the media. It is this content, i.e., the extracted feature vector, instead of the multimedia signal itself, that is authenticated in content-based authentication. The general structures for multimedia authenticator generation and verification for content-based authentication are shown in Figures 8 and 9, respectively. In authenticator generation, a feature vector is extracted from media, followed by an optional data-reduction stage and another optional lossless compression to reduce amount of data in the feature vector. The result is authenticated by MAC or a digital signature that is either tagged or embedded to the media with robust or semifragile watermarking. If watermarking is used, the media is typically partitioned depending on a secret key into disjoint signature and watermark subspaces for authentication data generation and embedding, respectively, without interference. If the two subspaces overlap, great care is necessary to avoid false alarm or reduced tolerance caused by watermarking distortion. Iteration is typically used in this case to reduce the adverse



▲ 8. Generating multimedia authenticator in content-based authentication.



▲ 9. Authenticity verification in content-based authentication.

impact of watermarking on feature extraction, but it is difficult to prove that such iteration converges. In the verification stage, the embedded authentication data is extracted, decrypted, and decompressed if necessary, to get the original feature vector, which is compared to the feature vector calculated from the challenged signal. If their difference measured in some metric such as L_2 norm is smaller than a preset threshold, the content of the signal is deemed authentic. Tamper locations may be found by measuring local distortion of the feature vectors.

The major challenge in content-based authentication is to define a computable feature vector that can capture the major content characteristics from a human perspective. This remains a research challenge. All proposed content-based authentication algorithms use very heuristic features to represent multimedia content. The features proposed for images include block histograms [22], averages [23], or lower-order moments [24]. Other features include image edges [24], [25], zero-crossings [26], and "salient" feature points extracted from the Mexican-Hat wavelet transform [27].

An elegant scheme to accept JPEG compression yet reject other manipulations is proposed in [28], which exploits the fact that an inequality between two DCT coefficients of the same frequency in two arbitrary blocks still holds or changes to equality after JPEG compression since the same quantization step is used to quantize the two coefficients. To generate a feature vector, image blocks are pseudorandomly grouped into pairs, and some DCT coefficients from each pair of blocks are selected to generate feature bits. For each pair of DCT coefficients, if the first coefficient is less than the second, 0 is generated; otherwise 1 is generated. The generated feature vector, together with image size and mean values of DCT coefficients at each selected frequency bins, are encrypted to generate the authentication data for the image. In authenticity testing, the feature bits generated from the challenged image are compared against the original feature bits. If any unmatched bits exist beyond a tolerance-bound $\tau \geq 0$; for example, if an original bit is 1, i.e., $\Delta_{i,j}^{\text{orig}}(f_k) \equiv \text{DCT}_i^{\text{orig}}(f_k) - \text{DCT}_j^{\text{orig}}(f_k) \geq 0$ for DCT blocks i and j at a frequency f_k, and the corresponding inequality for the test image is $\Delta_{i,j}^{\text{test}}(f_k) < -\tau$, then the scheme concludes that either block i or j or both in the testing image have been manipulated by some operation other than JPEG. The small tolerance-bound τ is used here to avoid small difference caused by rounding pixel values to integers, etc. The difference between two DCT coefficients can be further refined by comparing with some nonzero values due to the fact that an inequality still holds or changes to equality after JPEG compression if these nonzero values are also quantized in a certain manner by the same JPEG quantization table [28]. For example, let $\Delta_{i,j}^{\text{orig}}(f_k) > k$ for a fixed value k. After JPEG compression with a quantization table Q, we have $\Delta_{i,j}^{\text{JPEG}}(f_k) \geq \{[k_Q] - 1 + \delta(k_Q - [k_Q])\} \cdot Q(f_k)$, where $k_Q = k/Q(f_k)$, $[x]$ rounds x to the closest integer, and $\delta(x)$ is 1 for $x = 0$ and 0 for nonzero x. Here we have ignored the effect that quantization on DCT coefficients may result in fractional or out-of-bound pixel values in the spatial domain. A drawback for this refinement is that the JPEG quantization table used in JPEG compression is needed for verification, which requires test images in the JPEG format. Another drawback is that the second generation of JPEG compression is rejected if the first generation of JPEG compression is of lower quality than the second generation. The same approach was also extended to video authentication to accept MPEG compression [29]. Features in multiple-pass encoding in JPEG 2000, such as the state of passes of the most significant bits and a measure of the change of "1" in a given bit plane associated with each pass, have been used to generate feature vectors to reject any modifications other than the JPEG 2000 compression with bit rates above a preset lowest acceptable bit rate [30].

Content-based authentication depends on improvements in feature vectors that capture multimedia content from a human's perspective.

The aforementioned DCT-based authentication algorithm passes JPEG compression at any quality level yet rejects other modifications that may result in much better perceptual quality. This algorithm can be modified to reject JPEG compression below a certain quality level by applying a DCT-based quantization scheme to embed authentication data with the embedding quantization table to be one plus the JPEG quantization table at the coarsest acceptable JPEG quality [31]. Although the resulting watermarked image has a much higher PSNR value than that of the coarsest acceptable JPEG compression, thanks to less-modified DCT coefficients in watermarking, the perceptual quality of the watermarked image may be similar to the coarsest acceptable JPEG quality. This is because human eyes are sensitive to the worst artifact. Another drawback is that the scheme rejects JPEG compression with mismatched block partitions as compared to the authentication block partition. In addition, if the same key is used to authenticate multiple images, only $O(\log N)$ images, where N is the number of DCT blocks in an image, are needed to deduce the secret formation of pairs used to generate feature bits [32]. Once the pairs are known, an attacker can easily modify DCT coefficients yet keep the original relationships unchanged.

Soft authentication schemes require improvements to accept geometrical manipulations that preserve perceptual quality or semantic meaning.

In speech authentication, three types of features are proposed in [33]: pitch information, the changing shape of the vocal tract, and the energy envelope. They are extracted with the help of a CELP codec. The first three LSP coefficients—except the silent portion—are used as the pitch information. One pitch coefficient, used as the changing shape of the vocal tract, is obtained from each frame as the average of the "lag of pitch predictors" of all subframes except the nontonal part. The starting and ending points of silent periods and also nontonal regions are included in the authentication data. At the verification phase, distortion is calculated independently for each type of feature except silence periods for the first feature and nontonal regions for the second feature. A low-pass filter is applied to the resulting difference sequences before being compared with a threshold to determine a signal's authenticity.

Conclusion and Future Directions

We reviewed current MA technologies and classified them into two major types according to their integrity criteria. Hard authentication, usually based on fragile watermarks to detect modification to the underlying signal, has received the most coverage in the literature. Better tamper localization without sacrifice of security remains an issue to study. Soft authentication, on the other hand, is broken into two categories: quality based and content based. Quality-based authentication, often based on watermarks to measure signal modification within perceptual tolerance, still needs work to improve robustness to incidental changes yet remain sensitive to malicious modifications. Content-based authentication detects any manipulations that change signal's semantic meaning. To be robust to content-preserving modifications yet fragile to content-modifying manipulations, additional work is needed to define features that adequately describe the perceptual content of a multimedia signal. In soft authentication, lack of a clear-cut distinction between incidental and malicious modifications make it difficult to accurately characterize incidental manipulations from malicious manipulations. Many proposed schemes reject manipulations that may preserve better perceptual quality or semantic meaning than acceptable manipulations. A typical example is geometrical manipulations such as image scaling and rotations, which preserve perceptual quality but are likely to cause misalignment in verification and thus are rejected by most proposed authentication schemes. Differentiating between malicious and incidental manipulations in soft authentication remains an open research issue. More effort should be directed to develop soft authentication schemes to accept quality or semantic meaning preserving manipulations, such as incidental geometrical manipulations, and also to authenticate a portion of a multimedia—especially audio or video—signal.

Bin B. Zhu received his B.S. degree in physics from the University of Science and Technology of China in 1986 and M.S. and Ph.D. degrees in electrical engineering from the University of Minnesota in 1993 and 1998, respectively. From 1997 to 2001 he was a lead scientist at Cognicity, Inc., which he cofounded. He has been with Microsoft Research Asia since 2001. He has published two book chapters and about 30 peer-reviewed papers. He has six issued and ten pending U.S. patents. His current research interests include digital rights management and security, distributed multimedia networks, wireless communications, and multimedia authentication, watermarking, and compression. He is a Member of the IEEE.

Mitchell D. Swanson received the B.S. (summa cum laude), M.S., and Ph.D. degrees in electrical engineering from the University of Minnesota in 1992, 1995, and 1997, respectively. He started his career as a postdoctoral student at the University of Minnesota and went on to cofound Cognicity in 1997. He was a visiting assistant professor at the University of Minnesota during that time. In 2001, he joined General Dynamics Advanced Information Systems. He has published over 30 peer-reviewed papers and has a number of issued and pending patents. His current research interests include digital rights management, multimedia signal processing, and network and routing protocols for dynamic sensor networks. He is a Member of the IEEE.

Ahmed H. Tewfik received his B.Sc. degree from Cairo University, Egypt, in 1982 and his M.Sc., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1984, 1985, and 1987, respectively. He worked at Alphatech, Inc., Burlington, Massachusetts, in 1987. From 1997 to 2001, he was cofounder, president, and CEO of Cognicity, Inc. He was a Distinguished Lecturer of the IEEE Signal Processing Society in 1997–1999. He was awarded the IEEE Third Millennium Medal in 2000, the E.F. Johnson Professorship of Electronic Communications in 1993, a Taylor faculty development award from the Taylor foundation in 1992, and an NSF research initiation award in 1990. He was the first editor-in-chief of *IEEE Signal Processing Letters*

from 1993 to 1999 and a past associate editor of *IEEE Transactions on Signal Processing*. His current research interests are in signal processing for high performance wireless networks, pervasive datanomic computing, multimedia, and genomics. He is a Fellow of the IEEE.

References

[1] D.R. Stinson, *Cryptography, Theory and Practice*. Boca Raton, FL: CRC Press, 1995.

[2] B.B. Zhu and M.D. Swanson, "Multimedia authentication and watermarking," *Multimedia Information Retrieval and Management*, D. Feng, W.C. Siu, and H. Zhang, Eds. Springer-Verlag, 2003, ch. 7, pp. 148–177.

[3] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 432–441, 2000.

[4] M.M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE Int. Conf. Image Processing*, 1997, pp. 680–683.

[5] N. Memon, S. Shende, and P.W. Wong, "On the security of the yeung-mintzer authentication watermark," in *Proc. IS&T PICS Symp.*, Savannah, GA, Mar. 1999, pp. 301–306.

[6] J. Fridrich, M. Goljan, and N. Memon, "Further attacks on Yeung-Mintzer fragile watermarking scheme," in *Proc. SPIE Photonic West, Electronic Imaging 2000, Security and Watermarking of Multimedia Contents*, San Jose, CA, Jan, 24–26, 2000, pp. 428–437.

[7] J. Fridrich, M. Goljan, and A.C. Baldoza, "New fragile authentication watermark for images," in *Proc. IEEE Int. Conf. Image Processing*, 2000, pp. 446–449.

[8] J. Fridrich, "Security of fragile authentication watermarks with localization," in *Proc. SPIE Photonic West, vol. 4675, Electronic Imaging 2002, Security and Watermarking of Multimedia Contents*, 2002, pp. 691–700.

[9] S. Walton, "Information authentication for a slippery new age," *Dr. Dobbs J.*, vol. 20, no. 4, pp. 18–26, 1995.

[10] P.W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," *IEEE. Trans. Image Processing*, vol. 10, no. 10, pp. 1593–1601, 2001.

[11] D. Coppersmith, F. Mintzer, C. Tresser, C.W. Wu, and M.M. Yeung, "Fragile imperceptible digital watermark with privacy control," in *Proc. SPIE/IS&T Int. Symp. Electronic Imaging: Science and Technology*, San Jose, CA, 1999, vol. 3657, pp. 79–84.

[12] M.U. Celik, G. Sharma, E. Saber, and A.M. Tekalp, "A hierarchical image authentication watermark with improved localization and security," in *Proc. IEEE Int. Conf. Image Processing*, 2001, vol. 2, pp. 502–505.

[13] C.W. Honsinger, P.W. Jones, M. Rabbani, and J.C. Stoffel, "Lossless recovery of an original image containing embedded data," U. S. Patent No. 6,278,791, Aug. 21, 2001.

[14] J. Fridrich, M. Goljan, and R. Du, "Invertible authentication," in *Proc. SPIE, Security Watermarking of Multimedia Contents*, San Jose, CA, Jan. 2001, vol. 3971, pp. 197–208.

[15] J. Fridrich, M. Goljan, and R. Du, "Lossless data embedding—New paradigm in digital watermarking," *EURASIP J. Applied Signal Processing, (Special Issue Emerging Applications Multimedia Data Hiding)*, vol. 2002, no. 2, pp. 185–196, 2002.

[16] M.U. Celik, G. Sharma, A.M. Tekalp, and E. Saber, "Reversible data hiding," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 157–160, Sept. 22–25, 2002.

[17] B. Zhu, M.D. Swanson, and A.H. Tewfik, "A transparent authentication and distortion measurement technique for images," in *Proc. 7th IEEE Digital Signal Processing Workshop*, Loen, Norway, Sept. 1996, pp. 45–48.

[18] B. Zhu, "Coding and data hiding for multimedia," Ph. D. dissertation, Dept. of Electrical and Computer Engineering, Univ. of Minnesota, Twin Cities, MN, Dec. 1998.

[19] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proc. IEEE*, vol. 87, no. 7, pp. 1167–1180, 1999.

[20] N. Memon, P. Vora, B.-L. Yeo, and M. Yeung, "Distortion bounded authentication techniques," *Proc. SPIE, Security and Watermarking of Multimedia Contents II*, vol. 3971, pp. 164–174, 2000.

[21] C.W. Wu, "On the design of content-based multimedia authentication systems," *IEEE Trans. Multimedia*, vol. 4, pp. 385–393, Sept. 2002.

[22] M. Schneider and S.-F. Chang, "A robust content based digital signature for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, 1996, vol. 3, pp. 227–230.

[23] D.-C. Lou and J.-L. Liu, "Fault resilient and compression tolerant digital signature for image authentication," *IEEE Trans. Consumer Electron.*, vol. 46, no. 1, pp. 31–39, 2000.

[24] M.P. Queluz, "Content-based integrity protection of digital images," in *Proc. SPIE Conf. Security Watermarking Multimedia Contents*, Jan. 1999, vol. 3657, pp. 85–93.

[25] J. Dittmann, A. Steinmetz, and R. Steinmetz, "Content-based digital signature for motion pictures authentication and content-fragile watermarking," *IEEE Int. Conf. Multimedia Computing and Systems*, 1999, vol. 2, pp. 209–213.

[26] C.-T. Li, D.-C. Lou, and T.-H. Chen, "Image authentication and integrity verification via content-based watermarks and a public key cryptosystem," in *Proc. IEEE Int. Conf. Image Processing*, 2000, vol. 3, pp. 694–697.

[27] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. IEEE Int. Conf. Image Processing*, 1998, vol. 1, pp. 435–439.

[28] 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.

[29] C.-Y. Lin and S.-F. Chang, "Generating robust digital signature for image/video authentication," in *Multimedia Security Workshop ACM Multimedia 98*, Bristol, U.K., Sept. 1998. Available: http://www.ctr.columbia.edu/~cylin/publications.html.

[30] Q. Sun, S.-F. Chang, M. Kurato, and M. Suto, "A quantitative semi-fragile JPEG2000 image authentication system," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2002, vol. 2, pp. 22–25.

[31] C.-Y. Lin and S.-F. Chang, "Semi-fragile watermarking for authenticating JPEG visual content," in *Proc. SPIE Security and Watermarking of Multimedia Contents II*, San Jose, CA, Jan. 2000, pp. 140–151.

[32] R. Radhakrishnan and N. Memon, "On the security of the SARI image authentication system," in *Proc. IEEE Int. Conf. Image Processing*, 2001, vol. 3, pp. 971–974.

[33] C.-P. Wu and C.-C. J. Kuo, "Speech content authentication integrated with CELP speech coders," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME2001)*, Aug. 22–25, 2001, pp. 1009–1012.