Methods for Tamper Detection in Digital Images

Jiri Fridrich
State University of New York in Binghamton and Mission Research Corporation
fridrich@binghamton.edu

1. ABSTRACT

The purpose of this paper is to present a comprehensive overview of current steganographic techniques for tamper detection and authent ication of visual information. Fragile, semi-fragile, robust watermarks, and self-embedding are discussed as a means for detecting both malicious and inadvertent changes to digital imagery. Some attacks and security gaps are discussed.

1.1 Keywords

Authentication, tamper detection, fragile watermarks, semi-fragile watermarks, self-embedding, hybrid wate r-marks, feature authentication

2. Introduction

Image authentication using steganography is quite di f-ferent from authentication using cryptography. In cryptographic authentication, the intention is to protect the communication channel and make sure that the message received is authentic. It is typically done by appending the image hash (image digest) to the image and encrypting the result. Once the image is decrypted and stored on the hard disk, its integrity is not protected anymore. Steganography offers an interesting alternative to image integrity and authenticity problem. Because the image data is typicall very redundant, it is possible to slightly modify the image so that we can later check with the right key if the image has been modified and identify the modified portions. The integrity verification data is embedded in the image rathe than appended to it. If the image is tampered with, the embedded information will be modified thus enabling us to identify the modifications.

In the past, several techniques [1 –14] and concepts based on data hiding or steganography have been intro-duced as a means for tamper detection in digital images and for image authentication – fragile watermarks, semi-fragile watermarks, robust watermarks, and self-embedding. The visual redundancy of typical images e n-ables us to insert imperceptible additional information and make the images capable of authenticating themselves without accessing the originals. The goal is to prevent the possibility of creating a forgery that goes undetected. An example application would be a secure digital camera equipped with a watermarking chip that authenticates every image it takes before storing it on the flash card. The embedded information could be uniquely tied to the ca m-era's serial number thus creating a link between the images and the hardware that took them. Such smart images ma play an important role in detecting digital forgeries or e s-tablishing the origin of digital images.

Fragile watermarks [1 –4,10,13] are designed to detect every possible change in pixel values. They can be d e-signed to provide a very high probability of tamper detection while making it practically impossible to create a fo r-gery. However, since images are highly redundant, and their visual content is generally not modified under small perturbations, it may not be desirable to have this kind of sensitivity at least in some applications. Semi-fragile a-termarks [5,10] are moderately robust and thus provide a "softer" evaluation criterion. The value identifying the presence of the watermark (a correlation in most cases) can serve as a measure of tampering. A natural extension of the concept of a semi-fragile watermark is the robust spread spectrum watermark on medium sized blocks [7–9]. If an image feature comparable in size to the watermarking block is removed or added, the watermark in that block will no longer be present. On the other hand, typical image processing operations, such as filtering, gamma correction, or lossy compression will decrease the evidence for watermark presence more or less uniformly over all blocks. Consequently, one can distinguish malicious changes from innocent image processing operations. Such techniques [7–9,13,14] could be termed authentication of the visual content. They can be combined with fragile watermarks if the fragile watermark is inserted after the robust one. This hybrid watermark [13] combines the accuracy and precise localization of the fragile watermark with the robustness of the robust watermark.

The last category of image authentication techniques is called self-embedding. The image is embedded into itself in such a manner that it is later possible to not only detect tampered or cropped out portions of the image, but also to recover the original content. Due to very large payload requirements of self-embedding techniques, it is not poss i-ble to have a good reconstruction quality, watermark in-visibility, and robustness at the same time. Currently d e-veloped techniques [11,12] can be classified as fragile o semi-fragile watermarks with very high quality of the e-constructed image.

In this paper, we present the technical details of the above mentioned steganographic techniques for authenti-cation and tamper detection in digital images. We compare their performance, security, and outline research directions in this field. In section 3, we start with fragile watermarks, and continue with semi-fragile watermarks in Section 4. Robust watermarks and hybrid watermarks for tamper d e-tection are covered in Section 5. In Section 6, we describe algorithms for self-embedding techniques and conclude the paper in Section 7.

3. Fragile watermarks

If the inserted watermark is fragile so that any manipu-lation of pixels will disturb its integrity, one can readil detect the tampered areas by checking for presence of this fragile watermark. A very simple scheme is obtained b encrypting the seven most significant bit planes and hashing the result. This hash can then be inserted into the least significant bit plane of the image. With high pro b-ability, any change made to any bit plane will be detected. The localization properties of this simple scheme can be improved if it is applied to image blocks rather than the whole image.

One of the first fragile watermarking techniques p o-posed for detection of image tampering was based on i n-serting check-sums of gray levels determined from the seven most significant bits into the least significant bits (LSB) of pseudo-randomly selected pixels [1]. In this p a-per, we are going to describe one possible implementation of this idea. First, we choose a large numbe N that will be used for calculating the check sums. Its size directly influ-ences the probability of making a change that might go undetected. The image is then divided into 8×8 blocks, and in each block, a different pseudo-random walk through all 64 pixels is generated. Let us denote the pixels as p_1, p_2, …, p_{64}. We also generate 64 integers a_1, a_2, …, a_{64} compa-rable in size to N. The check sum S is calculated as

$$S = \sum_{i=1}^{64} a_i\, g(p_i) \bmod N ,$$

where $g(p_i)$ is the gray level of the pixel $p(i)$. It is then expressed in a binary form, encrypted, and embedded in the LSBs of the image block. Swapping pixels within one block will change the value of S because the two pixels will have different coefficients a_i. The random walk p_i and the coefficients a_i can be block dependent (using a secret key), thus making it impossible to swap entire blocks without making undetected changes. One weakness of this scheme is that it is possible to swap identically positioned blocks in two authenticated images, unless one does not make the watermark dependent on the image or at least the image order. This could be achieved for example using the robust bit extraction algorithm proposed in [16]. Anothe alternative to thwart this "collage" attack is to use ra n-domly placed pixels rather than publicly known 8×8 blocks. This may, however, somewhat negatively influence the ability to localize changes.

Yeung and Wong [2,3] proposed the following method for authentication of digital images. The process of image authentication starts with a secret key that is used to ge n-erate a key dependent binary valued function f, f: {0, 1, …, 255} → {0,1}, that maps integers from 0 to 255 to either 1 or 0. For color images, three such functions, f_R, f_G, f_B, one for each color channel, are generated. These binary func-tions are used to encode a binary logo L. The gray scales are perturbed to satisfy the following expression

$$L(i,j) = f_g(g(i,j)) \text{ for each pixel } (i,j).$$

For an RGB image, the three color channels are perturbed to obtain

$$L(i,j) = f_R(R(i,j)) \oplus f_G(G(i,j)) \oplus f_B(B(i,j)) \text{ for each pixel } (i,j),$$

where ⊕ denotes the excluded OR. Error diffusion is fur-ther employed to preserve the original colors. The image authenticity is easily verified by checking the relationship $L(i,j) = f_g(g(i,j))$ for each pixel (i,j).

There are some obvious advantages of this approach. First, the logo itself can carry some useful visual informa-tion about the image or its creator. It can also represent a particular authentication device or software. Second, b comparing the original logo with the recovered one, one can visually inspect the integrity of the image. Third, the authentication watermark is embedded not only in the LSBs of the image but somewhat deeper (± 5 gay scales). This makes it more secure and harder to remove. Fourth, the method is fast, simple, and amenable to fast and cheap hardware implementation. This makes it very appealing for still image authentication in digital cameras.

This method, however, has a serious security gap if the same logo and key are reused for multiple images. Given two images I_1 and I_2 with gray levels $g^{(1)}$ and $g^{(2)}$ water-marked with the same key and logo L, we have

$$f_g(g^{(1)}(i,j)) = L(i,j) = f_g(g^{(2)}(i,j)) \text{ for all } (i,j).$$

The last equation constitutes $M\times N$ equations for 256 un-knowns f_g. As reported in [17] only two images are needed on average to recover over 90% of the binary function f_g. Once the binary function is estimated, the logo can be easily derived. Actually, if the logo is a real image rather than a randomized picture, we can use this additional information to recover the rest of the binary function f_g. Although the situation becomes more complicated for color images, the method appears to have a serious security gap. Making the embedded information depend on the image index would not be too practical because one would have to search for the right index, which may in turn increase the complexity of the algorithm. Embedding the index in a robust manner in the image using a secret (camera) ke may alleviate this situation. As another approach, we can use the robust bit extraction algorithm [16] and make the logo and / or the binary function(s) a nontrivial function of the secret key and the image itself.

4. Semi-fragile watermarks

Another class of authentication watermarks is formed b semi-robust watermarks. Such watermarks are marginall robust and are less sensitive to pixel modifications. Thus, it is possible to use them for quantifying the degree of tamper and distinguish simple LSB shuffling from malicious changes, such as feature adding and removal. Van Schyndel et al. [4] modify the LSB of pixels by adding extended m-sequences to rows of pixels. For an $N \times N$ image, a sequence of length N is randomly shifted and added to the image rows. The phase of the sequence carries the watermark information. A simple cross-correlation is used to test for the presence of the watermark. Wolfgang and Delp [5,10] extended van Schyndel's work and improved the localization properties and robustness. They use bipolar m-sequences of –1's and 1's arranged into blocks and add them to corresponding image blocks. If $X(b)$ denotes the gray levels of the original image block b, the watermarked block $Y(b)$ is calculated as

$$Y(b) = X(b) + W(b).$$

The verification process used to test an image Z to see if the watermark is in the image is:

$$\delta(b) = Y(b) \cdot W(b) - Z(b) \cdot W(b).$$

A threshold test is then performed on the test statistic δ. If $\delta < T$, where T is a user-defined threshold, $Z(b)$ is considered genuine. Large values of T allow the toleration of changes to the marked image block $Y(b)$. If $Z(b) = Y(b)$, then $\delta = 0$.

Zhu et al. [6] propose two techniques based on spatial and frequency masking. Their watermark is guaranteed to be perceptually invisible, yet it can detect errors up to one half of the maximal allowable change in each pixel or f equency bin depending on whether frequency or spatial masking is used. The image is divided into blocks and in each block a secret random signature (a pseudo-random sequence uniformly distributed in [0,1]) is multiplied b the masking values of that block. The resulting signal depends on the image block and is added to the original block quantized using the same masking values. Errors smaller than one half of the maximal allowable change are readily detected by this scheme.

The authors apply this technique to small 8 ×8 pixel blocks. The block is DCT transformed, and the frequency masking values $M(i,j)$ for each frequency bin $P(i,j)$ are calculated using a frequency masking model. The values $M(i,j)$ are the maximal changes that do not introduce perceptible distortions. The DCT coefficients are modified to $P_S(i,j)$ according to the following expression

$$P_S(i,j) = M(i,j)\ \{\lfloor P(i,j) / M(i,j) \rfloor + r(i,j)\ \mathrm{sign}(P(i,j))\},$$

where $r(i,j)$ is a key-dependent noise signal in the interval (0,1), and $\lfloor x \rfloor$ rounds x towards zero. Since $|P(i,j) - P_S(i,j)| \le M(i,j)$, the modifications to DCT coefficients are imperceptible.

For a test image block with DCT coefficients $P_S'(i,j)$, the masking values $M'(i,j)$ are calculated. The error at (i,j) is estimated by the following equation

$$e' = P_S' - M'\ \{r\ \mathrm{sign}(P_S') + \lfloor P_S' / M' - (r-1/2)\ \mathrm{sign}(P_S') \rfloor\},$$

where all the values are evaluated at the same frequenc bin (i,j). The authors show that if the true erro e at (i,j) is smaller in absolute value than $M(i,j)/2$, and if $M'(i,j) = M(i,j)$, the estimated erro $e' = e$. It is further shown that the error estimates are fairly accurate for small distortions, such as high quality JPEG compression.

5. Authenticating the visual content

Fridrich [7,8] describes a technique in which an image is divided into medium-size blocks and a robust spread-spectrum watermark is inserted into each block. If watermarks are present in all blocks with high probability, one can be fairly confident that the image has not been tampered with in any significant manner (such as adding o removing features comparable in size to the block). If the watermark correlation is lower uniformly over all image blocks, one can deduce that some image processing ope ation was most likely applied. If one or more blocks sho very low evidence for watermark presence while othe blocks exhibit values well above the threshold, one can estimate the probability of tampering and with a high probability decide whether or not the image has been tampered with.

Kundur and Hatzinakos [9], propose a wavelet-based telltale image authentication. Because the watermark is localized both spatially and in the frequency domain, it provides spatial and frequency domain information on ho the signal was modified. For example, if certain frequencies in the image block have been untouched, they will be authenticated as credible. The image is first decomposed using the Haar transformation to the L-th level into the high frequency components $f_{k,l}(m,n)$ and the lowest resolution level $f_{a,L}$. The secret key is used to generate the subset of wavelet coefficients that will be modified. A special quantization function Q is used to assign binary values to wavelet coefficients f

$$Q_{\Delta,l}(f) = 0 \text{ if } \lfloor f/(\Delta 2^l) \rfloor \text{ is even,}$$
$$Q_{\Delta,l}(f) = 1 \text{ if } \lfloor f/(\Delta 2^l) \rfloor \text{ is odd}$$

at the quantization level l. If a wavelet coefficient $f_{k,l}(m,n)$ is chosen for watermark embedding, it is modified so that

$$Q_{\Delta,s}(f_{k,l}(m,n)) = w(i) \text{ XO } \; qkey(m,n),$$

where $w(i)$ is the i-th watermark bit and $qkey$ is a bit generated from the image and a secret key. The construction of the quantization function Q guarantees that one will never have to modify the coefficient at the level l by more than $\pm\Delta 2^l$. The watermark is extracted by evaluating the expression

$$w(i) = Q_{\Delta,s}(f'_{k,l}(m,n)) \text{ XO } \; qkey(m,n),$$

where f' is the wavelet coefficient of the potentially tampered image. The extent of tampering is evaluated using the number of correctly recovered watermark bits $w(i)$. The authors also provide an estimate of the probability that a random modification of the wavelet coefficients will go undetected. This probability is shown to decrease exponentially with the number of modified coefficients.

Schneider and Chang [14] propose a content-based signature for robust feature authentication. First an image is processed and a set of features is extracted. The result is hashed and encrypted using a public key. The encrypted information is finally embedded in the image. To authenticate an image, the embedded information is first extracted, decrypted, and compared to the hash of extracted features. Since the hash function is sensitive to every bit of its input, the feature extraction and normalization needs to be robust to achieve insensitivity to small modifications while being able to detect –large changes. The authors use image histograms on small blocks as features. Another possibilit would be to use special "robust" hash functions that are not sensitive to every input bit. The robust bit extraction algorithm [16] mentioned previously is one possible approach.

It appears that no single scheme can have both precise localization properties without being too sensitive. Indeed these two requirements are in conflict. On the other hand, it should be possible to combine a robust watermark or a feature authentication watermark with a fragile one, if the fragile watermark is embedded as the second one. The fragile watermark is usually very weak and should not influence the robust one in any significant manner. This hybrid watermark [13] can, therefore, enjoy the good properties of both watermarks. If a subtle change is made to a highly localized group of pixels, such as changing the eye color in a portrait photograph, the fragile watermark can be used to precisely localize the change. On the othe hand, a simple lossy compression or applying a filter to the image will be indicated as non-malicious tamper because the robust atermark will survive.

6. Self-embedding

The idea of self-embedding the image into itself enables not only detection of areas that have been tampered or damaged, but also recovering the missing information. The self-embedded information can be in a fragile or in a semi-fragile form. Thus, self-embedding is a means both for protecting the image content and for authentication. Because the image or its approximation needs to be embedded into itself without introducing visible artifacts, the embedded information cannot be robust. There is an obvious trade-off between the robustness of the self-embedded image and its visual quality.

Fridrich and Goljan [11] describe a self-embedding technique in which the image is first divided into blocks of 8×8 pixels. Setting the LSB of each pixel to zero, a DCT is calculated for each block. The DCT matrix is then quantized with the quantization matrix corresponding to a 50% quality JPEG. The resulting quantized matrix is then encoded using 64 bits and the code is inserted into the LSBs of a distant 8×8 block. The watermarking process on average modifies 50% of pixels by one gray level. The qualit of the reconstructed image is somewhat worse than 50% quality JPEG. If two LSBs are used for inserting the code (encoded quantized DCT coefficients) 128 bits can be used instead of 64 bits. For most blocks, this enables encoding almost all quantized DCT coefficients. Thus, the quality of the reconstructed image is roughly equivalent to a 50% JPEG compression.

The same authors [12] introduce another algorithm for self-embedding that is based on differential encoding. The differences between neighboring pixels are adjusted so that they mimic a decreased color depth approximation of the same image shifted by a third of the image dimension in a random direction. This provides a weak robustness to small noise adding while maintaining the quality of the embedded image quite reasonable (16-color approximation of the original). The embedded image gradually degrades with noise adding. The visual quality of the recovered image is still acceptable after adding noise of amplitude of 2 gray levels. However, the embedded information is lost after a 65% quality JPEG compression (the default setting in many commercial software products).

7. Conclusion

In this paper, we provide a comprehensive overview of steganographic techniques for tamper detection and authentication of digital images. The techniques are divided into several categories according to their ability to identify changes. Fragile watermarks can detect changes to every pixel and provide accurate information about the image integrity. However, it is not possible to distinguish small, innocuous changes due to common image processing operation from malicious changes, such as feature emoval or addition. Semi-fragile watermarks are more obust and allow "authentication with a degree". It is possible to set a threshold in those techniques so that images after high quality JPEG compression, or contrast/brightness adjustment will still be considered authentic to a high degree. In the third category, we put techniques that attempt to authenticate image features. Such techniques are even more robust and enable robust distinction between innocuous and malicious modifications at the expense of losing the sensitivity to small changes and sometimes the ability to localize modifications. However, it is possible to combine those watermarks with fragile atermarks if the fragile watermark is inserted as the second one. Such hybrid watermarks provide a much wider spectrum of protection against unauthorized modifications. The last category of data embedding techniques for tamper detection is called self-embedding. In those techniques, the image is embedded into itself in a judicious manner so that it is actually possible to later recover areas that have been cropped out, missing features, or identify newly added features. Self-embedding watermarks are fragile or semi-fragile and their disadvantage is that they cannot be combined with lossy compression.

One common problem of all watermarking techniques for tamper detection or authentication is that the watermark has to be image and key dependent in a non-trivial non-invertible manner to prevent creating forgeries that will go undetected. If the same key is used for authentication of multiple images, many block-based techniques will be vulnerable to a collage attack in which blocks from different images are combined. This is a serious issue if the authentication algorithm is expected to be implemented in digital cameras or surveillance video-cameras. It appears that "robust hash functions" also called robust bit extraction is the only practical way to solve this problem.

8. Acknowlegements

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under a grant number F30602-98-C-0009. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Air Force Research Laboratory, or the U. S. Government.

9. References

[1] S. Walton, “Information Authentication for a Slippery New Age”, *Dr. Dobbs Journal*, vol. 20, no. 4, pp. 18–26, Apr 1995.

[2] M. Yeung, and F. Mintzer, "An Invisible Watermarking Technique for Image Verification", *Proc. ICIP'97*, Santa Barbara, California, 1997.

[3] P. Wong, "A Watermark for Image Integrity and Ownership Verification", *Proc. IS&T PIC,* Portland, Oregon, 1998.

[4] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, “A Digital Watermark”, *Proc. of the IEEE Int. Conf. on Image Processing*, vol. 2, pp. 86–90, Austin, Texas, Nov 1994.

[5] R. B. Wolfgang and E. J. Delp, “A Watermark for Digital Images”, *Proc. IEEE Int. Conf. on Image Processing*, vol. 3, pp. 219–222, 1996.

[6] B. Zhu, M. D. Swanson, and A. Tewfik, “Transparent Robust Authentication and Distortion Measurement Technique for mages”, preprint, 1997.

[7] J. Fridrich, “Image Watermarking for Tamper Detection”, *Proc. ICIP ’98*, Chicago, Oct 1998.

[8] J. Fridrich, “Methods for Detecting Changes in Digital images”, ISPACS99, Melbourne, November 4^{th}–6^{th}, 1998.

[9] D. Kundur and D. Hatzinakos, "Towards a Telltale Watermarking Technique for Tamper Proofing", *Proc. ICIP*, Chicago, Illinois, Oct 4–7, 1998, vol 2.

[10] R. B. Wolfgang and E. J. Delp, "Fragile Watermarking Using the VW2D Watermark", *Proc. SPIE, Security and Watermarking of Multimedia Contents*, San Jose, California, Jan 25–27, 1999, pp. 204–213.

[11] J. Fridrich and M. Goljan, "Protection of Digital images Using Self-Embedding", *Symposium on Content Security and Data Hiding in Digital Media*, New Jersey Institute of Technology, May 14, 1999.

[12] J. Fridrich and M. Goljan, "Images with Self-Correcting Capabilities", *ICIP'99*, Kobe, Japan, October 25–28,1999.

[13] J. Fridrich, "A Hybrid Watermark for Tamper Detection in Digital Images", *ISSPA'99 Conf.*, Brisbane, Australia, August 22–25, 1999.

[14] M. Schneider and S .-F. Chang, “A Content-Based Approach to Image Signature Generation and Authentication”, *Proc. ICIP ’96* vol. III, pp. 227−230, 1996.

[15] M. Holliman, N. Memon, and M. M. Yeung, "On the Need for Image Dependent Keys for Watermarking", *Proc. Content Security and Data Hiding in Digital Media*, Newark, NJ, May 14, 1999.

[16] J. Fridrich, “Robust Bit Extraction From Images”, *ICMCS'99*, Florence, Italy, June 7−11, 1999.

[17] J. Fridrich and N. Memon, "Attack on a Fragile Watermarking Scheme", in preparation for *Electronic Imaging 2000, Security and Watermarking of Multimedia Contents*, San Jose, California, January, 2000.