

**IIT(ISM) DHANBAD**

**SV-based Semi-Fragile Watermarking for Image content Authentication**

**Name – Sneh Shrivastava Submitted to,**

**AN – 15JE001328(91) Dr. A K PAL**

**Introduction**

Fragile, semi-fragile, and self-recovering watermark-ing techniques have been considered as potential promising tech-niques for multimedia authentication. Fragile watermarking techniques aim to be fragile to any modifications so as to detect and localize these modifications. Semi-fragile watermarking techniques aim to resist acceptable content-preserving modifications and detect malicious content-altering modifications. Self-recovering watermarking techniques incorporate the content recovery property in either fragile or semi-fragile schemes to not only detect and localize the modifications but also recover the original content. However, self-recovering fragile techniques work well when the tampered area is not extensive and self-recovering semi-fragile techniques work well under mild or no content-preserving modifications.

Semi-fragile watermarking algorithms can be classified into spatial domain-based and transform domain-based schemes. In general, spatial domain-based schemes embed watermark into the host image by modifying a set of pixel values without causing obvious changes in its appearance. Transform domain-based schemes embed watermark into the host image by modifying transformed coefficients such as discrete cosine transform (DCT) coefficients and discrete wavelet transform (DWT) coefficients. Both schemes use extracted watermark to authenticate the digital content and localize tampered areas if possible. Transform domain-based schemes are better than spatial domain-based schemes since they tend to achieve better invisibility and more robustness. Here, we briefly review several representative semi-fragile watermarking schemes in two popular transform domains (DCT and DWT domains).

Ho and Li use the relationship of DCT coefficients in low and middle frequencies to protect the authenticity of a compressed watermarked image when the JPEG quality is higher than authors’ predefined lowest authenticable quality. Maeno et al. propose two semi-fragile authentication techniques. The first one incorpo-rates a random bias factor to the fixed decision boundary to detect malicious manipulations and keep the false alarm rate low. The second one uses a non-uniform quantization scheme to improve the encoding accuracy of relationships between paired DCT coeffi-cients and achieve higher alteration detection sensitivity. Zhou et al. propose to extract a signature from non-overlapping blocks of the original image and insert it into selected block-based wavelet coefficients. Hu and Han use the hash function to generate one watermark for classifying the intentional content modification and use the Sobel edge detector to generate the other watermark for indicating the modification location. These two watermarks are finally embedded in the middle-frequency wavelet coefficients. Zhu et al. apply the block-mean quantization strategy to embed inter-block and intra-block signatures in the finest scale of wavelet coefficients for tamper detection and localization, respectively. Yang and Sun integrate the human visual system (HVS) model in the embed-ding scheme to insert watermark by modifying vertical and hori-zontal subbands of image sub-blocks. Two measures are then used to judge whether a modification is malicious. Che et al. propose the HVS-based dynamic quantization approach to embed watermark in low-frequency wavelet coefficients. Cruz et al. propose to extract a robust signature from 16 16 non-overlapping blocks by thresholding projections onto random smooth patterns. They then employ the vector quantization method to embed this signature into the approximation sub-band of each image sub-block. Zhang et al. design a fast scheme to avoid the overflow checking of pixel intensity by adjust-ing wavelet coefficients and adopt the cumulative weighted voting method to reduce the false alarm pixels. Preda propose to embed a watermark bit in a group of randomly permuted wavelet coefficients by means of quantization. The embedded watermark is robust to mild to moderate JPEG compression by selecting appro-priate embedding parameters. Morphological operations are also used to improve detection results. Huo et al. Propose to embed general tampering and collage attack watermarks in the block-based DCT domain and detect tampered regions based on the two maps generated from two extracted watermarks.

However, all these schemes are only robust to moderate JPEG compression of higher than a 50% or 60% quality factor (QF). The false alarm rates for watermarking schemes proposed in are high under common image processing attacks. To address these shortcomings, two improved schemes have been recently proposed. Qi and Xin employ a non-traditional quantization method to embed watermark by modifying one chosen DWT approximation coefficient of each non-overlapping block. They then compute two measures to confirm the image content and localize possible tampered areas. Al-Otum embeds a random watermark bit sequence into the DWT domain using an adjusted expanded-bit multi-scale quantization index modulation (QIM) based technique. Two measures are further developed to classify the probe image as authenticated, incidentally, or maliciously attacked. However, the two measures in both schemes may fail to recognize tampered blocks if the watermarked image is also incidentally attacked by a moderate noise signal or a higher JPEG or JPEG2000 compression. Furthermore, the technique proposed in is limited in its use due to its inability to process color images as mentioned in the conclusion section.

an effective semi-fragile watermark-ing scheme that is capable of detecting various small to large con-tent modifications and localizing tampered areas even under mild to severe content-preserving modifications. It first generates secure watermark by performing the logical ‘‘xor’’ operation on content-dependent and content-independent watermarks. Here, content-dependent watermark is a robust signature extracted by SV-based features and content-independent watermark is a private-key-based random watermark. The proposed scheme then uses the adaptive quantization method to embed secure water-mark in the wavelet domain. It further utilizes a three-level authentication process to authenticate the image content and localize tampered areas. Specifically, the first-level process employs two measures (e.g., M1 and M2) derived from a binary error map to quickly classify the maliciously attacked water-marked image under mild content-preserving modifications. The second-level process employs another two measures (e.g., M3 and M4) derived from a binary strongly tampered error map to further classify the probe image as authenticated, incidentally, or mali-ciously attacked under moderate to severe content-preserving modifications. The third-level process employs the fifth measure (e.g., M5) derived from the post-processed candidate tampered area to further validate the nature of the attack and produce a clean tampered result if applicable. This scheme also possesses all the desired properties (e.g., invisibility, tamper detection, secu-rity, identification of tampered areas, oblivion with no transmis-sion of any secret information, and discrimination of incidental distortion and malicious tampering) for an effective authentication watermarking scheme.

**The Purposed Scheme:-**

The proposed scheme consists of four components: secure watermark generation, watermark embedding, watermark extrac-tion, and watermark authentication. In the following subsections, we explain each component in detail.

**Secure watermark generation:-**

Using relationships of SVs among three sub-blocks of each 4 4 non-overlapping JPEG quantized block of the original image, we generate content-dependent watermark CW that represents intrin-sic algebraic image properties to facilitate the authentication pro-cess. In order to increase the security, we also generate random content-independent watermark IW by using the Mersenne Twister algorithm and a private key k1. Secure watermark SW is finally generated by performing the logic ‘‘xor’’ operation on CW and IW. This secure watermark encodes the image content and therefore is more fragile to content altering operations and moderate to severe common image processing attacks.

The quantized image q-I is generated by modifying coefficients (i.e., Blki(x, y)) in each 8 8 block Blki of original image I to an inte-gral multiple (i.e., modified\_Blki(x, y)) of the quantization matrix Q as follows:

modified\_Blki(x, y)=round(Blki(x, y)/Q(x,y)) \* Q(x,y)

where round() represents the rounding operation, Q is the quantiza-tion matrix specified in the JPEG standard (1 6 x, y 6 8), and opera-tions / and are the element-wise division and multiplication, respectively.

For each 4 4 quantized block q\_Blki, the following operations are performed:

1. Divide the block into 2 2 sub-blocks to obtain subblocki\_0, sub-blocki\_1, subblocki\_2, and subblocki\_3 in the raster scan order.
2. Apply SVD on subblocki\_1 to obtain three matrices Ui\_1, Si\_1, and Vi\_1, where subblocki\_1 = Ui\_1 Si\_1 VTi\_1. Similarly, apply SVD on subblocki\_2 and subblocki\_3 to obtain two sets of three matri-ces, i.e., Ui\_2, Si\_2, and Vi\_2, and Ui\_3, Si\_3, and Vi\_3, respectively.
3. Generate a watermark bit based on the relationship of the most notable and stable SVs of subblocki\_1, subblocki\_2, and subblocki\_3, which correspond to three values, i.e., Si\_1(1, 1), Si\_2(1, 1), and Si\_3(1, 1). The rules for generating content-dependent watermark bit CWi are:

3.1 Generate bit B1 using the relationship between Si\_1(1, 1) and Si\_2(1, 1) by:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| B1 = | 1 | if Si 1 >= Si 2 |  |  |
| 0 | Otherwise |  |  |

3.2. Similarly, generate bit B2 using the relationship between Si\_2(1, 1) and Si\_3(1, 1) and generate bit B3 using the relation-ship between Si\_1(1, 1) and Si\_3(1, 1), respectively.

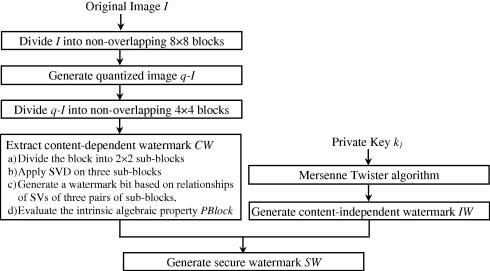
3.3 Generate CWi by applying a series of ‘‘xor’’ operations:

CWi = xor(xor(B1,B2), B3)

3.4 Evaluate the intrinsic algebraic image property of the block by:

PBlocki = B1 + B2 + B3

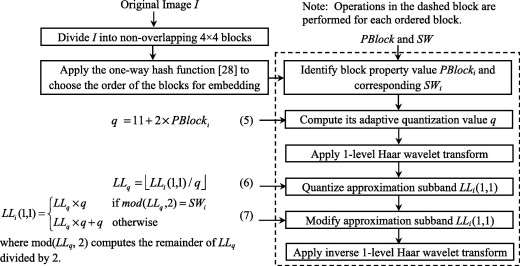
Chosen SVs to generate CW due to its stability in terms of a small perturbation, its invariant algebraic and geometric proper-ties, and its compact representation of algebraic properties of an image. The relationship of SVs between sub-block pairs mea-sures the luminant difference of each pair. The PBlocki value for each quantized block measures the luminant change patterns of each pair, which is used to automatically determine the block’s quantization step. The subblocki\_0 is not used to generate the watermark. Instead, it is used for embedding watermark to ensure the robustness of generating CW and SW.



**Watermark Embedding:-**

We divide the original image into non-overlapping 4 4 blocks and embed secure watermark SW in the wavelet domain of each unique randomly chosen 4 4 block.

We choose the wavelet domain as the embedding media mainly due to its excellent spatial-frequency localization and its compati-bility with the JPEG2000 coding standard. Specifically, we utilize the parity of the quantized value at the approximation subband to embed watermark. To ensure the watermark’s invisibility and increase robustness against common image processing attacks, we choose the upper-left value of the approximation subband (e.g., LLi(1, 1)), which corresponds to subblocki\_0, to embed a water-mark bit. The embedding strategy is to ensure that the parity of the modified LLi(1, 1) is consistent with the embedding bit. It should be noted that the bigger the q is, the bigger the changes, consequently, the worse the quality of the watermarked image, and the stronger the robustness. In our system, q is adaptively chosen for each block Bi based on PBlocki. A larger q value is assigned to a block with a higher PBlocki value. The value of q can be 11, 13, 15, or 17 for four PBlocki values (i.e., 0, 1, 2, and 3), respectively. These four quantiza-tion values are determined based on the tradeoff among invisibil-ity, robustness, and fragileness. They have been shown to be effective when using either of them as the fixed quantizer in the proposed system.



**Watermark Extraction:-**

The watermark extraction process is similar to the watermark embedding process except that it uses the parity of the quantized upper-left value of the approximation subband of each non-overlapping 4 4 block to extract the watermark bit. The upper-left value LL(1, 1)0 of each block is quantized by the adaptive quantizer q calculated by as follows:

LL0 = round(LL(1,1)’/q)

The watermarked bit EWi for each block is extracted as mod(LLq0, 2).

**Watermark Aunthentication:-**

we perform the first-level authentication using mea-sures M1 and M2 computed from the binary error map ErrorMap. This map is generated by mapping the absolute difference between extracted and regenerated secure watermarks (e.g., |EWi SWi0|) onto its corresponding 4 4 block with 0’s and 1’s indicating match and mismatch, respectively. In other words, any pixel with the value of 1’s is an error pixel. In the proposed system, we clas-sify tampered error pixels in a 3 3 window in ErrorMap into two categories: strongly and mildly tampered error pixels. An error pixel is strongly tampered if at least five of its eight neighbors are error pixels; and an error pixel is mildly tampered if at most four of its eight neighbors are error pixels. We define two authen-tication measures by:

M1= No. of error pixels in ErrorMap/No of pixels in error map

M2= No. of strongly tempred error pixels in Error Map/No of tampred error pixel in error map

Here, M1 measures the similarity between extracted and regener-ated secure watermarks and M2 measures the clustering level of tampered error pixels. M2 is set to 0’s if the count of tampered error pixels is zero. This authentication process aims to quickly identify tampered regions for the probe image undergoing content-altering modifications together with mild content-preserving modifications. We apply median filtering on ErrorMap when M1 is less than or equal to Tmedian to accommodate the small amount of distortions caused by mild modifications. This filtering keeps clustered error pixels intact and makes scat-tered mildly tampered and isolated error pixels disappear. As a result, the small malicious attack leads to a larger M2 value due to the removal of mildly distorted error pixels.

Third, we perform the second-level authentication using mea-sures M3 and M4 computed from the binary error map STErrorMap, which is generated by marking strongly tampered error pixels in ErrorMap as 1’s. We extract connected components in STErrorMap and define two other authentication measures as follows:

M3=No of error pixels in STErrorMap/No of connected components in STErrorMap

M4=std(size of all the connected components

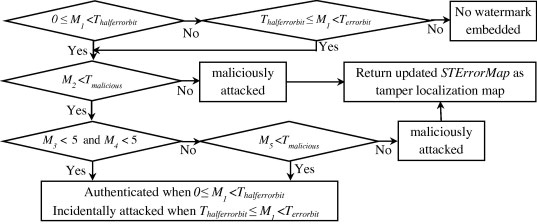
M3 measures the average size of connected components and M4 measures the size variation of all connected components. This authentication process aims to distinguish between tampered regions caused by content-altering modifications and tampered regions caused by moderate content-preserving modifications. It is triggered when M2 is smaller than Tmalicious, which indicates that there is either a significant amount of mildly tampered error pixels (caused by moderate content-preserving modifications) or a small amount of strongly tampered error pixels (caused by minor content-altering modifications). Small M3 and M4 values reflect that connected components are randomly spread out, relatively small, and similar in size. As a result, we conclude that tampered regions are caused by moderate content-preserving modifications and the probe image therefore undergoes incidental attacks.

Fourth, we perform the third-level authentication using mea-sure M5 computed from the processed STErrorMap and the pro-cessed binary error map MTErrorMap, which is generated by marking mildly tampered error pixels in ErrorMap as 1’s. Specifically, we remove the component(s) in STErrorMap and the mildly tampered error pixels in MTErrorMap whose distance to the centroid of the largest connected component is larger than an adaptively computed threshold SizeC (e.g., square root of the size of the largest connected component). The processed STErrorMap and MTErrorMap contain tampered error pixels that are near the largest connected component in STErrorMap. Measure M5 re-estimates the clustering level of tampered error pixels around the largest connected component area and is com-puted as follows:

M5=No of error pixels in processed STErrorMap/No of error pixels in processes STErrorMap and MTErrorMap

This authentication process aims to identify tampered regions when the probe image undergoes content-altering modifications together with moderate to severe content-preserving modifications. It is triggered when M3 or M4 is large, which indicates that a relatively large component corresponding to malicious distortion may exist. The post-processing operation removes the effect from content-preserving modifications and keeps the most important distortion area intact. Consequently, M5 measures the clustering level of tampered error pixels in the estimated localized area. In this way, a larger M5 indicates the presence of the maliciously tampered area.

Finally, we design a quantitative method to decide the authenticity of the probe image based on the three-level five authentication measures. The flow chart of this process is summarized in Fig. The involved thresholds, i.e., Terrorbit, Thalferrorbit, and Tmalicious, are determined based on the predefined false negative probability of 10 6. Interested readers may refer to [22] for detailed derivation of these values. In the proposed system, we set Terrorbit as 0.4837, Thalferrorbit as 0.2419, and Tmalicious as 0.6085.



**Conclusions**

We present an effective SV-based semi-fragile watermarking scheme for image content authentication with tamper localization. The contributions of the proposed scheme are:

Utilizing relationships of SVs among three sub-blocks of each 4 4 non-overlapping JPEG quantized block to extract content-dependent watermark in both watermark embedding and authentication processes. This extracted watermark is able to capture the changes in SV-based relationships that likely cor-respond to content changes in non-embedding sub-blocks.

Generating secure watermark by performing the logical ‘‘xor’’ operation on SV-based content-dependent and private-key-based content-independent watermark. This secure watermark encodes image content and therefore is more fragile to content-altering operations and moderate to severe content-preserving operations.

Merging relationships of SVs among three sub-blocks of each 4 4 non-overlapping JPEG quantized block to choose its adap-tive quantizer q for both embedding and extraction processes. Such variation reduces the possibility of misclassification and improves the quality of the watermarked image.

Applying the adaptive quantization method to embed secure watermark in the wavelet domain so that a majority of image distortions, which cause the intensity shift by a value larger than a half of q or cause image content to change, can be detected in the authentication process.

Defining a three-level authentication process involving five measures to quantitatively detect the authenticity of the probe image and prove tampering, with M1 measuring the similarity between extracted and embedded watermarks, M2 measuring the clustering level of tampered error pixels, M3 measuring the average size of connected components formed by strongly tampered error pixels, M4 measuring the variation in the sizes of connected components, and M5 measuring the clustering level of tampered error pixels around the potential distorted area.

Using five authentication measures derived from three binary maps (e.g., ErrorMap, STErrorMap, and MTErrorMap) to compen-sate possible misclassification and ensure that the proposed scheme is capable of capturing distortions and localizing tam-pered areas when a moderate noise signal or a higher compres-sion is also involved.

Extensive experimental results show that the proposed scheme successfully distinguishes malicious attacks from non-malicious tampering of image content. It also accurately localizes maliciously tampered regions even under moderate to severe JPEG compressions. The proposed scheme is more robust to various acceptable content-preserving operations and more fragile to mali-cious distortions than five semi-fragile watermarking schemes and its two variant schemes. It can be easily extended to color images.