12-2011

A General Attack Method for Steganography Removal Using Pseudo-CFA Re-interpolation

Pradhumna Shrestha
University of Nebraska Lincoln

Michael Hempel
University of Nebraska-Lincoln, mhempel2@unl.edu

Tao Ma
University of Nebraska Lincoln, tma@unlnotes.unl.edu

Dongming Peng
University of Nebraska-Lincoln, dpeng2@unl.edu

Hamid Sharif
University of Nebraska-Lincoln, hsharif@unl.edu

A General Attack Method for Steganography Removal Using Pseudo-CFA Re-interpolation

Pradhumna Lal Shrestha, Michael Hempel, Tao Ma, Dongming Peng, Hamid Sharif

Computer and Electronics Engineering Department
University of Nebraska – Lincoln, Omaha, NE 68182
{plshrestha, mhempel, tma, dpeng, hsharif}@unlnotes.unl.edu

Abstract: Watermarking and steganography are two of the most researched topics in multimedia forensics. However, easy availability of tools and technology has made their misuse a serious concern. To counteract this development some effective tools are necessary to remove malicious steganography. In this work, we introduce a novel watermark attack method which can destroy hidden information embedded in images based on the principle of re-interpolation of Color Filter Array (CFA) artifacts. In digital cameras and scanners, CFA filters are used to acquire low-resolution physical color channel information and produce a high-quality image by subsequent interpolation. We propose to emulate a similar scheme in which image information is removed and reconstructed through interpolation and thereby destroy any hidden information without impairing the visual quality. Most importantly, our presented method is general and does not assume any knowledge of the used watermarking methods, the hidden message or the host image.

Keywords: Multimedia; Watermark Embedding; Watermark Attack; Color Filter Array; Re-interpolation

I. INTRODUCTION

With advances in Internet connectivity, computing capabilities and easy availability of software, online multimedia content can easily be pirated, illegally distributed, and abused for malicious activities. Verifying the authenticity and protecting the ownership of multimedia has become a challenge. To counteract this, watermarking has emerged as one of the most powerful and well researched tools, which facilitates content protection and data integrity.

Watermarking has been successfully applied to all forms of multimedia [1-12]. However, due to ease of availability of watermarking tools [8], it has also been widely misused for steganography, hiding information within other data, predominantly multimedia files. With these tools, malicious groups can carry out secret communication with potentially devastating results. Informants and industrial espionage operatives often release sensitive information such as governmental, corporate, or trade secrets, by embedding them within multimedia files and posting those multimedia carriers online. Hence, it is of utmost importance to have an efficient tool that is able to remove steganographic information from multimedia content. However, any such removal operation on multimedia data should not reduce the quality of the content. When talking about watermarks or hidden information within this paper, we are referring to malicious information whose exchange should be prevented. We will limit our scope to image-based watermarking, which is one of the most prevalent forms of multimedia data on the Internet.

A lot of work has been done in the area of image watermarking attacks [13-19]. Watermarking attack algorithms can be broadly classified into signal processing attacks, noise attacks, geometric attacks and statistical attacks. Signal processing attacks include low pass filtering, histogram equalization, image compression, smoothing/blurring, sharpening, and any other prevalent "image correction" techniques. Noise attacks consist of adding small amounts of noise to the image. Geometric attacks include all forms of affine transforms and cropping attacks. Most modern watermarking algorithms are claimed to be resilient against the first three types of attacks [9-12]. Statistical attacks, which use properties of the image and watermarking methods, have been found to be efficient to destroy watermarks. However, none of the statistical attack methods available have been generalized and are claimed to be effective against only very specific watermarking algorithms.

In [13], the authors proposed destroying hidden information based on sparse decomposition. They treat the host image as the transmitted signal and the embedded watermark as the noise added to the signal and then use sparse analysis to estimate the image. However their process is computationally intensive and uses a set of images for training. Also it is only applicable to a specific type of watermarking algorithm based on discrete wavelet transform.

The authors of [14] proposed a new version of sensitivity attack for watermark removal. But the attack is designed for spread spectrum methods of watermarking and is essentially linked with the optimum detector of the decoding system, thus assuming clear knowledge of the specific watermarking algorithm used. Moreover, it requires information about the binary output of the detector.

In [15], the authors proposed attacking quantization-based watermarking schemes. The removal is achieved by changing enough components of the watermarked signal by half of the estimated quantization step. However the narrow scope of the application of the algorithm renders it

inadequate to be broadly applicable in real-world scenarios.

The work reported in [16] claims that even though the power spectral density of the watermark signal and the host signal are the same globally, they may not be the same locally and the attack estimates the watermark based on this difference. They propose to attack watermarks that do not satisfy this power spectrum condition on a local level by applying a Wiener attack on a block-by-block level. Furthermore it is assumed that the attacker knows the power spectral density of the host image and the watermark, which is virtually impossible to properly estimate in practice.

In [18], the authors first estimate the Eigen-image energy of the watermark signal and the watermarked image and use these values to reconstruct the host image while removing the watermark in the process. But they have used special binary random independently and identically distributed watermarks, which represent only a very small fraction of watermarks in reality.

In [19], the authors proposed counterfeiting attacks against block-wise independent watermarking schemes in which they forge the embedded content of a watermarked image into another image. But the work is more concerned with watermark security rather than watermark removal.

In summary, despite of the remarkable work done in the area of watermark attacks and removal, no low-complexity method that would systematically destroy embedded messages in multimedia content has yet been proposed. The published works suffer from one or more of the flaws listed below.

(i). They are not general and target only specific types of watermarking schemes.
(ii). They assume or require some information about watermarking schemes, watermark and host image which are generally not available to the removal process.
(iii). They are more concerned with cryptographic aspects of watermarking and watermark security rather than destroying the watermark to make it irrecoverable.

In this paper, we propose a general method of watermark removal in which as much as 75% of the host image pixels are systematically removed from the image and the removed pixels are reconstructed by re-interpolation, which approximates the host image to such a degree that the visual quality of the host image is maintained. This process produces significant watermark removal, as shown by our results.

The rest of the paper is arranged as follows. Section II will describe the operating process of the algorithm. Section III will outline our algorithm in detail. Section IV will provide the theoretical proof for the method. Section V will describe the experiment performed. Section VI will present our results demonstrating the effectiveness of our method and discuss its implications. Finally, we conclude our work in section VII.

II. System Model

Our proposed watermark removal algorithm is based on CFA filtering and interpolation, most commonly used in digital cameras and scanners. These devices use light sensitive elements to measure light intensity, combined with CFA filters that pass only one of the three primary colors (red, green and blue), depending on the filter. In a 2x2 matrix of light sensors, a common element arrangement is for 1xred, 2xgreen, 1xblue as shown in figure 1. This results in a color channel resolution decrease of up to 75 %. However, CFA-based devices restore image quality and fidelity through interpolation. We adopt this principle and create a virtual CFA filtering and re-interpolation environment, in which significant information from the watermarked image is discarded and subsequently the image quality is restored through re-interpolation using information from all color channels.

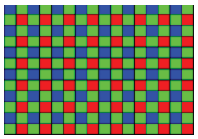

Figure 1: Color Filter Array showing colors passing through respective pixels

This method is applicable to both grayscale and color images. If the watermarked image is grayscale, it is simply first converted to an RGB representation.

The CFA's filtering method is used as the pixel removal scheme. A re-interpolation scheme is utilized to then restore the lost information. However, any filtering and a corresponding re-interpolation scheme can theoretically be used for this purpose. Since the principle of CFA filtering revolves around color images, this attack method will perform particularly well in the case of color images, especially if only one of the channels is watermarked. Considering that the majority of digital images captured by cameras are color images, this benefit of our proposed method is a significant factor in removing watermarks.

III. Algorithm

The algorithm utilizes the principle of CFA filtering and interpolation, allowing it to discard up to 75% of information from a particular channel and then restore it through re-interpolation to the original image's quality. The removal of pixels in this case represents a watermark desynchronization attack. The details of the proposed algorithm are shown below.

Step 1. If the image is grayscale, convert it into a color image. This step can be skipped, if the image is color.

$$if\ \left(image_{type} \neq RGB\right)$$

$$I_{Color}(x,y,S) = I(x,y), for\ all\ x,y\ and\ S \quad (1)$$

The content of the grayscale image is copied into all three color channels *S* to *colorize* the image.

Step 2. Apply filtering to remove pixels from all three channels.

$$I_S(x,y) = \begin{cases} I(x,y,S), & f_S(x,y) = 1 \\ 0, & otherwise \end{cases}$$
$$for\ all\ x,y\ and\ S \quad (2)$$

The CFA filter is a binomial filter which either passes without scaling or blocks the corresponding color intensity for any particular channel. It should be noted that at each location *(x,y)* only one of the color channel values is retained by the filter, which is necessary for re-interpolation in the following stages.

Step 3. For each channel estimate the removed intensity values using neighborhood information.

$$J_S(x,y) = \begin{cases} I_S(x,y), & f_S(x,y) = 1 \\ \sum_u \sum_v \alpha(u,v) I_S(x+u, y+v), & otherwise \end{cases}$$
$$for\ all\ x,y\ and\ S \quad (3)$$

For each location of each color channel, if the intensity is available, the value is retained. If the intensity was filtered out, the value is estimated by using the neighborhood pixel intensities that were passed by the CFA. In other words, the estimated value of the blocked intensity is now a function of the neighboring pixels' intensities. The weight of the neighbors is determined by the coefficient α. In this step, we have assumed the function to be a linear interpolator.

Step 4. Convert the image back to grayscale. Skip this step if the input image is a colored.

$$J(x,y) = \sum_S \beta_S J_S(x,y)\ \ for\ all\ x,y\ and\ \beta \quad (4)$$

where β_S is the weight given to channel S.

Inherent in this process is a tradeoff between retaining image quality and the effectiveness of the watermark removal approach. A better re-interpolation and restoring process that completely restores the image results in the highest possible image quality but at the same time has no effect on watermark. On the other hand, a lack of image reconstruction will best remove the watermark, but the image quality will be very poor. Therefore, we allow our approach to be adaptive by iteratively manipulating the coefficients α and β, making sure the resulting image satisfies some threshold defined acceptable to the end user.

IV. Theoretical Analysis

Since our method can be applied to any watermarked image independent of the utilized watermarking algorithm, it is impossible to specify to what extent the embedded watermark is destroyed. It will entirely depend on the decoding algorithm of an individual watermarking method and, to some extent, the statistics of the image under investigation as well.

However we can empirically establish some quantitative measure to present the validity of our approach. In this section, we address the issue of improving visual quality of the image by using the virtual CFA filtering and re-interpolation process.

Let $I(x,y)$ be the watermarked image. Then, the filtered image is,

$$I_S(x,y) = \begin{cases} I(x,y); & if\ f_S(x,y) = 1 \\ 0; & otherwise \end{cases}$$
$$= I(x,y) * f_S(x,y) \quad (5)$$

where *S* represents one of the three color channels (R, G or B) and $f_S(x,y)$ represents the matrix that determines which pixels are removed from channel S. After re-interpolation,

$$J_S(x,y) = \begin{cases} I_S(x,y);\ if\ f_S(x,y) = 1 \\ \sum_u \sum_v \alpha(u,v) I_S(x+u, y+v); otherwise \end{cases}$$
$$= \begin{cases} I(x,y); if\ f_S(x,y) = 1 \\ \sum_u \sum_v \alpha(u,v) I_S(x+u, y+v); otherwise \end{cases} \quad (6)$$

where $\alpha(u,v)$ are the interpolating function coefficients. After conversion to grayscale,

$$J(x,y) = 0.2989\, J_R(x,y) + 0.5870\, J_G(x,y) + 0.1140\, J_B(x,y) \quad (7)$$

Since color channels have high interchannel correlation [20], due to the inclusion of information from all the three color channels, it is evident that $J(x,y)$ is able to restore the loss of information during filtering process to some extent. Therefore, it is clear that,

$$PSNR\big(I(x,y), J(x,y)\big) > PSNR\ (I(x,y), J_S(x,y)) \quad (8)$$

This clearly demonstrates that using the process of re-interpolation to estimate the removed pixels and then combining the output of the three virtual channels will significantly improve the PSNR.

Since the process of re-interpolation cannot replicate the image precisely, we can see that the watermark is effectively removed.

$$I(x,y) \neq J(x,y) \quad (9)$$

V. EXPERIMENT PERFORMED

Several test images were used for experimentation in this work. These images were first watermarked using the algorithms published in [1, 21, 22]. These algorithms were chosen with regards to the variety in their approaches to embed watermark data in images, as well as their wide acceptance as representative watermarking algorithms by the research community.

In [1], the authors have proposed a technique for embedding messages in multimedia. This paper revolutionized the area of watermarking by first proposing watermark signal spreading across the spectrum of the image, in the same fashion as is used by CDMA spread spectrum communication in which the desired signal is spread across a much wider carrier spectrum. This made the watermark imperceptible and robust. Though many modifications have been proposed since it was published, most of the watermark algorithms, even some modern ones, are still based on this method. Hence, this algorithm was an important candidate for our tests. We embedded an 8x8 watermark image into our set of 1024x1024 host images.

In [21], the authors have presented an algorithm to embed data by quantizing the so-called wavelet tree of images. Basically, the image is wavelet transformed and wavelet coefficients are spread and grouped together to form trees. The values of the coefficients are quantized based on a desired error threshold. This algorithm incorporates both spreading and quantization principles and hence was chosen as a representative prospect for our tests. For each 512x512 benchmark image, an 8x8 watermark has been embedded.

In [22], the authors have presented a remarkably different method by hiding message data in the histogram of the host images. Unlike most watermarking algorithms, in which encoding and decoding is basically a function of how pixels are related to their neighborhood, this method is based on the intensity counts of the image pixels. Since our method is based on destroying neighborhood relation by changing pixel values by a small amount and in turn impacting the histogram of the image, it was of interest to determine the effectiveness of our method on this algorithm. Since this algorithm has a lower data hiding capacity, a 5x5 watermark was embedded.

The embedded images were then attacked using our proposed Pseudo-CFA filtering and re-interpolation attack. The attacked carrier images are then passed though their respective decoders in an attempt to extract the watermark. After being attacked, the extracted watermarks were compared with the originally embedded watermark. The results of the comparison are presented in section VI.

VI. RESULTS AND DISCUSSION

Figures 2-7 illustrate an example taken from our test image set: Figure 2 shows the original 512x512 host image intended to be watermarked. Figure 3 shows the watermark message that is to be embedded into the host image. Figure 4 and 5 respectively show the watermarked image using the algorithm in [21] and the extracted watermark before the watermarked image has been attacked. It can be seen that the extracted watermark is a perfect replica of the embedded signal. Figures 6 and 7 show the image after Pseudo-CFA filtering and re-interpolation and the watermark signal extracted from it. It

Figure 2: Original 512x512 image

Figure 4: Watermarked image using algorithm in [21]

Figure 6: Watermarked Image after CFA filtering and re-interpolation

Figure 3: 8x8 Watermark Signal, magnified

Figure 5: Extracted Watermark, magnified

Figure 7: Extracted Watermark after attack, magnified

Table 1. Statistics of recovered watermark (in format of PSNR (dB) /Correlation/Bit Error Rate) and impact on image after attacking

Image	Watermarked using [1]			Watermarked using [21]			Watermarked using [22]		
	No attack	Pseudo-CFA filtering	PSNR of attacked image (dB)	No attack	Pseudo-CFA filtering	PSNR of attacked image(dB)	No attack	Pseudo-CFA filtering	PSNR of attacked image(dB)
Sample: Baboon	∞, 1, 0%	9.81, 0.38, 30%	36.49	∞, 1, 0%	8.50, 0.10, 50%	25.59	∞, 1, 0%	8.09, 0.12, 35%	26.68
Sample: Peppers	∞, 1, 0%	14.31, 0.85, 7%	46.59	∞, 1, 0%	8.51, 0.18, 44%	29.98	∞, 1, 0%	6.73, 0.16, 48%	32.88
Sample: Barbara	∞, 1, 0%	14.97, 0.81, 12%	37.60	30.07, 0.99, 0.2%	8.33, 0.14, 48%	27.25	∞, 1, 0%	9.06, 0.22, 32%	28.60
Sample: Lena	∞, 1, 0%	12.08, 0.64, 15%	47.92	∞, 1, 0%	8.03, 0.09, 45%	31.97	∞, 1, 0%	11.59, 0.36, 20%	39.92
Image Set Average	∞, 1, 0%	11.67, 0.63, 19%	41.33	∞, 1, 0%	8.27, 0.12, 47%	28.55	∞, 1, 0%	8.79, 0.21, 41%	31.87

is evident that the extracted watermark bears no resemblance to the originally embedded watermark. Also it should be noted that all the host images (original, watermarked and attacked) look similar and hence the image integrity has not been sacrificed throughout this process. However, we have successfully destroyed the hidden signal embedded within an image. For sake of space, all the images have not been presented in this paper.

In table 1, the effect of attacking our benchmark host image set of 80 images, each watermarked using the algorithm in [1, 21, 22], have been presented. The table provides the quality of the extracted watermark before and after our Pseudo-CFA re-interpolation attack, in terms of the watermark PSNR, the watermark correlation, and the watermark BER. Furthermore, the impact of the operation on the attacked image is also presented in terms of its PSNR with respect to the watermarked image. Though the effect of the attacks on the images varies from image to image, the statistics consistently show that enough bit errors have occurred for the watermark to have become irretrievable. These fluctuations are expected since the embedding algorithm will impact different images in different fashions depending on the texture of the image. Similarly, the correlation between the attacked and original watermark, as well as the watermark PSNR, has been significantly degraded and the extracted watermark bears no resemblance to the embedded watermark.

Most importantly, our algorithm did not depend on, nor utilize, any information about the image, watermark signal or the embedding procedure while attacking the image. This clearly demonstrates that our method is widely applicable for the fully-blind removal of malicious information hidden within images.

VII. CONCLUSION

Watermarking is a very important tool in the area of multimedia distribution as it helps to maintain copyright protection and content integrity. However, it can also easily be abused, with often severe consequences. Therefore, the removal of malicious information hidden in images using watermarking or steganography is an important aspect in multimedia security. Any such algorithm, however, needs to be designed to attack the watermark without compromising the quality of the image itself.

Recently, several watermark attacking methods have been published in literature. However, the major problem with those methods is that they are not generic, are only effective against very specific watermarking methods, and oftentimes require knowledge about the hidden information content, the embedding algorithm, or the original image, which are typically inaccessible to the watermark removal algorithm.

In this work, we have presented a comprehensive watermark attack method based on the principle of a Pseudo-Color Filter Array filtering and re-interpolation process, which is used to successfully remove hidden information from the images without impairing their visual quality. The method is effective in removing watermarks from color images as well as grayscale images. We have shown that the method is of low complexity and applicable to a wide range of watermarking approaches, yet very effective in removing hidden information as clearly demonstrated by our results.

REFERENCES

[1]. Ingemar J. Cox, Kilian, F. Thomson Leighton, and Talal Shamoon, "Secure Spread Spectrum Watermarking for Multimedia", Transactions on Image Processing, vol. 6, no. 12, December, 1997

[2]. B. Chen and G.W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding", ISIT 2000

[3]. A Koz and A.A. Alatan, "Oblivious Spatio-Temporal Watermarking of Digital Video by Exploiting the Human Visual System", IEEE Circuits and Systems for Video Technology 2008

[4]. M. Noorkami and R.M. Mersereau, "Digital Video Watermarking in P-Frames With Controlled Video Bit-Rate Increase", IEEE Transactions on Information Forensics and Security, 2008

[5]. Pik Wah Chan, M.R Lyu and R.T. Chin, "A novel scheme for hybrid digital video watermarking: approach, evaluation and

experimentation", IEEE Transactions on Circuits and Systems for Video Technology, 2005

[6]. Kang Xiangui; Yang Rui and Huang Jiwu, "Geometric Invariant Audio Watermarking Based on an LCM Feature", IEEE Transactions on Multimedia, 2011

[7]. Wang Liang, S. Emmanuel and M.S Kankanhalli, "EMD and psychoacoustic model based watermarking for audio", IEEE International Conference on Multimedia and Expo, 2010

[8]. Thi Hoang Ngan Le, Kim Hung Nguyen and Hoai Bac Le, "Literature Survey on Image Watermarking Tools, Watermark Attacks, and Benchmarking Tools", 2010 Second International Conferences on Advances in Multimedia

[9]. Wo Yan, Han Guo-Qiang, Zhang Jian-Wei and Zhang Bo, "Image Authentication Resilient to Translation, Rotation and Scaling", International Conference on Machine Learning and Cybernetics, 2006

[10]. S. Pholsomboon and S. Vongpradhip, "Rotation, scale, and translation resilient digital watermarking based on complex exponential function", IEEE Region 10 Conference, 2004

[11]. Xie Gui and Shen Hong, "Robust wavelet-based blind image watermarking against geometrical attacks", IEEE International Conference on Multimedia and Expo, 200

[12]. Zhao Yu-xin, Liu Guang-jie, Dai Yue-wei and Wang Zhi-quan , "A RST-Resilient Watermarking Scheme Based on Invariant Features", International IEEE Conference on Signal-Image Technologies and Internet-Based System, 2007

[13]. Su-min Yang, Zheng-bao Zhang , Wen-bo Wang and Hong-yan Han, "A novel watermark blind attacking algorithm based on spase analysis and ICA", International Conference on Multimedia Technology, 2010

[14]. P. Comesana, L. Perez-Freire, F. Perez-Gonzalez, "Blind newton sensitivity attack", IEEE Proceedings Information Security, 2006

[15]. Mohamed F. Mansour and Ahrned H. Tewfik, "Attacks On Quatnization-Based Watermarking Schemes", International Symposium on Signal Processing and its Applications, 2003

[16]. Hiroshi Ito, "Local Weiner Attack for Additive Watermarks", IEEE International Symposium on Consumer Electronics, 2009

[17]. Chunlin Song, Sud Sudirman, Madjid Merabti and David Llewellyn-Jones, "Analysis of Digital Image Watermark Attacks", IEEE CCNC 2010

[18]. Te-Cheng Hsu, Wen-Shyong Hsieh and Tung-Shih Su, "A New Watermark Attacking Method Based on Eigen-Image Energy", Intelligent Information Hiding and Multimedia Signal Processing, 2008

[19]. Matthew Holliman and Nasir Memon , "Counterfeiting Attacks on Oblivious Block-wise Independent Invisible Watermarking Schemes", IEEE Transactions on Image Processing, vol. 9, no. 3, March 2000

[20]. Bahadir K. Gunturk, Yucel Altunbasak, and Russell M. Mersereau, "Color Plane Interpolation Using Alternating Projections", IEEE Transactions on Image Processing, Vol. 11, No. 9, September, 2002

[21]. S. Wang and Y. Lin, "Wavelet Tree Quantization for Copyright Protection Watermarking", IEEE Transactions on Image Processing vol.13, Feb. 2004

[22]. Shijun Xiang, Hyoung Joong Kim and Jiwu Huang, 'Invariant Image Watermarking Based on Statistical Features in the Low-Frequency Domain", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 6, June 2008