Countering Ambiguity Attacks Against Digital Image Watermarking Schemes

Randima Hettiarachchi [1] and Chandana Gamage [2]
[1] Central Bank of Sri Lanka, Colombo, Sri Lanka
[2] Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

Abstract—**The ambiguity attack, or invertibility attack, was discovered in 1998 as a simple but a powerful attack against digital watermarking schemes. Identifying the potential threat of this attack, a number of ambiguity attack resistant watermarking schemes have been proposed in the past literature. However, later on some of these schemes were proven to be failing under the ambiguity attack. In this paper, we study the concept of the ambiguity attack, evaluate different ambiguity attack resistant watermarking schemes and the scenarios under which they fail to provide necessary security against ambiguity attacks. We employ the secure spread spectrum method as the underlying watermarking scheme to implement and evaluate these ambiguity attack resistant watermarking schemes.**

I. Introduction

Early days, duplication of art work was quite complicated and required a high level of expertise for the counterfeit to look like the original. In modern times, in the digital era it is possible for almost anyone to duplicate or manipulate digital data and not lose data quality.

Artists in the past creatively signed their paintings with a brush to claim copyrights. The solution for artists of today is digital watermarking. Digital watermarking is the process of embedding ownership information into a digital object in a way that is difficul to remove or alter.

It was Craver et al. [1], who initiated the concept of inversion attacks that aim to fin a forged watermark and a corresponding fake original from watermarked work. His argument was that when Alice, the original owner of the image proves her ownership of the watermarked image I' by producing her original image I_a and her watermark W_a, an attacker, Bob will also prove his ownership of the watermarked image I' by producing his original image I_b and his watermark W_b.

Moreover both of them will be able to prove that by embedding their watermarks W_a and W_b in their original images I_a and I_b respectively, they both can produce the same watermarked image I'.

As a solution to ownership dispute and resolution during ambiguities, Craver et al. proposed to impose the invertibility requirement on the embedding/detection scheme. This initiated a series of work that aimed at devising non-invertible schemes that were mainly built on conventional (embedding/detection) techniques such as spread spectrum and discrete wavelet transformation techniques.

R. Hettiarachchi is with the Central Bank of Sri Lanka, Colombo, Sri Lanka. Email: randishi@yahoo.com

C. Gamage is with the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. Email: chandag@uom.lk

In order to achieve non-invertibility, Craver et al. [1] proposed to include one-way (trapdoor) functions along the path of watermark generation, so that it is not possible to reverse the process. In other words, the watermark is generated by applying a one-way hash function on the original work. An attacker would have to break the underlying one-way hash function to launch an invertibility attack.

The loopholes of the scheme proposed by Craver et al. were discussed in some subsequent papers [2], [6]. Ramkumar et al. [2] give an algorithm to break the scheme proposed by Craver et al. [1] as well as an improved scheme. Later on Adelsbach et al. [5] introduced a more general form of the inversion attacks proposed by Craver et al., called ambiguity attacks, where an attacker is not required to generate a fake original to prove his ownership of the watermarked image, rather a forged watermark and the key to generate the same.

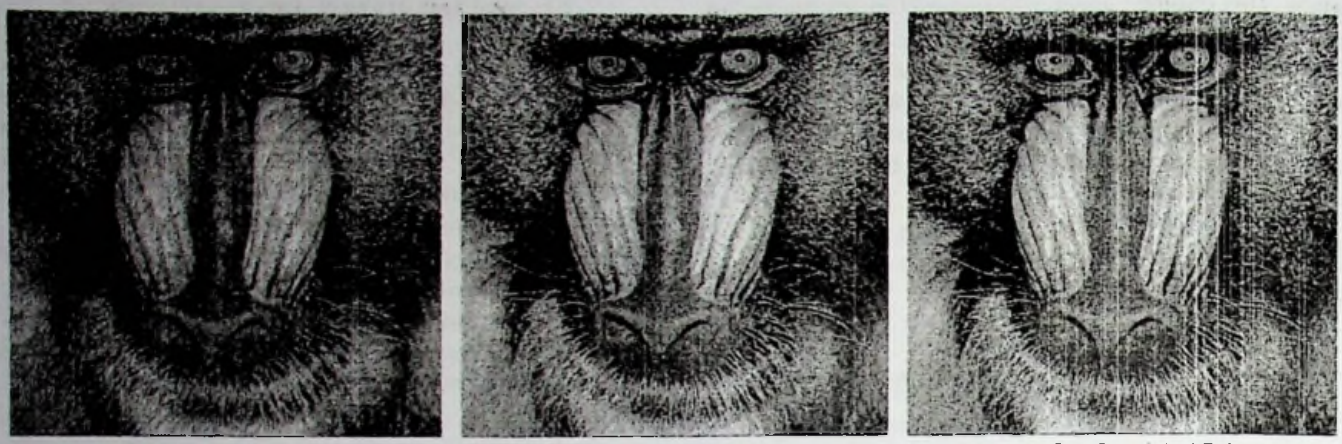
As a solution to ambiguity attack, Adelsbach et al. [5] proposed a provably secure non-invertible scheme with the help of trusted third party. However, according to Adelsbach et al. [5], this scheme suffers from another form of ambiguity attack called interactive ambiguity attack, as the trusted third party cannot distinguish between a true author and an attacker.

Moreover involvement of trusted third party in watermarking scheme should be avoided as much as possible since trusting a trusted third party itself creates problems. Taking all these facts into consideration, Li et al. proposed the firs stand-alone provably secure non-invertible watermarking scheme in "On the Possibility of Non-Invertible Watermarking Schemes" [3].

In this scheme, the watermark is statistically independent of the original image and the original image is not used during the watermark detection process. Therefore, Li et al. argue that the drawbacks of the scheme proposed by Craver et al. [1], which were brought out by Ramkumar et al. in [2], are being addressed by their stand alone watermarking scheme in [3].

It is interesting to know whether the watermarking schemes discussed above, have addressed the threat of ambiguity attacks on watermarking schemes in full scale. Unfortunately, some popular watermarking schemes proposed in [7], [9], [11] were later claimed to be susceptible to ambiguity attack in [8], [10], [12] respectively.

In this paper, we explore the models of ambiguity attack against watermarking schemes and the evolution of ambiguity attack resistant watermarking schemes to counteract those attacks. We broadly analyze the ambiguity attack resistant watermarking schemes proposed by Craver et al. [1] and Li et al. [3] and evaluate their resistance to ambiguity attacks.



(a) Original Baboon image (b) Watermarked Baboon image (c) Constructed fake for original Baboon image

Fig. 1. Baboon images under standard watermarking

TABLE I
VISUAL QUALITY VALIDATION VIA CORRELATION MEASUREMENTS BETWEEN ORIGINAL IMAGE, WATERMARKED IMAGE AND FAKE ORIGINAL IMAGE

Image	$\frac{\text{OriginalImage}}{\text{WatermarkedImage}}$	$\frac{\text{FakeOriginalImage}}{\text{WatermarkedImage}}$
Baboon	0.9969	0.9969
Lena	0.9956	0.9955
Peppers	0.9956	0.9955
Ruwanweliseya	0.9976	0.9978
Sandakadapahana	0.9972	0.9969
Parakramabahu	0.9958	0.9958

B. Watermarking Scheme Based on Original Image

Unlike in the standard scheme, in this scheme the watermark is generated by using pseudo random number generator by giving image hash as the seed. MD5 hash algorithm was used in order to calculate the image hash and the firs 16 bits out of the 128-bit long image hash value was used as the key. In this setting, the watermark embedding and detection processes are similar to that of the standard watermarking scheme.

However, the ambiguity attack against the watermarking scheme based on the original image is quite different from the attack explained in the above section. In this case, first the watermarked image I' was modifie by embedding a some random noise such that the difference in PSNR between the watermarked image I' and the modifie image I_m was around 30 - 40 dB, keeping the visual similarity unchanged. We didn't employ the histogram modificatio as proposed by Ramkumar, because it resulted in detection correlations much far from the expected value. Next, the difference between I' in transform domain and I_m was calculated as follows:

$$I_{dt} = \frac{v_i' - v_i}{\alpha v_i} \tag{21}$$

Then a watermark was generated by taking the modifie image I_m as the initial fake original image using equations 22 and eqn23.

$$K = H(I_m) \tag{22}$$

TABLE II
SUMMARY OF THE AMBIGUITY ATTACK AGAINST WATERMARKING SCHEME BASED ON ORIGINAL IMAGE

Image	Correlation at Detection $(\frac{I}{I'})$	Correlation at Detection $(\frac{I_m}{I'})$	PSNR$(\frac{I'}{I_m})$
Baboon	0.9802	0.7248	37.24 dB
Lena	0.9965	0.9575	34.95 dB
Peppers	0.9965	0.9612	33.64 dB
Ruwanweliseya	0.9905	0.6573	32.04 dB
Sandakadapahana	0.9945	0.2096	32.54 dB
Parakramabahu	0.9831	0.3881	37.10 dB

TABLE III
SUMMARIZED RESULTS OF THE WATERMARK DETECTION PROCESS OF WATERMARKING SCHEME BASED ON SECRET KEY

Image	Correlation at Detection $(I' \cdot W)$	Correlation at Detection $(I \cdot W)$
Baboon	36.3308	18.9916
Lena	41.9985	5.1658
Peppers	8.2684	-23.7809
Ruwanweliseya	-10.5586	-12.9018
Sandakadapahana	15.5067	14.8588
Parakramabahu	24.4402	16.3352

$$W' = f(K) \tag{23}$$

As an attacker, our objective is to fin a fake original, which would result in a W' that has a reasonably high correlation with I_{dt}. Therefore, a number of watermarks were generated by tweaking the 1 -2 LSBs of I_m and correlated with I_{dt}. Furthermore, a new I_{dt} was calculated each time I_m was tweaked to obtain a new fake original.

The data in table II illustrates the results of the ambiguity attack against the watermarking scheme based on original image. By observing figur 2 and table II, it is evident that successful ambiguity attacks can be performed on the watermarking scheme based on original image, maintaining the visual similarity.

(a) Original Peppers image (b) Watermarked Peppers image (c) Constructed fake for original Peppers image

Fig. 2. Peppers images under watermarking scheme based on original image

TABLE IV
SUMMARIZED RESULTS OF AMBIGUITY ATTACKS AGAINST WATERMARKING SCHEME BASED ON SECRET KEY

Image	Correlation at Detection $(I' \cdot W)$	Correlation at Detection $(I' \cdot W')$
Baboon	36.3308	21.2396
Lena	41.9985	45.2032
Peppers	8.2684	30.6084
Ruwanweliseya	-10.5586	-9.9136
Sandakadapahana	15.5067	6.6645
Parakramabahu	24.4402	19.7067

C. Watermarking Scheme based on a Secret Key

In our setting, a 16-bit long secret key was used in order to generate the watermark in the watermarking scheme based on a secret key. Next, this watermark was embedded in to the original image's largest 1000 coefficien by using equation 24. Note that I should be converted into normal distribution $N(0,1)$ before embedding the watermark.

$$v_i' = v_i + \alpha w_i \tag{24}$$

The major difference in this scheme compared to the previous two scheme is basically the detection process, because in this case the correlation between the watermarked image I' and the extracted watermark W is measured by calculating the inner product of those two. Both I' and W were converted into normal distribution $N(0,1)$ before computing $I' \cdot W$. By looking at the equation 25, it is clear that the expected value of $I' \cdot W$ should be positive as $W \cdot W$ results in a positive expected value.

If the image I' is watermarked using W:

$$I' \cdot W = I \cdot W + W \cdot W \tag{25}$$

If the image I' is not watermarked using W:

$$I' \cdot W = I \cdot W \tag{26}$$

Results presented in table III clearly show that the scheme proposed by Li et al. satisfie the suggestions made by Ramkumar et al. in [3]. That is, from table III, it is evident that the correlation between watermarked image and the watermark $(I' \cdot W)$ is higher than the correlation between the original image and the watermark $(I \cdot W)$.

In order to implement the ambiguity attack against this scheme, watermarks were generated by varying the key and the presence of each watermark in the watermarked image I' was tested by using the equation 15. The data in table IV illustrates the results of successful ambiguity attacks against the watermarking scheme based on secret key.

From the results in table IV, it is evident that in some cases an attacker is capable of detecting his forged watermark with a higher correlation than the correlation, the original owner gets for his true watermark. Therefore, when the false alarm rate of the underlying scheme is high, a successful ambiguity attack can be performed on the scheme based on secret key.

IV. REMARKS AND FUTURE WORK

The results presented in the previous section clearly show that the well known ambiguity attack resistant watermarking schemes also fail to provide the necessary security against ambiguity attacks under some scenarios. Therefore, there is more room for further research on how can we improve the existing schemes and how can we come up with a new scheme, which foils the drawbacks of the existing schemes and provide much more robustness against ambiguity attack.

During the analysis of the watermarking scheme based on secret key, it was revealed that the false alarm of the underlying scheme is high. Hence, it is worthwhile to explore ways of reducing the false alarm of the underlying scheme.

V. CONCLUSION

Many watermarking schemes have been proposed in past literature, which claimed to be ambiguity attack resistant. In this paper, we have evaluated the performance of two popular ambiguity attack resistant watermarking schemes: watermarking scheme based on original image proposed by Craver et al. [1] and watermarking scheme based on secret key proposed by Li et al. [3].

We have shown that it is possible to break the watermarking scheme based on original image by the attack proposed by Ramkumar et al. [2]. Furthermore, we can conclude that a successful ambiguity attack can be performed on the scheme based on secret key, when the false alarm rate of the underlying scheme is high.

References

[1] Craver S., Memon N., and Yeo B. *Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications.* In IEEE Journal on Selected Areas of Communications, 1998. Online: http://vada1.skku.ac.kr/ClassInfo/microsystem/multimedia/watermarking/JB00231998010012.pdf

[2] Ramkumar M. and Akansu A. *Image Watermarks and Counterfeit Attacks: Some Problems and Solutions.* In Proceedings of the Symposium on Content Security and Data Hiding in Digital Media, pages 102-112, 1999. Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.9438&rep=rep1&type=pdf

[3] Li Q. and Chang E. *On the Possibility of Non-Invertible Watermarking Schemes.* In Proceedings of the Information Hiding Workshop, volume 3200 of LNCS, pages 13-24, 2004. Online: http://profs.sci.univr.it/ giaco/download/Watermarking-Obfuscation/non-invertible%20watermarking.pdf

[4] Li Q. and Memon N. *Practical Security of Non-InvertibleWatermarking Schemes.* In Proceedings of the IEEE Internaltional Conference on Image Processing, 2007. Online: http://isis.poly.edu/ qiming/publications/icip07.pdf

[5] Adelsbach A., Katzenbeisser S., and Veith H. *Watermarking Schemes Provably Secure Against Copy and Ambiguity Attacks.* In Proceedings of the 3rd ACM Workshop on Digital Rights Management, pages 111-119, 2003.

[6] Cox I., Kilian J., Leighton T., and Shamoon T. *Secure Spread Spectrum Watermarking for Images, Audio and Video.* In Proceedings of the IEEE International Conference on Image Processing, volume 3, pages 243-246, 1996.

[7] Liu J., Lou D., Chang M., and Tso H. *A Robust Watermarking Scheme using Self-reference Image.* In Computer Standards & Interfaces, 2006. Online: http://163.13.127.161/cht/courses/Yen/05Fall/project/paper/p2.pdf

[8] Ting G., Goi B., and Heng S. *Attacks on a Robust Watermarking Scheme based on Self-reference Image.* In Computer Standards & Interfaces, 2008.

[9] Ganic E. and Eskicioglu A. *Robust DWT-SVD Domain Image Watermarking: Embedding Data in All Frequencies.* In Proceedings of the International Multimedia Conference - Workshop on Multimedia and Security, 2004. Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70.7656&rep=rep1&type=pdf

[10] Ting G. *Ambiguity Attacks on the Ganic-Eskicioglu Robust DWT-SVD Image Watermarking Scheme.* In Proceedings of the International Conference on Information Security and Cryptology (ICISC), 2005.

[11] Lu C., Sun S., Hsu C., and Chang P. *Media Hash-dependent Image Watermarking Resilient Against Both Geometric Attacks and Estimation Attacks Based on False Positive-Oriented Detection,* Technical Report, 2005. Online: http://www.iis.sinica.edu.tw/page/library/TechReport/tr2005/tr05002.pdf

[12] Lu C. and Yu C. *On the Security of Mesh-Based Media Hash-Dependent Watermarking Against Protocol Attacks,.* Online: http://www.iis.sinica.edu.tw/papers/lcs/1748-F.pdf