SECURE IMAGE HASHING VIA MINIMUM DISTORTION ESTIMATION

Chao-Yung Hsu,[1,2] *Chun-Shien Lu,*[2,*] *and Soo-Chang Pei*[1]

[1]Graduate Institute of Communication Eng., National Taiwan University, Taipei, Taiwan, ROC
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC

ABSTRACT

Security is still a relatively unexplored issue in image hashing. In this paper, we address this problem and present a new metric, called minimum distortion estimation (MDE), by demonstrating its appropriateness over entropy for secure hashing evaluation. We investigate the relationship between bit changing rate and content distortion via statistical analyses of MDE. This provides a guideline for our design of a new secure image hashing method. The feasibility of our method is further demonstrated via comparisons with two known image hashing methods.

Keywords: Distortion, Entropy, Image hashing, Robustness, Security

1. INTRODUCTION

Satisfaction of both security and robustness are of paramount importance for a media hashing scheme to be feasible. So far, several image hashing methods were developed [1, 2, 4, 9]. However, we find that their common disadvantage is the limited robustness against geometric distortions. In view of this fact, our previous work [3] addressed this challenging issue and presented a robust mesh-based image hashing scheme achieving robustness against extensive geometric distortions (e.g., standard benchmarks like Stirmark3.1 and Stirmark4.0 [6]). In [3], the feature points are extracted using Harris detector in the wavelet domain to form meshes through Delaunay tessellation. For each mesh, a kind of energy difference strategy, a commonly adopted in the media hashing community, is used for hash generation. In [7], the authors proposed to generate image hash based on scale-invariant feature transform (SIFT) [5], which has been extensively used in many applications due to its promising invariance to geometric transforms.

On the other hand, security in media hashing is also important but is still a relatively unexplored issue. In [8], Swaminathan *et al*. addressed the security problem of media hashing from the differential entropy point of view. More specifically, they argue that a hash with high differential entropy can bring high security. However, we argue that differential entropy is indeed not a reliable indicator for measuring security of a media hashing method since it can be increased intentionally. More specifically, the maximum entropy can be achieved if a hash sequence is XORed with a random sequence that exhibits the maximum entropy. This point will be demonstrated in Sec. 2.1.

In this paper, we focus on the security issue of image hashing. We first study the inappropriateness of entropy in evaluating the security of image hashing methods. Then, we present a new metric, called minimum distortion estimation, and provide statistical analyses to demonstrate its feasibility in secure image hashing. Based on this new metric, a new secure and robust image hashing scheme is proposed. Comparisons with two known methods are provided to verify the feasibility of our method.

2. PROPOSED SECURE IMAGE HASHING METHOD

In this section, we first study how to evaluate the security of image hashing methods by presenting a security metric based on minimum distortion estimation. Then, we propose a new image hashing scheme with security guaranteed with maximization of the minimum distortion. As opposed to our previous work [3], in this paper we exploit scale-invariant feature transform (SIFT) for extracting feature points from which robust hashes are generated. Due to limited space, the robustness part of proposed method is omitted here and only secure hashing is discussed.

2.1. Security Metric: Minimum Distortion Estimation

Differential entropy has been used as a metric for measuring the security of image hashing [7, 8]. Here, we first study to find that it is not a suitable metric for indicating the security of media hashing. Then, we present a new security metric, termed Minimum Distortion Estimation (MDE), which states that the cost an attacker needs to pay will be significant. In other words, when the hash of an image is modified

*Corresponding Author: Dr. C. S. Lu (lcs@iis.sinica.edu.tw)

to successfully defeat an image hashing scheme, the cost will be severe degradation of image quality.

Let H_1 and H_2 be two n-bits hash sequences with a distribution of each bit given as:

$$P(H_j(i) = 1) = p_j,\ i = 1, 2, ..., n;\ j = 1, 2. \quad (1)$$

The probability mass function (pmf) of a hash sequence can be calculated as:

$$P(H_j = h_j) = p_j^k (1-p_j)^{n-k},\ 0 \le k \le n;\ h_j \in [0, 2^{n-1}]. \quad (2)$$

Let H_3 be the third sequence, which is generated by XORing H_1 and H_2. The pmf of H_3 can be derived as:

$$\begin{aligned} P(H_3 = h_3) &= [p_1 p_2 + (1-p_1)(1-p_2)]^k \cdot \\ &\quad [(1-p_1)p_2 + (1-p_2)p_1]^{n-k}. \quad (3) \end{aligned}$$

When the entropy of H_2 is maximum (*i.e.*, $p_2 = \frac{1}{2}$), the entropy of H_3 in Eq. (3) becomes $P(H_3 = h_3) = \frac{1}{2^n}$, which achieves the maximum value. From the above derivations, we can find that the entropy of a hash sequence can be intentionally increased by XORing a high entropy sequence. Therefore, we verify that entropy is not a good indicator for security evaluation.

In fact, from an adversary's viewpoint, the goal will be properly modify the media content and change its hash values accordingly such that the modified media content can escape from copy detection. Under this circumstance, the modifications will introduce content distortions to affect visual quality. Therefore, it is interesting to know the trade-off between content distortions and the degree of hash changes. Even an image hash can be successfully counterfeited or modified, if the introduced distortion is significant, then such an attack will be meaningless. In view of this, we argue that the content distortion is a feasible indicator for security evaluation of an image hashing method.

2.2. Minimum Distortion via Energy Difference

In this paper, minimum distortion estimation is proposed as a metric to measure the security of an image hashing scheme. Since energy difference in the transform domain is a common way for hash generation, the minimum distortion discussed here will also be defined for this kind of methods. However, the concept of minimum distortion can be extended to other kinds of schemes.

Let V_A and V_B be a pair of i.i.d. feature vectors (such as wavelet coefficients) with length n from which hash bits are produced. Assume V_A and V_B are uniformly distributed in $[0, c]$. In general, a hash sequence H can be generated by the rule of energy difference as follows:

$$H(i) = \begin{cases} 1, & V_A(i) - V_B(i) \ge 0 \\ 0, & V_A(i) - V_B(i) < 0 \end{cases},\ i = 1, 2, ..., n. \quad (4)$$

Hence, the minimum distortion, mD, required for modifying a hash sequence in order to escape from detection is relevant to the problem of how many feature pairs defined in Eq. (4) needing to be inverted.

As shown in Fig. 1, the pmf of absolute energy difference is near a right-triangular distribution. This is reasonable because the pmf $f_D(d)$ of two i.i.d. random variables' energy difference d is a shift version of the convolution result between pmfs of the two random variables. The convolution result normally centralizes to zero such that most of the energy differences are small enough. Thus, an adversary can exploit this characteristic to easily modify the hash of an image by inverting the relationship corresponding to small energy differences without being detected, and keep its quality acceptable. We call this kind of attack, minimum distortion attack, which is related to the security of an image hashing method.

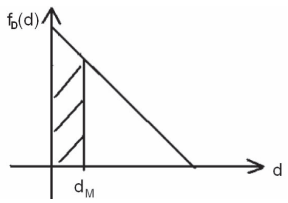

Fig. 1. The density of energy difference between two i.i.d. uniform random variables.

From the adversary's viewpoint, the process of changing (flipping) a hash bit is equivalent to inverting the relationship in Eq. (4). One way he/she can do in order to reduce the introduced distortions is to change the features as:

$$\begin{aligned} V_A'(i) &= \frac{V_A(i) + V_B(i)}{2} \pm \epsilon, \\ V_B'(i) &= \frac{V_A(i) + V_B(i)}{2} \mp \epsilon, \end{aligned} \quad (5)$$

where V_A' and V_B' denote modified feature vectors corresponding to V_A and V_B, and ϵ is a small value. For computational simplification, ϵ is omitted in the following derivations. The distortion, measured in terms of absolute energy difference, for each pair of features can be calculated as:

$$\begin{aligned} D_A(i) &= |V_A(i) - V_A'(i)| = |\frac{V_A(i) - V_B(i)}{2}|, \\ D_B(i) &= |V_B(i) - V_B'(i)| = |\frac{V_B(i) - V_A(i)}{2}|. \end{aligned} \quad (6)$$

Similarly, based on Eq. (6) the distortion required to change the i-th hash bit is calculated as:

$$D(i) = \sqrt{D_A^2(i) + D_B^2(i)} = \frac{|V_A(i) - V_B(i)|}{\sqrt{2}}. \quad (7)$$

The probability distribution function (pdf) of $D(i)$ can be derived from the i.i.d. random variables, V_A and V_B, which

are uniformly distributed in $[0, c]$ as:

$$F_D(d) = P(D(i) \leq d) = \frac{2d}{c} - (\frac{d}{c})^2,\ d \in [0, c]. \quad (8)$$

The probability mass function of $D(i)$ can be derived by differentiating $F_D(d)$ as:

$$f_D(d) = \frac{\partial}{\partial d} F_D(d) = \frac{2}{c}(1 - \frac{d}{c}). \quad (9)$$

To successfully attack an image with minimum quality distortion, a smart attacker will pick the hash bits, which are determined from small energy differences, to modify. Therefore, the minimum distortion for a successful attack can be derived from Eqs. (8) and (9) as:

$$mD = \frac{1}{F_D(d_M)} \sum_{d=0}^{d_M} d^2 f_D(d), \quad (10)$$

where $d_M \leq c$ represents the maximum cost in distortion that an attacker needs to pay for one bit changing and $\frac{1}{F_D(d_M)}$ denotes the percentage of hash bits that are modified.

2.3. Secure Hash Generation

After SIFT is executed, several feature points, each of which owns a 128-dimensional feature vector, are generated. In our robust hash design, each 128-dimensional SIFT feature vector is equally divided in to s sub-vectors, V_j $(1 \leq j \leq s)$, of length $\frac{128}{s}$, and a hash sequence of length $N = \frac{64}{s}$ is generated accordingly. For each V_j, it is divided into two parts of length $\frac{64}{s}$, V_{j1} and V_{j2}, for hash bit generation via energy difference. Next, the sequence of energy difference, ED_j, for sub-vector V_j can be calculated as:

$$ED_j(i) = |V_{j1}(i) - V_{j2}(i)|,\ 1 \leq i \leq N;\ 1 \leq j \leq s. \quad (11)$$

In order to achieve security without being easily attacked by paying few distortion, the hash bit must be generated from significant energy difference. Based on this observation, a new feature vector is defined as:

$$K(i) = V_{k1}(i) - V_{k2}(i),\ k = \arg_j \max_{j=1}^{s} \{ED_j(i)\}. \quad (12)$$

According to Eq. (12), a robust and secure image hash sequence is finally generated as:

$$H(i) = \begin{cases} 1, K(i) \geq 0 \\ 0, K(i) < 0 \end{cases},\ 1 \leq i \leq N. \quad (13)$$

Based on Eq. (13), it is found that the distortion that an adversary needs to pay for changing the i-th hash bit is:

$$D(i) = |K(i)|. \quad (14)$$

In order to calculate the minimum distortion that an attacker needs to pay for a successful attack, the minimum distortion is characterized by the pdf of $D(\cdot)$, $i.e.$,

$$\begin{aligned} F_D(d) &= P(D(i) \leq d) = P(|K(i)| \leq d) \\ &= P(Max\{V_1(i), V_2(i), ..., V_s(i)\} \leq d) \\ &= P(V_1(i) \leq d, V_2(i) \leq d, ..., V_s(i) \leq d). \end{aligned} \quad (15)$$

Since the elements of random variables V_j's are independent, Eq.(15) can be rewritten according to Eq. (8) as:

$$\begin{aligned} F_D(d) &= P(V_1(i) \leq d) \cdot P(V_2(i) \leq d) \cdot ... \cdot P(V_s(i) \leq d) \\ &= (\frac{2d}{c} - (\frac{d}{c})^2)^s. \end{aligned} \quad (16)$$

The probability mass function of Eq. (16) can be derived as:

$$\begin{aligned} f_D(d) = P(D(i) = d) &= \frac{\partial}{\partial d} F_D(d) \\ &= \frac{2}{c}(1 - \frac{d}{c})(\frac{2d}{c} - (\frac{d}{c})^2)^{s-1}. \end{aligned} \quad (17)$$

Substituting Eqs. (16) and (17) into Eq. (10), the minimum distortion required to defeat our method can be obtained.

Fig. 2 depicts the results of Eq. (17) with various values of s. It can be observed by comparing Fig. 1 and Fig. 2 that in our method larger s makes the resultant minimum distortion larger. This superiority will be demonstrated in Sec. 3 by comparing with two known methods.

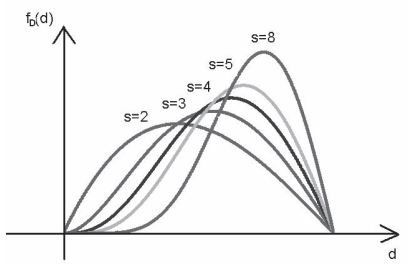

Fig. 2. The pmfs of distortions under various values of s in our method.

3. COMPARISONS AND EXPERIMENTS

In this section, experiments and comparisons with two known methods [7, 8] were conducted. We evaluate the security of these image hashing methods in terms of minimum distortion by changing r hash bits, which denotes the target bit changing rate $(0 < r \leq 1)$.

3.1. Security of our method

As we have described in Sec. 2.3 regarding our proposed method, the pdf of energy difference and target hash bit

changing rate r can be related as:

$$P(D(i) \leq d_M) = (\frac{2d_M}{c} - (\frac{d_M}{c})^2)^s = r, \qquad (18)$$

where the upper bound of distortion that an attacker is willing to pay for flipping a hash bit is d_M. After d_M is derived, the minimum distortion defined in Eq. (10) is obtained as the average distortion within $[0, d_M]$.

Fig. 3 shows the bit changing rate vs. minimum distortion under different values of s. It can be observed that larger s leads to larger distortion, which is the guideline for our design of image hash described in Sec. 2.3.

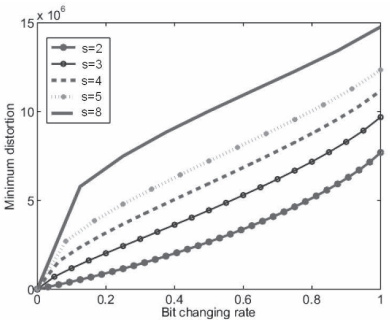

Fig. 3. Bit changing rate vs. minimum distortion under different values of s in our method.

3.2. Security of Roy and Sun's method

Roy and Sun's method [7] also propose to extract hashes based on SIFT. The major difference distinguishing our scheme from their method is that they used a random threshold to generate hash sequences from SIFT feature vectors. The solid curve in Fig. 4 shows the result of minimum distortion vs. bit changing rate in [7]. It is found that the distortion required to pay for [7] is significantly less than that for our method under the same r. The reason is that the random threshold makes the distribution of distortion be a near Gaussian, as previously discussed in Sec. 2.2.

3.3. Security of Swaminathan *et al.*'s method

Swaminathan *et al.*'s scheme [8] uses Fourier coefficients as the features to generate hash sequences, where a hash value is obtained from random linear combination of the sum of coefficients, followed by quantization and gray-coding. The relationship between bit changing rate, r, and minimum distortion, $mD_{Wu}(r)$, can be derived as:

$$mD_{Wu}(r) = t_w(\frac{Q \cdot N_w \cdot r}{t_w})^2, \qquad (19)$$

where Q is the quantization interval, N_w is the total number of hash bits, and t_w is the number of coefficients that are used for image hashing. The dash curve in Fig. 4 shows the result of minimum distortion vs. bit changing rate in [8]. It is found that the distortion required to pay for [8] is significantly less than that for our method under the same r.

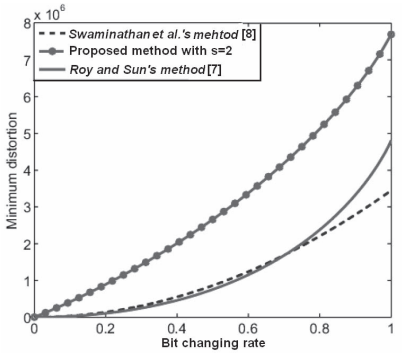

Fig. 4. Comparisons of minimum distortion vs. bit changing rate among our method and [7, 8]. Note that the worst result in our method is used for comparisons.

4. CONCLUSION

We address the issue of security regarding image hashing by presenting a security metric, termed minimum distortion evaluation. We statistically analyze how to achieve secure image hashing. Based on this, a secure and robust image hashing scheme is proposed. In the future, more image hashing methods will be included for security comparisons.

5. REFERENCES

[1] J. Fridrich, "Visual Hash for Oblivious Watermarking," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, 2000.

[2] F. Lefebvre and B. Macq, "RASH : RAdon Soft Hash algorithm," *Proc. European Signal Processing Conference*, Toulouse, France, 2002.

[3] C. S. Lu and C. Y. Hsu, "Geometric Distortion-Resilient Image Hashing Scheme and Its Applications on Copy Detection and Authentication," *ACM Multimedia Systems Journal, special issue on Multimedia and Security*, Vol. 11, No. 2, pp. 159-173, 2005.

[4] C. Y. Lin and S. F. Chang, "A Robust Image Authentication Method Distinguishing JPEG Compression from Malicious Manipulation," *IEEE Trans. on Circuits and Systems for Video Tech.,* Vol. 11, No. 2, pp. 153-168, 2001.

[5] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Intl. Journal of Computer Vision,* 2004.

[6] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on Copyright Marking Systems," *Proc. Int. Workshop on Information Hiding,* LNCS 1575, pp. 219-239, 1998.

[7] S. Roy and Q. Sun, "Robust Hash for Detecting and Localizing Image Tampering," *Proc. IEEE Int. Conf. Image Processing,* 2007.

[8] A. Swaminathan, Y. Mao, and M. Wu, "Robust and secure image hashing," *IEEE Trans. on Information Forensics and Security*, pp. 215-230, 2006.

[9] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin, "Robust Image Hashing," *Proc. IEEE Int. Conf. Image Processing,* 2000.