Journal of Network Intelligence ©2021 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **6**, Number **4**, November **2021**

Information Leakage Source Tracing Based on Robust Image Watermarking Against Mobile Instant Messaging

Xiao-Dan Jiang[1,2], Shi-Ming Yu[2], De-Feng He[2], Yong-Liang Liu[3] and Zhe-Ming Lu*[4]

[1]College of Electrical and Information Engineering
Quzhou University
Quzhou, 324000, P. R. China
xdjiang_515@126.com

[2]College of Information Engineering
Zhejiang University of Technology
Hangzhou, 310023, P. R. China
ysm@zjut.edu.cn

[3]Alibaba Group
Hangzhou, 311121, P. R. China
yongliang.lyl@alibaba-inc.com

[4]School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310013, China
*Corresponding Author: zheminglu@zju.edu.cn

Received April 2021;Revised July 2021
(Communicated by Zhe-Ming Lu)

Abstract. *Leaking sensitive information via transmitting screenshots over mobile instant messaging has become a serious and urgent problem to be solved. This paper presents a source tracking scheme for detecting information leakers based on robust image watermarking. When the mobile APP detects a screenshot operation, the APP embeds the user's ID robustly in the screenshot image. Then the screenshot image may be transmitted by the user via instant messaging software to other users. Thus, by detecting the watermark in the screenshot image can detect the leaker. Considering the main attacks of instant messaging software on the image are JPEG compression and scaling operations, this paper designs a robust image watermarking scheme in the DCT domain named scaling-resilience quantization index modulation with template patterns (SRQIMTP). Experimental results demonstrate the effectiveness and superiority of our scheme compared with other schemes.*

Keywords: Image watermarking, Robust watermarks, Information leakage source tracking, Mobile instant messaging.

1. **Introduction.** The severity of the data leakage problem is increasing year by year. According to the Risk Based Security (RBS) report in the Q3 quarter of 2020, there were 2,953 publicly disclosed data breaches from January to September 2020, which is 49% of the number of incidents in the same period in 2019 (6021); however, The number of leaked data records involved was 36.107 billion, an increase of 332.21% compared with the leaked records of the same period in 2019 (8.354 billion), a record high. In general, the global data breach situation in 2020 is not optimistic.

In recent years, data leakage incidents are not only related to hacker attacks and improper server configuration, but internal personnel leakage has also become an important cause. For example, in April 2020, a Zhejiang Rural Commercial Bank was punished for leaking user information in violation of regulations; in May of the same year, Jiangsu police cracked a case involving internal employees selling personal financial information of the bank, involving more than 50,000 records; in August, an investigation found YTO's internal employees colluded with external criminals, resulting in the leakage of 400,000 pieces of personal information. In addition, the personal information collected during the epidemic has been frequently leaked due to internal personnel actively sending out data. For example, in January 2020, the personal information of more than 7,000 returnees from Wuhan was leaked, including citizens' ID numbers and telephone numbers, Specific home addresses, train information, etc.; in July of the same year, the list of more than 6,000 people in Qingdao Jiaozhou Central Hospital in Shandong Province was leaked, involving detailed personal information of patients. The value of data and the ability to monetize data have led to more rampant data black and gray production, and the dark web is active in all kinds of transactions that leak data every day. Leakage tracing is the key to root out the problems of black and gray production and data leakage. On the one hand, traceability can help companies understand the weaknesses of internal security management and technical measures, and on the other hand, it can act as a psychological deterrent to leakers who commit criminal acts, thereby effectively reducing the occurrence of similar incidents. However, in the face of data leakage incidents in environments such as the dark web or public networks, in most cases it is impossible to accurately trace the source-who leaked it? Where did it leak? When did it leak?

With the development of the Internet and smart phones, the issue of information leakage has been more and more serious. Through the instant messaging software in smart phones, secret or sensitive information in mobile APPs can be easily leaked by screen shooting and transmitting the screenshot to others. Thus, it is critical to find ways to detecting and preventing information leakage [1, 2] or tracing the information leakers [3, 4]. In this paper, we focus on a special application scenario: for some mobile APPs, high authority managers may leak some sensitive information that can only be seen in their login page to other normal users by making a screenshot of their page and transmitting the screenshot image via instant messaging software such as Twitter and WeChat. A simple solution to tracing the source is displaying the manager's ID in the login page. However, the leaker will definitely remove this information from the screenshot. Another solution is to spread the manager's information in different parts of the login page by visible watermarks. However, these visible watermarks will reduce the aesthetics of the login page.

In the past two decades, information hiding has been widely used in copyright protection, transaction tracing, content authentication and covert transmission [5, 6]. Digital Watermarking technology is to embed some identification information (i.e., digital watermark) directly into a digital carrier (including multimedia, documents, software, etc.) or indirectly express (modifying the structure of a specific area), and does not affect the use of the original carrier value is not easy to be detected and modified again. But it can be identified and identified by the manufacturer. Through the information hidden in the carrier, the purpose of confirming the content creator, purchaser, transmitting secret information or judging whether the carrier has been tampered can be achieved. Digital watermarking is an effective way to protect information security, realize anti-counterfeiting traceability, and copyright protection. It is an important branch and research direction in the field of information hiding technology. Thus, we can adopt robust image watermarking schemes to solve the above situation. The embedded watermark should be robust against scaling and JPEG compression operations, since the original screenshot image will be often compressed by JPEG with a relatively low quality factor (e.g., QF=70) and resized to a smaller image before transmission via instant messaging software. Recently, many robust image watermarking schemes have been proposed [7-9]. Chen and Wornell [7] introduced an important class of embedding methods termed quantization index modulation (QIM) and developed a realization form named dither modulation. Ali et al. [8] proposed a robust image watermarking technique using SVD and differential evolution in the DCT domain. Reference [9] proposed a robust image watermarking method using DCT and YCoCg-R color space. Although the methods in [8,9] are robust to known-angle rotation attacks, they are either with high time complexity [8, 9] or not robust to unknown-size scaling attacks [7-9]. There are also some other methods [10,11] can be considered to solve the security problem. In this paper, considering the time complexity, we adopt QIM in the DCT domain with special pre-processing steps to make it robust to scaling and JPEG compression.

2. **Proposed SRQIMTP Scheme.** To trace the information leaker via screenshot based on watermarking, three aspects should be considered. First, the embedding process should be as fast as possible which is performed once the screenshot operation is detected in mobile apps. Second, the embedded watermark should be very robust to scaling and JPEG compression. Third, the extraction result should belong to one of three cases, i.e., correct leaker's ID, failure, without watermark. Quantization index modulation (QIM) is a computationally efficient method of watermarking with side information. As we know, the most serious disadvantage of QIM has been its extreme sensitivity to valumetric scaling. Even small changes in the brightness of an image, or the volume of a song, can result in dramatic increases in the bit-error rate (BER). And the conventional QIM is a typical blind robust watermarking scheme that is not robust to scaling attacks, although the complexity is relatively low. To make it robust to scaling operations, our SRQIMTP scheme performs a resizing step on the original screenshot image before embedding and also on the screenshot image to be detected before extraction. To make QIM robust enough to JPEG compression, our scheme performs QIM on 8×8 DCT blocks. To make our scheme can distinguish different extraction results and also make the scheme as robust as possible, we perform a special watermark template pattern mapping on the original ID to be embedded before embedding and also perform the inverse mapping on the extracted watermark. To make our scheme as fast as possible, we also perform many speed-up operations on the whole watermarking systems such as using the integer DCT and fast interpolation in resizing.

Assume that a mobile APP has the function of detecting screenshot operations, when a high authority manager logs in this mobile APP, he make a screenshot of the login page and would like to leak the information to other users. Once the APP captures the screenshot operation, the APP launches the watermark embedding operation. Assume that the manager's ID is encoded with 32bits composed of bits '0' and '1', and the original screenshot image is of size $M \times N$. Based on these assumptions and above considerations, the embedding process of our SRQIMTP algorithm can be illustrated as follows:

Step 1: Pre-processing of the watermark information. Segmenting the 32bits binary ID into 8 segments, each having 4 bits. Each segment is an integer in [0,15], which is mapped into one of 16 template patterns. Here, each pattern is a binary image of size 16×16. Thus, the final watermark to be embedded becomes 8 template patterns, i.e. $8 \times 16 \times 16$ bits. An example of 16 patterns is shown in Fig. 1. The patterns can be optimally designed in order to distinguish each other.

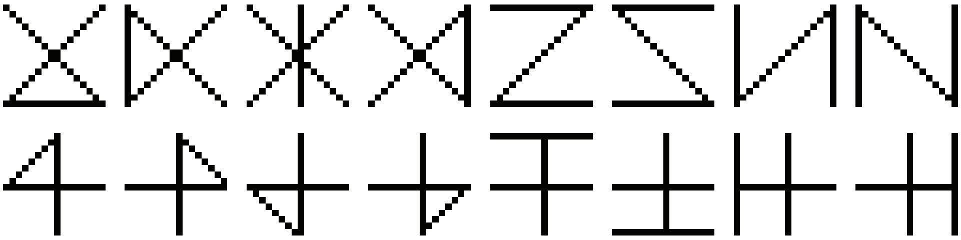


Figure 1. Example of 16 template patterns

Step 2: Pre-processing of the original screenshot image. The original screenshot image is resized into a fixed width W (e.g., $W = 720$) preserving the aspect ratio based on linear interpolation. This step is an essential step to achieve the scaling-resilience performance.

Step 3: Watermark embedding.

Step 3.1: The Y component of the original screenshot image is segmented into non-overlapping blocks of size 8×8,

Step 3.2: The integer DCT is performed on each block to obtain 64 coefficients. Scan these coefficients in the zigzag manner and select two mid-frequency coefficients to embed watermark bits. That is, each block can be embedded with 2 watermark bits.

Step 3.3: The $8 \times 16 \times 16$ watermark (i.e., 8 template patterns) bits are sequentially embedded into DCT blocks using the dither modulation technique [7]. These watermark bits can be redundantly embedded with multiple times (at most N times, e.g., $N = 25$) as long as there are enough DCT blocks to be embedded in order to achieve the best robustness.

Step 3.4: Perform the inverse transform on these watermarked DCT blocks, we can obtain the final watermarked screenshot image. Assume that we have a suspect screenshot image to be detected, the extraction process of our SRQIMTP method can be described as follows:

Step 1: Pre-processing of the suspect screenshot image. The suspect screenshot image is resized into a fixed width W (e.g., $W = 720$) preserving the aspect ratio based on linear interpolation.

Step 2: Watermark extraction.

Step 2.1: The Y component of the suspect screenshot image is segmented into non-overlapping blocks of size 8×8,

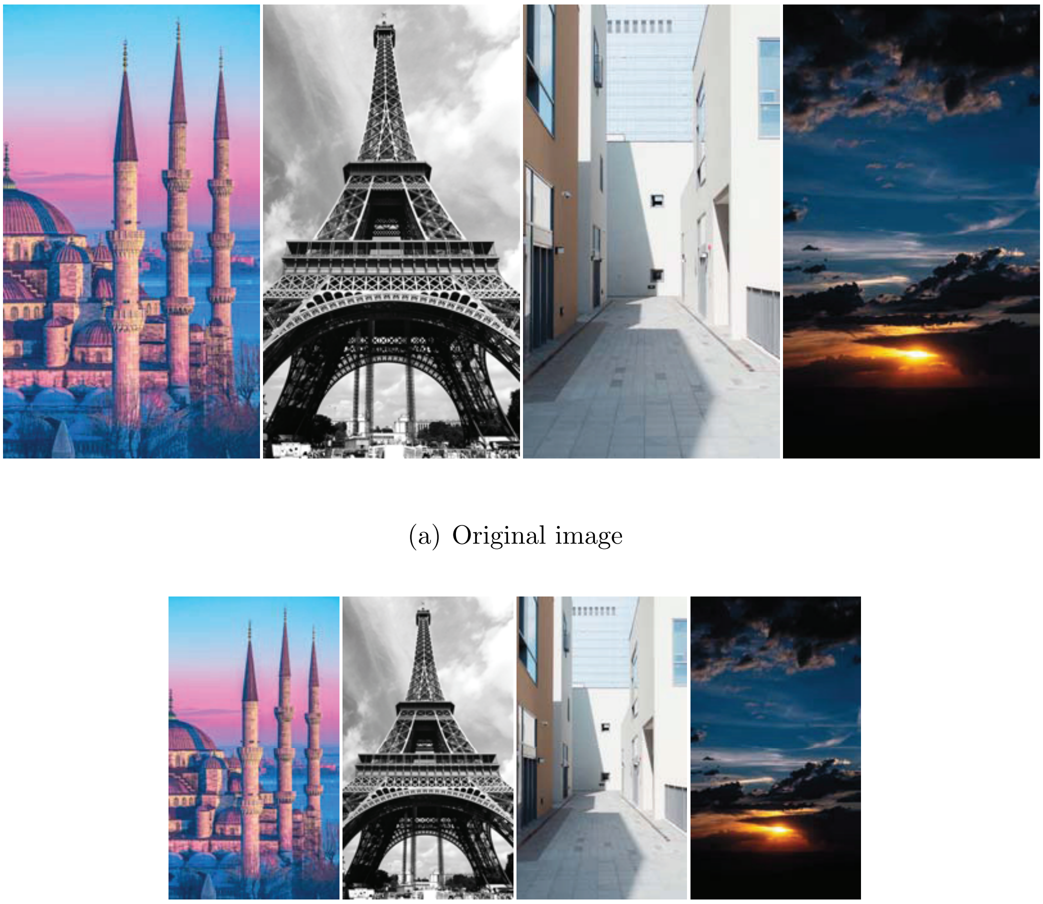
Step 2.2: The integer DCT is performed on each block to obtain 64 coefficients. Scan these coefficients in the zigzag manner and extract two watermark bits based on the QIM technique from two mid-frequency coefficients that were used to embed watermark bits.

Step 2.3: Assume that we have extracted S bits from the DCT blocks, then $S/(8 \times 16 \times 16)$ watermarks can be obtained, we just average these watermarks to obtain the final watermark (i.e., 8 patterns) .

Step 4: Post-processing and Decision.

Step 4.1: Post-processing of the 8 obtained patterns. Each obtained pattern is compared with 32 template patterns using template matching, and the template pattern with maximum similarity $s_i (i = 1, 2, \ldots, 8)$ is output as the resulting pattern.

Step 4.2: Decision. If all of the 8 similarities are not less than the threshold t, then the indices of 8 corresponding template patterns compose the final detected 32-bit ID.



(a) Original image

(b) Attacked watermarked image

Figure 2. Watermark embedding and attacking examples of our scheme

If one of similarities s_i is smaller than t, then the ID detection fails. If all of the eight similarities are smaller than t, then there is no watermark embedded.

3. **Experimental Results and Discussions.** To demonstrate the effectiveness of the proposed scheme (SRQIMTP), we adopted 1000 screenshot images in PNG format to evaluate the robustness of our scheme to JPEG compression and scaling operations, as well as some other image processing operations. Fig.2 shows an example of embedding a 32-bit ID (10100101010001001000001010100000) into four screenshot PNG images of size 1080 × 1920 (Fig.2(a)) and extracting the correct ID from there JPEG-compressed (QF=70) and scaled version Fig.2(b) (reduced into 2/3 original size, i.e., 720 × 1280). From this example, we can see that our scheme can effectively extract the correct ID from the commonly attacked image. To show the superiority of our scheme under the simulated instant messaging based image transmission environment, we compare it with other three schemes, i.e., original QIM [7], Ali et al.'s [8] in 2014 and Liu et al.'s [9] in 2017, in terms of average PSNR of 800 watermarked images (Y component), detection accuracy and average embedding time complexity (compared with our scheme). Here, 1000 screenshot images of different sizes from different types of mobile phones are used. Among them, 200 images are not embedded with information, 800 images are embedded with a randomly generated 32-bit ID (For our scheme, the 32-bit ID is transformed into 8

Table 1. Comparisons of our scheme with other three schemes in terms of average PSNR,detection accuracy and average embedding time complexity

Method	[7]	[8]	[9]	Our
Average PSNR (dB)	39.62	35.12	39.95	39.54
Detection accuracy	0	0	0	98.5%
Average time complexity	0.95	2.91	1.53	1.0

Table 2. Comparisons of our scheme with other three schemes in terms of detection accuracy under different attacks.

Method	[7]	[8]	[9]	Our
JPEG(QF=80)	100%	100%	100%	100%
JPEG(QF=60)	99.6%	99.9%	100%	99.9%
Scaling(3/4 original size)	0	0	0	100%
Scaling(1/2 original size)	0	0	0	99.8%
Gaussian filtering ($\sigma = 1$)	99.7%	95.1%	99.7%	99.7%
Cropping left-upper corner	100%	99.9%	85%	100%
Median filtering (3×3)	99.5%	99.3%	99.5%	99.6%

binary template patterns and then embedded in each image for at most N times, while for methods [7-9], the 32-bit ID is embedded directly in each image for at most 99N times, here N=19, t=0.6). Each of 1000 images is then attacked by JPEG compression (QF=70) together with scaling (reducing into 2/3 original size). The detection accuracy means the number of exactly detected correct IDs divided by 800. Table 1 shows the comparison results. From these results, we can see that our scheme is overall the best, other schemes are unable to restore the correct ID after the composite attack, i.e., scaling + JPEG compression. Note that here we don't know the original image size in the watermark extraction stage, that is, the scaled image is directly sent into the extraction process without size restoration.

To further show the superiority of our scheme for various single attacks in terms of detection accuracy, we compare it with the methods in [7-9] based on 1000 screenshot images. We perform different kinds of attacks on the watermarked images, including JPEG compression (QF=80), JPEG compression (QF=60), scaling (reducing into 3/4 original size), scaling (reducing into 1/2 original size), Gaussian filtering ($\sigma = 1$), cropping the 1/4 left-upper corner, and 3×3 median filtering. Table 2 shows the comparison results, which can demonstrate the superiority of our scheme.

4. **Conclusion.**

This paper presents an effective source tracking scheme for detecting information leakers based on robust image watermarking. We adopt QIM in the DCT domain with special pre-processing steps. Our scheme is robust to the composite attack of scaling and JPEG compression. Experimental results show the effectiveness and superiority of our scheme.

Acknowledgment. This work is partially supported in part by the National Key Research and Development Program of China under Grant No.2020AAA0140004, the Natural Science Foundation of Zhejiang Provincial under Grant LQ20F030010, the New 115 Talent Project of Quzhou City, and the Project for Public Interest Research Projects of Science and Technology Program of Zhejiang Province, China(No.LGG21F020005, NO. LGF21F010002, NO. LGN21C130001 and NO. LGG18F010013).

REFERENCES

[1] J. Choi, W. Sung, C. Choi, and P. Kim, Personal information leakage detection method using the inference-based access control model on the Android platform, *Pervasive and Mobile Computing*, vol. 24, pp. 138–149, December 2015.

[2] B. Hauer, Data and information leakage prevention within the scope of information security, *IEEE Access*, vol. 3, pp. 2554–2565, December 2015.

[3] C. Liu, H. Ling, F. Zou, L. Yan, X. Ou, Efficient digital fingerprints tracing, *IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, pp. 431–435, 2014.

[4] D. Megías, A. Qureshi, Collusion-resistant and privacy-preserving P2P multimedia distribution based on recombined fingerprinting, *Expert Systems with Applications*, vol.71, pp. 147–172, April 2017.

[5] Z. M. Lu, D. G. Xu, S. H. Sun, Multipurpose image watermarking algorithm based on multistage vector quantization, *IEEE Transactions on Image Processing*, vol. 14, no.6, pp. 822–831, June 2005.

[6] D. Li, G. Zhai, X. Yang, M. Hu, J. Liu, Perceptual information hiding based on multi-channel visual masking, *Neurocomputing*, vol. 269, pp. 170–179, December 2017.

[7] B. Chen, G. W. Wornell, Quantization index modulation: a class of provably good methods for digital watermarking and information embedding, *IEEE Transactions on Information Theory*, vol. 47, no.4, pp. 1423–1443, May 2001.

[8] M. Ali, C. W. Ahn, M. Pant, A robust image watermarking technique using SVD and differential evolution in DCT domain, *Optik*, vol. 125, no.1, pp. 428–434, January 2014.

[9] M. Mohammad, E. Gholamhossein, An improved robust image watermarking method using DCT and YCoCg-R color space, *Optik*, vol. 140, pp. 975–988, July 2017.

[10] T. Y. Wu, X. Fan, K. H. Wang, J. S. Pan, C. M. Chen, Security analysis and improvement on an image encryption algorithm using Chebyshev generator, *Journal of Internet Technology*, vol. 20, no. 1, pp. 13–23, 2019.

[11] P. Wang, C. Chen, S. Kumari, M. Shojafar, R. Tafazolli, Y. Liu, HDMA: Hybrid D2D Message Authentication Scheme for 5G-Enabled VANETs, *IEEE Transactions on Intelligent Transportation Systems*, pp.1-10, 2020.