Survey of Secure and Compression Encryption Key Based Textual Data Deduplication using Digital Watermarking

Chandra Bhan Yadav
M. Tech. Scholar
Department of Electronics and Communication
TIT, Bhopal

Prof. Hema Singh
Professor
Department of Electronics and Communication
TIT, Bhopal

Abstract: - Digital watermarking is one of the important digital image processing techniques and it is a popular research area. Due to development in multimedia technology and vast use of Internet, the watermarking technique has developed rapidly and it plays a major role in digital data security. The introduction to telemedicine, e-diagnosis and tele-diagnosis, leads to creation of copy, transmit and distribute digital data over networks which involves security risks. Providing security and authentication to the medical data has become an essential aspect in the field of medical sciences. This paper studied of compressed textual data deduplication using encryption key and digital watermarking technique. The symmetric Key is also called as session key and each symmetric key is used only once. The encrypted message and its session key are sent to the recipients of the email message. The session key must be sent to the email recipients so that, they know how to decrypt the message. The public key encryption algorithm is used to securely share the session key with the recipients of the email.

Keywords: - Digital Watermarking, Encryption Key, Secure Data

I. INTRODUCTION

A digital watermark is a kind of marker covertly embedded in a noise-tolerant signal such as audio or image. It is typically used to identify the ownership of the copyright of such signal.
Computer aided hiding of the given digitized information in a carrier is called watermarking. Digital watermarks may be used to verify the authenticity or integrity of the carrier signal or to show the identity of the source [1]. It is prominently used for tracing copyright infringements and for bank note authentication. Like traditional watermarks, digital watermarks are perceptible only under certain conditions. If a digital watermark distorts the carrier in a way that it becomes perceivable, then it is of no use. Traditional watermarks may be applied to visible media (like images or video), whereas in digital watermarking, the signal may be audio, pictures, video, texts or 3D models. A signal may carry several different watermarks at the same time. Unlike some techniques in which metadata were added to the carrier signal, a digital watermark does not change the size of the source image at the receiving end.
A watermarking system is usually divided into three distinct stages: embedding, attack and retrieval as shown in Figure 1.

In the embedding part, an algorithm accepts the host image and the watermark data to be embedded and produces a watermarked signal [2]. The watermarked digital signal is then transmitted or stored. While the modification may not be malicious, the term attack arises from the copyright protection application, wherein third parties may attempt to remove the digital watermark through modification. There are many possible modifications, for example, a lossy compression of the data (in which resolution is diminished), cropping an image or video or intentionally adding noise.

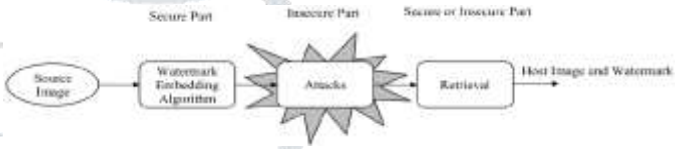

Figure 1: General Watermarking Systems

This might lead to further duplication and re-distribution leaving the rights holders powerless and royalty-less [3]. To enhance the security of audio data, digital watermarking and steganography techniques complement cryptography for protecting content even after it is deciphered [4].
The study of multimedia security [5] therefore includes not just encryption but also watermarking and steganography. Steganography and Watermarking almost interchangeably, refers to hiding secondary information into the primary multimedia source. The primary multimedia sources can be audio, image, and video. There are unique techniques associated with each type of primary perceptual sources depending on their inherent redundancy and perceptual properties. These techniques have been proposed as alternative methods to enforce the intellectual property rights and protect digital media from tampering [6]. In this thesis work the primary multimedia source is image.
The word steganography was originated from Greek which means covered writing. Steganography is the oldest form of covert channel. A famous illustration of steganography is Simmons' Prisoners' Problem [7]. Audio Steganography is the act of embedding a secret message within a larger message so that others cannot discern the presence of the secret message [8]. Steganography can be used to hide a message intended for later retrieval by a specific individual or group. Audio watermarking involves a process of embedding into host audio signal a perceptually transparent digital signature, carrying a message about the host data in order to mark its ownership. The aim in watermarking systems is to ensure the robustness of the hidden message; the presence of the embedded message itself does not have to be secret [9].
The watermark is always present in the signal, even in illegal copies of it and the protection that is offered by the watermarking system is therefore of a permanent kind. Although the process of watermark embedding and steganography are similar, there are some basic differences between the two techniques. Steganography methods assume

that the existence of the covert communication is unknown to third parties and are mainly used in secret one-to-one communication between authorized users. On the other hand, watermarking is to hide message in one-to-many communications. Steganography methods usually do not need to provide strong security against removing or modification of the hidden message. Whereas, watermarking methods need to be very robust to attempts to remove or modify a hidden message.

II. LITERATURE REVIEW

Ali Miri et al. [1], as the need for storage has grown exponentially in recent years, cloud storage has been providing a solution to this need by providing users expanded capacity and access. Providing adequate security and privacy, and lowering storage costs are some of the key challenges facing this solution. A common practice used by cloud service providers (CSPs)- data deduplication - identifies identical copies of users' data, and removing all, but one copy to lower required storage overhead. However, this can result in serious privacy concerns. In this paper, we formulate a new secure deduplication scheme for textual data. Our proposed method uses data encoding and compression techniques that not only result in reduce storage space required, but also in saving in required transmission bandwidth. The security of the data against the semi-honest CSP and malicious users is ensured by using Burrows Wheel Transform encoding scheme. The encoded data is further compressed to gain effective savings in terms of storage and reduced size of the data. Data encoding and data compression techniques are combined together to realize secure and efficient data deduplication. Through our scheme, the CSP will not only achieve huge storage space savings through data compression and data deduplication, but can also provide the users a satisfactory level of security for their data in the cloud.

Awdhesh K. Shukla et al. [2], a high limit information concealing technique utilizing lossless pressure, propelled encryption standard (AES), modified pixel esteem differencing (MPVD), and least significant bit (LSB) substitution is exhibited. Number juggling coding was connected on a mystery message for the lossless pressure, which gave 22% higher implanting limit. After pressure and encryption, the LSB substitution and MPVD are connected. It is tentatively settled that with the proposed strategy, significant upgrade in installing limit was accomplished and additional bits than existing strategies could be implanted because of the utilization of number juggling pressure and MPVD. The MPVD and number juggling coding together came about into 25% upgraded implanting limit than the prior strategies. The proposed technique likewise furnishes elevated amounts of visual quality with a normal of 36.38 dB at 4.00 bpp.

S. Thakur et al. [3], in this paper, we present a powerful and secure watermarking approach utilizing change space method for tele-wellbeing applications. The patient report/personality is inserting into the host medicinal picture with the end goal of verification, explanation and recognizable proof. For better secrecy, we apply the disorder put together encryption calculation with respect to watermarked picture in a less unpredictable way. Exploratory outcomes obviously shown that the proposed strategy is exceedingly hearty and adequate secure for different types of assaults with no huge contortions among watermarked and spread picture. Further, the execution assessment of our strategy is discovered better to existing cutting edge watermarking strategies under thought. Besides, quality investigation of the watermarked picture is assessed by abstract measure which is valuable in quality driven social insurance industry.

R. Srivastava et al. [4], this paper introduce a computationally productive joint intangible picture watermarking and joint photographic specialists gathering (JPEG) pressure conspire. As of late, the transmission and capacity of computerized archives/data over the unbound channel are gigantic concerns and almost the majority of the advanced reports are compacted before they are put away or transmitted to spare the transfer speed prerequisites. There are numerous comparable computational activities performed amid watermarking and pressure which lead to computational excess and time delay. This requests advancement of joint watermarking and pressure conspire for different interactive media substance. In this paper, we propose a strategy for picture watermarking amid JPEG pressure to address the ideal exchange off between significant execution parameters including implanting and pressure rates, strength and installing changes against various realized flag handling assaults.

D. S. Chauhan et al. [5], this paper present a protected restorative picture watermarking system applying spread-range idea in wavelet change space is proposed. In the initial step, discrete wavelet transform(DWT) deteriorates the spread restorative picture into four recurrence sub-groups utilizing Mexican cap as mother wavelet and after that comparing to every pixel of the parallel watermark a couple of Pseudo-Noise (PN) is installed into a level (HL) and a vertical (LH) sub-band. So as to keep up the indistinctness of the watermarked picture, quality of the produced PN arrangement pair is balanced by indicated report to watermark proportion (DWR). For the extraction the watermark, measurable profile of DWT coefficients of watermarked picture is resolved and the got likelihood dispersion work (pdf) is used for planning the watermark location system.

III. DIGITAL WATERMARKING

Watermarking basically refers to information hiding. Information or digital signal in the form of images, audio, video or text is hidden or inserted. This information to be hidden is termed as Watermark. The watermark can be hidden in cover/host/carrier signal. The host popularly can be text file, image, audio file or video file. Depending on the type of host, watermarking can be categorized as:

- Text watermarking
- Digital image watermarking,
- Digital audio watermarking and
- Digital video watermarking

To have efficient copyright protection, watermarking algorithms must possess certain characteristics. Depending on the application requirement different characteristics can be primary objectives. The most desirable characteristics [2] are listed below:

Robustness- Robustness refers to difficulty in removing or destroying watermark from host image when watermarked image is subjected to image processing attacks.

Imperceptibility- Imperceptibility dictates the inability to notice the existence of watermark in host image and retained visual quality of host image after embedding watermark into it.

Capacity- Capacity refers to amount of information that can be embedded in host image. Capacity depends on the application and the image.

Security- Watermarking algorithm is secure if knowing the algorithm to embed and extract the watermark does not help an unauthorised party to detect the presence of watermark.

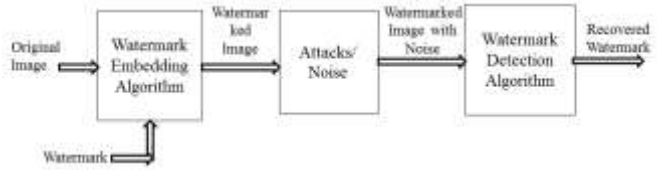

Figure 2: General digital watermark life-cycle phases with embedding-, attacking-, and detection and retrieval functions

All these characteristics cannot be achieved simultaneously as there is always a trade-off between them. For example, robustness and imperceptibility are contradictory to each other. Watermarking algorithm having high robustness usually sacrifices imperceptibility and vice versa. For higher robustness increased capacity is desired. But increased capacity leads to compromising imperceptibility. Watermarking methods introduced in proposed work aim to provide higher robustness as well as imperceptibility.

IV. COMPRESSION

An efficient representation of integers is essential in many applications such as text compression, image compression, fast query evaluation, fast searching, and fast file access, In all these applications, one of the simple algorithms such as Golomb coding, Elias coding, Fibonacci coding etc, is used to represent an arbitrary integer compactly. The selection of the algorithm depends on the type of application and the probability distribution of the integers in that application. We can also construct variable length codes using the above mentioned simple algorithms without knowing the probability of the integers in advance. While many different representations have been developed, it is not always obvious in which circumstances a particular code is to be preferred.
The proposed new variable length code, called Extended Golomb Code (EGC), is presented to code the given non-negative integer N. In EGC, a divisor (d) is selected and the integer N to be coded is divided successively M times by d until the quotient q becomes zero. In each division, the remainders r_i (i = 1 to M) are retained. The integer N is then coded by coding M and the M remainders as

$$\text{Code}(M) = \text{Code}(r_M, r_{M-1} \ldots r1)$$

M is coded in unary and the remainders (r_M, r_{M-1} …r1) are coding using a unique coding scheme. The bit length bl of EGC follows the inequality:

$$bl \le \left(\left|\frac{\log_{10} N}{\log_{10} d}\right| + 1\right) * (1 + Log_2 d)$$

In general, when an integer N is divided by a divisor d, there are d possible remainders when the quotient (q) is greater than 0, and d-1 possible remainders when q is equal to 0.

Algorithm for Encoding
The integers in file to be compressed are encoded using following steps:

1. Select an optimized divisor (d) for the probability distribution of the integers in that file.
2. Divide each integer N successively by d, until the quotient (q) becomes zero. Count the number of divisions made as M. Retain the remainders in each division as r_1, r_2, r_3… r_M. Code r_1, r_2, r_3… r_{M-1} in $\log_2 d$ bits, and r_M in $\log_2 (d-1)$ bits.

Algorithm for Decoding
The following steps are used to decode the data in the compressed file.

1. Read the next bit 0 until bit 1 is encountered and count the no. of reads made so far including the bit 1 as M.
2. Read the bits further and decode M remainders as per the code given for the given divisor d and reconstruct r_i.
3. Then obtain the integer N using the following procedure

```
If      d > 2
{
        Set N = 0
        For i = M to 1
            N = N*d + ri
}
Else
{
        Set N=1
For         i = M-1 to 1
            N = N*d + ri
}
```

Repeat the steps 1 and 2 to obtain all Ns in the compressed file.

V. ENCRYPTION AND DECRYPTION METHODOLOGY

The email message originated via MS outlook email client is encrypted with the new symmetric key encryption algorithm on click of the Encrypt button under Security tab of the compose window. The user on opening the compose window fills the 'To' field, Subject and then adds the message in the message body followed by attachments, if any.

Figure 3: Encryption Process

If the message needs to be encrypted, the user has to click on the encrypt button under the Security tab. On click of the encrypt button the following process begins:

1. A 256 bit length session key is generated.
2. The message body of the compose window is copied to a file, named ”msgbody.txt” and file is appended as one more attachment to the email. On completing the above action, ”msgbody.txt” is deleted from the local computer. The message body is replaced with the following text. “The Message is encrypted. Please click on the decrypt button to read the message”.

The attachments and the generated session key are passed as an input to the new symmetric key encryption algorithm. For this, first all the attachments are copied to a local directory, created for this purpose. The following arguments needs to be passed to the new algorithm: • The session key • The input file (the message that needs to be encrypted/decrypted) • The location of output file, where the encrypted/decrypted data needs to be saved • E/D(E for encrypt and D for decrypt) • Electronic Codebook/Cipher Block Chaining(ECB/CBC cipher modes) The encrypted attachments (including message body) are stored at a specified location. All these encrypted files are appended as an attachment to the mail item. The original attachments (in plain text) are deleted from the mail item so that the compose window will only contain the encrypted attachments and encrypted message body as an attachment.

4. The public key of all the recipients (from ‘To’ filed) is fetched from the SQL Express server one after another.
5. The generated Session key and Public key of the each user is passed as an input to the RSA encryption algorithm. The email address of the user is appended to the output of the encryption algorithm (in the format as shown below) and written to a text file. This process is repeated for all the recipients of the email message. The same text file is appended with the email address and encrypted session key for every user.

Decryption Process:- The message body of the encrypted message will contain the following text. “This message is encrypted. Please click on the decrypt button to read the message”. On trying to open the attachment before decryption, the user will see, only the junk information. The encrypted message will contain the encrypted attachments, encrypted message body, encrypted session key as an attachment to the mail item.

Figure 4: Decryption Process

VI. CONCLUSION

Outsourcing offers the data owner scalability and a low initial investment. It consists of a data owner, a trusted query user and a dubious server. Here, the data owner is the medical organization that consists of a preserved medical image database. A trusted query user is an authenticated doctor who has access to medical database. A dubious server is who makes a prediction of the input image without knowing any details of the process. The data are to be revealed only to trusted users and not to the service provider or to anyone else.

We are encrypting the email using the symmetric key algorithm and share the symmetric (session) key with the recipients of the email using asymmetric encryption algorithm. This is because the asymmetric encryption algorithms are quite computationally intensive and some algorithms can produce cipher text up to twice the size of the original text and therefore it is not used for encrypting the email message.

REFRENCES

[1] Ali Miri and Fatema Rashid, “Secure Textual Data Deduplication Scheme Based on Data Encoding and Compression”, IEEE 2019.

[2] Awdhesh K. Shukla, Akanksha Singh, Balvinder Singh, And Amod Kumar, “A Secure and High-Capacity Data-Hiding Method using Compression, Encryption and Optimized Pixel Value Differencing”, Received July 5, 2018, accepted August 25, 2018, date of publication September 3, 2018, date of current version October 8, 2018.

[3] S. Thakur, A. K. Singh, S. P. Ghrera, and M. Elhoseny, ``Multi-layer security of medical data through watermarking and chaotic encryption for tele-health applications," in *Multimedia Tools and Applications*. New York, NY, USA: Springer, 2018, pp. 1_14.

[4] R. Srivastava, B. Kumar, A. K. Singh, and A. Mohan, “Computationally efficient joint imperceptible image watermarking and JPEG compression: A green computing approach," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 16447_16459, 2017.

[5] D. S. Chauhan, A. K. Singh, A. Adarsh, B. Kumar, and J. P. Saini, “Combining Mexican hat wavelet and spread spectrum for adaptive water-marking and its statistical detection using medical images," in *MultimediaTools and Applications*. New York, NY, USA: Springer, 2017, pp. 1_15.

[6] N. Senthil Kumaran, and S. Abinaya, “Comparison Analysis of Digital Image Watermarking using DWT and LSB Technique”, International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[7] Aase, S.O., Husoy, J.H. and Waldemar, P. (2014) A Critique of SVD-Based Image Coding Systems, IEEE International Symposium on Circuits and Systems VLSI, Orlando, FL, Vol. 4, Pp. 13-16.

[8] Ahmed, F. and Moskowitz, I.S. (2014) Composite Signature Based Watermarking for Fingerprint Authentication, ACM Multimedia and Security Workshop, New York, Pp.1-8.

[9] Akhaee, M.A., Sahraeian, S.M.E. and Jin, C. (2013) Blind Image Watermarking Using a Sample Projection Approach, IEEE Transactions on Information Forensics and Security, Vol. 6, Issue 3, Pp.883-893.

[10] Ali, J.M.H. and Hassanien, A.E. (2012) An Iris Recognition System to Enhance E-security Environment Based on Wavelet Theory, Advanced Modeling and Optimization, Vol. 5, No. 2, Pp. 93-104.

[11] Al-Otum, H.M. and Samara, N.A. (2009) A robust blind color image watermarking based on wavelet-tree bit host difference selection, Signal Processing, Vol. 90, Issue 8, Pp. 2498-2512.

[12] Ateniese, G., Blundo, C., De Santis, A. and Stinson, D.R. (1996) Visual cryptography for general access structures, Information Computation, Vol. 129, Pp. 86-106.

[13] Baaziz, N., Zheng, D. and Wang, D. (2011) Image quality assessment based on multiple watermarking approach, IEEE 13th International Workshop on Multimedia Signal Processing (MMSP), Hangzhou, Pp.1-5.

[14] Bao, F., Deng, R., Deing, X. and Yang, Y. (2008) Private Query on Encrypted Data in Multi-User Settings, Proceedings of 4th International Conference on Information Security Practice and Experience (ISPEC 2008), Pp. 71-85, 2008.

[15] Barni, M. and Bartolini, F. (2004) Watermarking systems engineering: Enabling digital assets security and other application, Signal processing and communications series, Marcel Dekker Inc., New York.

[16] Barni, M., Bartolini, F. and Piva, A. (2001) Improved Wavelet based Watermarking Through Pixel-Wise Masking, IEEE Transactions on Image Processing, Vol. 10, Pp. 783-791.