IJDCST @ Nov-Dec, Issue- V-3, I-1, SW-064
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Obtaining Embedded Data in a Spectrum Domain of Digital Media

Veera V Chantibabu Yasam[1], Sk Haseena Parveen[2]

[1]M.Tech (CSE), Nimra College of Engineering and Technology, A.P., India.

[2]Asst. Professor, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.

Abstract— Several ways have been adopted and as well as invented over the years to transfer the data securely under the aegis of digital medium. Consider the problem of extracting blindly data embedded over a wide band in a spectrum (transform) domain of a digital medium (image, audio, and video). We develop a novel multicarrier/ signature iterative generalized least-squares (M-IGLS) core procedure to seek unknown data hidden in hosts via multicarrier spread-spectrum embedding. Neither the original host nor the embedding carriers are assumed available. Experimental studies on images show that the developed algorithm can achieve recovery probability of error close to what may be attained with known embedding carriers and host autocorrelation matrix.

Keywords — Authentication, annotation, blind detection, covert communications, data hiding, information hiding, spread-spectrum embedding, steganalysis, steganography, watermarking.

I. INTRODUCTION

The field of embedding data in digital media is an information technology growing rapidly commercial as well as national security interest. Applications may vary from annotation, copyright-marking, and watermarking, to singlestream media merging (text, audio, image) and covert communication [1].

In annotation, secondary data are embedded into digital multimedia to provide a way to deliver side information for various purposes; copyright-marking may act as permanent "iron branding" to show ownership; fragile watermarking may be intended to detect future tampering; hidden low-probabilityto-detect (LPD) watermarking may serve as identification for confidential data validation or digital fingerprinting for tracing purposes [2]-[4]. Covert communication or steganography, which literally means "covered writing" in Greek, is the process of hiding data under a cover medium such as image, video, or audio, to establish secret communication between trusting parties and conceal the existence of embedded data [5]. As a general encompassing comment, different applications of information hiding, such as the ones identified above, require different satisfactory tradeoffs between the following four basic attributes of data hiding [6]: (*i*) Payload - information delivery rate; (*ii*) robustness - hidden data resistance to noise/disturbance; (*iii*) transparency - low host distortion for concealment purposes; and (*iv*) security - inability by unauthorized users to detect/access the communication channel.

II. RELATED WORK

Recently, developing data embedding technologies are being seen to pose a threat to personal privacy, commercial, and national security interests [7]. In this work, we focus our attention on the blind recovery of secret data hidden in medium hosts via multi-carrier/signature direct-sequence spread-spectrum (DS-SS) transform domain embedding [8]. Neither the original host nor the

embedding carriers are assumed known (fully blind data extraction). This blind hidden data extraction problem has also been referred to as "Watermarked content Only Attack" (WOA) in the watermarking security context [9].

While passive detection-only of the presence of embedded data is being intensively investigated in the past few years [10], active hidden data extraction is a relatively new branch of research. In blind extraction of SS embedded data, the unknown host acts as a source of interference/disturbance to the data to be recovered and, in a way, the problem parallels blind signal separation (BSS) applications as they arise in the fields of array processing, biomedical signal processing, and code-division multiple-access (CDMA) communication systems [11]. Under the assumption that the embedded secret messages are independent identically distributed (i.i.d.) random sequences and independent to the cover host, independent component analysis (ICA) may be utilized to pursue hidden data extraction [12]. However, ICA-based BSS algorithms are not effective in the presence of correlated signal interference as is the case in SS multimedia embedding and degrade rapidly as the dimension of the carrier (signature) decreases relative to the message size.

In [13], an iterative generalized least squares (IGLS) procedure was developed to blindly recover unknown messages hidden in image hosts via SS embedding. The algorithm has low complexity and strong recovery performance. However, the scheme is designed solely for single-carrier SS embedding where messages are hidden with one signature only and is not generalizable to the multicarrier case. Realistically, an embedder would favor multicarrier SS transform-domain embedding to increase security and/or payload rate.

In this paper, we develop a novel multi-carrier iterative generalized least squares (M-IGLS) algorithm for SS hidden data extraction that, to the best of the authors' knowledge, appears for the first time in the broad communication theory and systems literature. For improved recovery performance, in particular for small hidden messages that pose the greatest challenge, experimental studies indicate that a few independent M-IGLS re-initializations and executions on the host can lead to hidden data recovery with probability of error close to what may be attained with known embedding carriers and known original host autocorrelation matrix. Applications of the developed algorithm are, of course, not limited to attacking steganographic covert communications by recovering the secret embedded messages. Since the carriers are also jointly estimated with the embedded data, the developed scheme can also be used for complete message removal or tampering attack as well by reinserting a fabricated message in place of the original. From the opposite data embedding point of view, the developed algorithm can be treated as a tool to test security robustness of SS data hiding schemes.

III. MULTI-CARRIER SS EMBEDDING AND EXTRACTION: PROBLEM FORMULATION

Consider a host image H ∈ MN1×N2 where M is the finite image alphabet and N1×N2 is the image size in

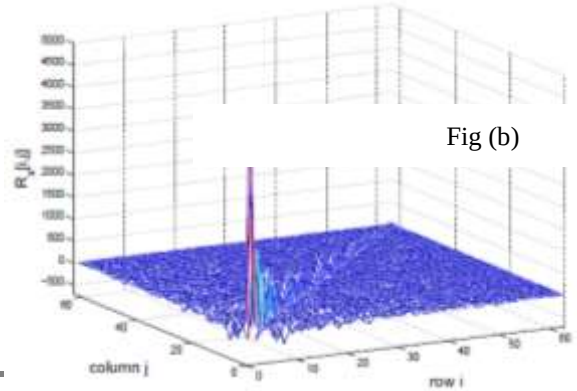

Fig (b)

pixels. Without loss of generality, the image H is partitioned into M local non-overlapping blocks of size N1N2 M . Each block, H1,H2, ...,HM, is to carry K hidden information bits (KM bits total image payload). Embedding is performed in a 2-D transform domain T (such as the discrete cosine transform, a wavelet transform, etc.). After transform calculation and vectorization (for example by conventional zig-zag scanning), we obtain T (Hm) ∈ R N1N2 M ,m = 1, 2, . . . ,M. From the transform domain vectors T (Hm) we choose a fixed subset of L ≤ N1N2 M coefficients (bins) to form the final host vectors x(m) ∈ RL, m = 1, 2, . . . ,M. It is common and appropriate to avoid the dc coefficient (if applicable) due to high perceptual sensitivity in changes of the dc value.

The autocorrelation matrix of the host data x is an important statistical quantity for our developments and is defined as

$$\mathbf{R}_{\mathbf{x}} \triangleq \mathbb{E}\{\mathbf{x}\mathbf{x}^T\} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}(m)\mathbf{x}(m)^T.$$

It is easy to verify that in general Rx ≠ αIL, _ > 0; that is, Rx is *not* constant value diagonal or "white" in field language. For example, 8×8 DCT with 63-bin host data formation (excluding only the dc coefficient) for the 256×256 gray-scale Baboon image in Fig. 1(a) gives the host autocorrelation matrix Rx in Fig. 1(b) [14].

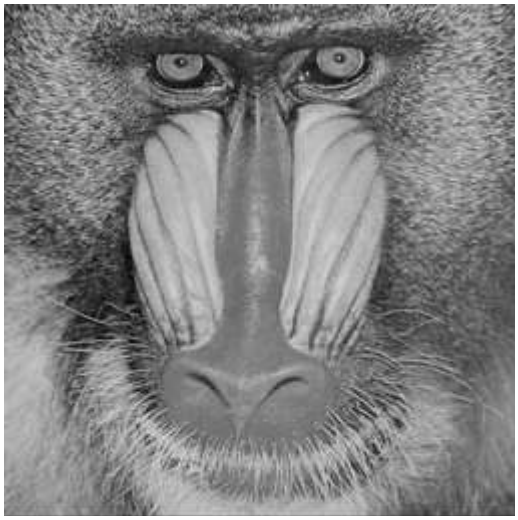

Fig (a)

Fig. (a) Baboon image example
H € {0, 1, ..., 255}$^{256\times256.}$

Fig. (b) Host data autocorrelation matrix
(8 × 8 DCT, 63-bin host) [14].

IV. HIDDEN DATA EXTRACTION

If Z were to be modeled as Gaussian distributed, the joint maximum-likelihood (ML) estimator of V and decoder of B would be

$$\widehat{\mathrm{V}},\widehat{\mathrm{B}} = \arg \min_{\substack{\mathrm{B}\in\{\pm 1\}^{K\times M} \\ \mathrm{V}\in\mathbb{R}^{L\times K}}} \|\mathrm{R}_{\mathrm{z}}^{-\frac{1}{2}}(\mathrm{Y}-\mathrm{VB})\|_F^2$$

where multiplication by $R_z^{-1/2}$ can be interpreted as prewhitening of the compound observation data. If aussianity of Z is not to be invoked, then (9) can be simply referred to as the joint generalized least-squares (GLS) solution[2] of V and B.

The global GLS-optimal message matrix B in above function can be computed independently of V by exhaustive search over all possible choices under the criterion function

$$\widehat{\mathrm{B}} = \arg \min_{\mathrm{B}\in\{\pm 1\}^{K\times M}} \|\mathrm{R}_{\mathrm{z}}^{-\frac{1}{2}}\mathrm{YP}_{\perp\mathrm{B}}\|_F^2$$

The computational complexity of the P-times re-initialized MIGLS is, of course,$O(PD(2K^3 +2LMK +K^2(3L+M)+ L^2K))$ where D represents the number of internal iterations in d.

V. EXPERIMENTAL STUDIES

A technically firm and keen measure of quality of a hidden message extraction solution is the difference in bit-error-rate (BER) experienced by the intended recipient and the analyst. The intended recipient in our studies may be using any of the following three message recovery methods: (*i*) Standard carrier matched-filtering (MF) with the known carriers sk, k = 1, ...,K; (*ii*) sample-matrix-inversion MMSE (SMI-MMSE) filtering with known carriers sk and estimated host autocorrelation matrix bRy (see (3)); and (*iii*) ideal MMSE filtering with known carriers sk and known true host autocorrelation matrix Rx, which serves as the ultimate performance bound reference for all methods. In terms of blind extraction (neither sk nor Rx known), we will examine: (*iv*) The developed MIGLS algorithm in Table I with P = 20 re-initializations and, for comparison purposes, the performance of two typical independent component analysis (ICA) based blind signal separation (BSS) algorithms (*v*) FastICA [44], and (*vi*) JADE [45].

An encompassing conclusion over all executed experiments is that M-IGLS remains a most effective technique to blindly extract hidden messages, while extraction becomes more challenging as the length of the hidden message per used embedding carrier decreases or the number of hidden messages (number of used carriers) increases. It is also worth pointing out that, in these experimental studies, M-IGLS may outperform (in moderate to high distortion values) SMI-MMSE in which the true carriers/signatures are known. This is because SMI-MMSE suffers from performance degradation due to small-sample-support adaptation (estimation of matrix Ry). The unsatisfactory performance of the ICA-based methods is due to the interference from high-amplitude (low-frequency) host coefficients. To demonstrate this point, in Fig. 10 we repeat the exact same experiment of Fig. 2 using this time only the L = 20 highest-frequency DCT coefficients as our host vector. It can be observed that, in this moderate host interference environment, ICA-based methods can provide

satisfactory performance (not superior to M-IGLS, however). Of course, we may not expect that data are always embedded exclusively in low-amplitude coefficients alone.

Next, for the sake of enhanced experimental credibility, we examine the average performance of the proposed MIGLS algorithm over a large image database. The experimental image data set combined, consists of more than 11, 500 8-bit gray-scale photographic images which have great variety (e.g., outdoor/indoor, daylight/night, natural/manmade) and different sizes. We embed one up to five messages, K ∈ {1, 2, . . . , 5}, via multi-carrier SS embedding with arbitrary carriers and payload between 0.016 and 0.078 bits per pixel (bpp). The length of the embedding carriers varies between 30 and 63, L ∈ {30, 31, . . . , 63}. Recovery performance plots are given in Fig. 11. Similar conclusions can be drawn as in the previous individual image host experimentations.

While our blind data extraction algorithmic development was based on the most common SS embedding form (1) for convenience in presentation, the developed algorithm can also be applied to more advanced SS embedding schemes such as improved spread-spectrum (ISS) and correlation-aware improved spread-spectrum (CAISS). We go again over the whole [15] databases under ISS embedding and under CAISS embedding (with amplitude proportion parameter _ = 0.7)[4].

VI. CONCLUSION

In this paper we considered the problem of blindly extracting unknown messages hidden in image hosts via multi- carrier/signature spread-spectrum embedding. Neither the original host nor the embedding carriers are assumed available. We developed a low complexity multi-carrier iterative generalized least-squares (M-IGLS) core algorithm. Experimental studies showed that M-IGLS can achieve probability of error rather close to what may be attained with known embedding signatures and known original host autocorrelation matrix and presents itself as an effective countermeasure to conventional SS data embedding/ hiding.

References

[1] Michael Armbrust, Armando Fox, Rean Gri_th, Anthony D. Joseph, Randy Katz, Andy Kon- winski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. A view of cloud computing. Commun. ACM, 53(4):50{58, April 2010.

[2] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding: A survey," Proc. IEEE (Special Issue on Identification and Protection of Multimedia Information), vol. 87, pp. 1062-1078, July 1999.

[3] I. J. Cox, M. L. Miller, and J. A. Bloom, Digital Watermarking. San Francisco, CA: Morgan-Kaufmann, 2002.

[4] F. Hartung and M. Kutter, "Multimedia watermarking techniques," Proc. IEEE (Special Issue on Identification and Protection of Multimedia Information), vol. 87, pp. 1079-1107, July 1999.

[5] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk, "Watermarking digital image and video data: A state-of-the-art overview," IEEE Signal Processing Magazine, vol. 17, pp. 20-46, Sept. 2000.

[6] N. F. Johnson and S. Katzenbeisser, "A survey of steganographic techniques," in Information Hiding, S. Katzenbeisser and F. Petitcolas Eds. Norwood, MA: Artech House, 2000, pp. 43-78.

[7] Y. Wang and P. Moulin, "Perfectly secure steganography: Capacity, error exponents, and code constructions," *IEEE Trans. Inform. Theory*, vol. 54, pp. 2706-2722, June 2008.

[8] Federal plan for cyber security and information assurance research and development, Interagency Working Group on Cyber Security and Information Assurance, Apr. 2006.

[9] H. S. Malvar and D. A. Florencio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Proc.*, vol. 51, pp. 898-905, Apr. 2003.

[10] L. P´erez-Freire and F. P´erez-Gonz´alez, "Spread-spectrum watermarking security," IEEE Trans. Inform. Forensics and Security, vol. 4, pp. 2-24, Mar. 2009.

[11] S. Lyu and H. Farid, "Steganalysis using higher-order image statistics," IEEE Trans. Inform. Forensics and Security, vol. 1, pp. 111-119, Mar.2006.

[12] D. G. Manolakis, V. K. Ingle, and S. M. Kogon. Statistical and adaptive signal processing: Spectral estimation, signal modeling, adaptive filtering and array processing. Boston, MA: McGraw-Hill, 2000.

[13] P. Bas and F. Cayre, "Achieving subspace or key security for WOA using natural or circular watermarking," in Proc. ACM Multimedia and Security Workshop, Geneva, Switzerland, Sept. 2006.

[14] M. Gkizeli, D. A. Pados, S. N. Batalama, and M. J. Medley, "Blind iterative recovery of spread-spectrum steganographic messages," in Proc. IEEE Intern. Conf. Image Proc. (ICIP), Genova, Italy, Sept. 2005, vol. 2, pp. 11-14.

[15] M. Gkizeli, D. A. Pados, and M. J. Medley, "Optimal signature design for spread-spectrum steganography," *IEEE Trans. Image Proc.*, vol. 16, pp. 391-405, Feb. 2007.

[16] G. Schaefer and M. Stich, "UCID–An uncompressed colour image database," in Proc. SPIE, Storage and Retrieval Methods and Application for Multimedia, San Jose, CA, Jan. 2004, pp. 472-480.