Global Information Assurance Certification Paper

Zsolt Szabo
GSEC Practical v1.4b

Watermarks and Hidden Information in Communications: An Introductory Guide to Contemporary Methods

Abstract:

Spurred by advances in high quality, low bit-rate media compression techniques and availability of inexpensive residential broadband connectivity that combined make rapid and wide-spread distribution of high quality audio and visual content a reality today, the field of steganography and media watermarking has in recent years emerged as a hot topic of interest among entities seeking copyright enforcement of digital media through fingerprinting and proof of ownership technologies.

This paper introduces the reader to the principles behind watermarking and steganography. It points out and contrasts requirements stipulated by the design objectives of the two concepts. Further, it presents techniques for watermarking or information-hiding in different types of media and summarizes strengths and weaknesses related to these. Emphasis is placed on watermarking/fingerprinting digital audio and video, since secure distribution of these types of media has been the predominant driving force in the field.

The word *steganography* has its roots in the Greek language, translating in literal meaning to “covered writing”[1]. According to Webster's Revised Unabridged Dictionary, steganography is defined as the science of publishing information within a container message, such that the presence and content of this information are hidden from third parties[2]. Some additional technical terms used in the field are as follows:

-*cover medium* This is the carrier data which holds the hidden content; it can either be used merely as a communication medium for exchanging hidden information (i.e. its own contents are meaningless), or it can entail data which is to be protected through fingerprinting/watermarking
-embedded message This is the information which is to be hidden in the cover medium, either as a hidden message from sender A to recipient B, or as a watermark/fingerprint
-stego-medium This is the output obtained once the message has been embedded in the cover medium
-stegokey This is a key that may be used in encoding hidden information within the cover medium; when used for encoding, it is subsequently required afterwards in order to extract the information and is typically a shared secret between the two communicating endpoints

Presently, steganography finds two distinct uses:

-*classical steganography* is the practice of hiding information within a carrier message, such that the resulting output appears trivial to a third party observer

-*invisible digital watermarking/fingerprinting* is the method of fingerprinting media (i.e. adding information in order to enable the originator to trace copies after initial distribution of the media) or watermarking it (i.e. adding information in order to assert ownership over the media)

In the former case, the focus is on *secrecy*; the objective is to exchange data between entities A and B, such that it is difficult for entity C intercepting the message to detect or decipher the hidden content. For this application, both the originator A and the recipient B of the message have an equal interest in maintaining a high fidelity reproduction of the stego-medium (SM), since it is used as a communication channel between them and hence both entities require fidelity with respect to the original. A further objective is achieving maximum data rate in transmitting the embedded message (EM) via the cover medium (CM), since the SM is in this case only used as a transmission channel and is therefore contextually irrelevant. Correspondingly, *robustness* of the embedded message against deletion becomes secondary in importance. Still, it is possible for C to attempt to modify the SM in order to inhibit communication between A and B and techniques will be presented later that address this issue.

In the case of invisible digital watermarking and fingerprinting, the principal objective is *robustness*. The originator (A) of the content wishes to encode either ownership information via the EM within the media, or information that ties the SM to its recipient (B) such that further distribution of it by B can be traced back to B. Simultaneously, it is often in B's interest to either delete the EM from the SM in the former scenario, or to modify it in a predictable fashion, such that the resulting SM contains a forged EM that is able to pass validation (e.g. so that it can be proved that it is owned by B, or some alternate third party specified by B). In both of these cases, the interests of A and B are asymmetric and hence it is A's intent that a technique be applied in generating the SM in a fashion which makes it impervious to modification or deletion by B. In this case, robustness is placed above data rate in importance, as the data transmitted merely needs to contain enough information to tie the underlying CM to A and/or B. Therefore, the relative levels of importance of the CM and the EM are reversed, relative to classical steganography. The CM itself becomes predominant, while the importance of the EM is reduced to guaranteeing traceability and ownership of the CM. Given the importance of the CM itself, a further requirement is that relatively high fidelity be maintained in the SM, relative to the CM. Specifically, since typical applications include digital audio and video transmissions, it is desired that differential changes from the CM to the SM be subtle enough so as not reduce the *perceived* quality/value of the CM[3].

The primary objective of classical steganography is information *hiding*, given a presentation oriented context. By its nature, it therefore exploits deficiencies in various aspects of human perception, which by nature are conditioned to be selective in information processing. Digital media steganography has received much recent attention from academia as well as the industry:

> The main driving force is concern over copyright; as audio, video and other works become available in digital form, the ease with which perfect copies can be made may lead to large-scale unauthorised copying, and this is of great concern to the music, film, book, and software publishing industries.[4]

As the bit resolution of digital media increases, it eventually reaches a threshold beyond which further subdivisions in resolution are no longer detected by human perception. As an example, a technique of information hiding called least significant bit insertion makes use of this limit. Other techniques rely on similar phenomena to insert noise below the threshold of perceptibility into the data; these will be described in more detail.

Distributed media manifests itself in three categories: digital audio data, bit mapped still images, and digital video. One significant factor in audio and visual presentation is that the human auditory tract has a superior performance when compared to that of the visual system; hence, transformations on audio data are often more readily discerned in a subjective assessment of the underlying encoding technique, compared to visual encoding using the same algorithm[1]. Furthermore, still images can be regarded as non-causal data, while audio and video streams are often comprised of real-time multicasts, which are causal in nature. As such, algorithms must be optimized given an application-specific context. Methods of watermarking and known attacks are next presented for each.

Bit mapped Still Images

Still images are non-causal in nature, meaning they are not dependent on the flow of time. The advantage of this is that the entire data set comprising the image can be accessed and relationally modified at any given point in time. The same holds true for pre-recorded audio and video streams. In comparison, live broadcasts of audio and video are causal in nature, since accessibility of the data stream is in those cases limited to what is being broadcast at the present time; future information is not yet available and hence cannot be pre-processed. Given the nature of non-causal data, it is clearly more suitable for deterministic modification; that is, the EM can be embedded given a priori knowledge of the data set, instead of requiring heuristic data estimation.

Most of the research on steganography has centered on this type of

media, so it will be treated in greater detail here than the other types. Furthermore, some of the basis of still image steganography can be reused in causal video steganography, where context is – after differential encoding - treated on a frame-by frame basis. Hence, similarities will be pointed out during the discussion of video encoding.

Least Significant Bit (LSB) Insertion is a technique whereby the CM is broken up into bitwise components; subsequently, for each individual pixel value in the EM, the LSB is changed to correspond to successive bits of the EM bit map. Given a 24 bit image map as the CM (using 8 bits for each of the three primary colors), it is possible to embed the data in the two LSBs of the original, while maintaining fidelity with respect to the original[5]. The result is a bit rate of 6 EM bits per 24 CM bits, or in other words, four CM bytes per each encoded EM byte. This comparatively high bit rate is diminished by one major factor: the encoding is only valid as long as the transmitted SM is left unmodified. Even the simplest image transformation (e.g. change in image brightness) can result in an inversion (in the case of 1 LSB modification) or some other permutation (in the case of 2 LSB modification) when the EM is extracted from the SM. Further image transformations (e.g. scaling/rotation/cropping) completely, or at least partially negate the inherent EM content. Therefore, LSB insertion in conjunction with prior encryption of the EM (which results in message secrecy even if the SM is intercepted and decoded) is suitable for classical steganography, where SM transformation is presumed to be of no concern.

A suitable form of encryption can also be utilized to ascertain that the message has not been tampered with, as subsequent decoding of the EM would otherwise result in garbage. Lastly, because even LSB insertion adds “noise” or pseudo-randomness (caused by the EM), a CM with a lot of high frequency content is desirable, since large, uniformly colored areas typically give away tampering to the naked eye, even when only slightly modified.

Spread Spectrum Image Steganography (SSIS) is a step up from LSB insertion. Here, a carrier signal with its fundamental frequency determined by a pseudo-random number generator with known initial seed point (secret key) is modulated by the EM. In other words, the EM is distributed uniformly across the entire frequency band in a semi-random fashion. The seed point is known by A only in the case of watermarking/fingerprinting, or by A and B for steganography and is required to initialize the pseudo-random number generator. Extraction is performed by filtering the SM with a noise reduction filter whose threshold is set just below the noise level added by the SSIS generator (which requires a priori knowledge of the SSIS noise threshold); this results in an estimate of the original CM, which is then subtracted from the SM to derive the EM.

With a proper choice of the seed point and a generator exhibiting a uniform distribution, it becomes a hard problem to extract the EM without knowledge of the seed value. If the EM is encrypted prior to encoding, this further increases the difficulty of decoding and subsequent message extraction. Furthermore, redundancy can be added to the EM prior to encoding (e.g. through

error correction or parity/cyclic redundancy checks), so that the amount of noise generated by the encoder can be set below a desired threshold (typically so that the resulting SM exhibits a low overall noise content, which makes encoding difficult to discern by visual inspection). Thus, SSIS can be used for both information hiding (fingerprinting/watermarking), as well as secret information transmission (steganography). Since the EM is encoded in the spectral domain, attacks must either alter the frequency map of the image (e.g. by breaking it into pieces, or by non-linear surface transformations of the image, which also alter spectral coefficients), or they must destroy the EM content by using noise reduction filters whose threshold lies below that of the noise level added by the SSIS encoder (which would subsequently result in an estimate of the original CM, sans the EM).

With a proper choice of redundancy (and low noise level) and encryption of the EM prior to encoding, it is possible to obtain a level of robustness that is acceptable, with respect to other known steganography techniques. Notably, if SSIS is used for watermarking or fingerprinting and the EM is sufficiently small, the technique may be robust against image cropping attacks, provided either that the resulting crop retains a sufficient portion of the EM for unique identification, or that the watermark is inserted in multiple locations in the CM and is thus retained in its entirety in at least one of the cropped parts. In this case, to obtain the EM from the altered image, a corresponding crop of the original CM is subtracted from the estimate after the cropped CM is filtered. The data rate is variable, based on the redundancy added to the EM prior to encoding, as well as the desired noise threshold inherent within the SM.

As noted above, a universal benchmark is a non-linear image transform, e.g. using StirMark. It is shown, e.g. In an attack against Digimarc(TM)'s watermarking technology (which is based on SSIS), that this technique is able to successfully render EM extraction difficult or impossible[7].

Orthogonal Projection (OP) Coefficients Manipulation is a generalized version of the discrete cosine transform (DCT) coefficients manipulation[8]. Whereas the latter relies on a DCT of the image prior to EM embedding, OP uses a pseudo-random generator with a seed value (given as a key, known to A only, or A and B) to generate a set of n vectors, which are subsequently orthonormalized using the Gramm-Schmidt process. Similar to the technique involving the DCT, the image is then projected onto these basis vectors, resulting in n vector coefficients. It should be noted that the coefficients are entirely dependent on the vector set created and as such, a different seed point will result in a different set of coefficients, given constant n. Because of the orthonormality of the basis set and given a value of n sufficiently large, a subsequent inverse transform will yield the original set of bitmap values.

After the initial transform, the largest m OP coefficients (where m represents the number of EM bits) are selected and modified by adding m EM bits (one EM bit per OP coefficient), scaled by a factor which is bound by the desired noise threshold of the output signal. This is in principle similar to the SSIS technique, with two differences: first, an arbitrary basis set is selected

instead of the basis set comprising the frequency spectrum; secondly, instead of uniformly spreading the EM throughout the entire spectrum using a carrier with a pseudo-random frequency, the largest m OP coefficients are chosen and modified with the EM information. The key difference is that in the former case, the secret key is used to initialize the pseudo-random generator, while in the latter it is chosen to initialize the set of basis vectors. Given that different keys result in OP coefficients of varying magnitude and that optimum noise reduction in EM encoding results when a set of high amplitude OP coefficients is chosen, it is thus possible to select an appropriate key which optimizes coefficient amplitude given a particular CM. For extraction, a similar noise reduction technique as that in SSIS is used to estimate the original CM. This is subsequently subtracted from SM to obtain EM. Once again, the proper basis set must be generated to produce the right set of OP coefficients, which can then be used to regain the embedded EM bits.

Unlike the technique presented, DCT assumes a basic DCT transform, which prohibits the creation of a secret basis set. As such, the DCT requires EM encryption prior to encoding, to make steganography feasible. Nevertheless, given an encrypted EM, both techniques are suitable for steganography and furthermore, given their relative resilience to a variety of attacks, to watermarking and fingerprinting as well.

Texture Block Coding works "by copying a region from a random texture pattern found in a picture to an area that has similar texture"[10]. For extraction, the image is auto-correlated with itself, which shows peaks at the higher correlation points corresponding to the copied pixel map data. Subsequently, the original content is shifted and subtracted from these regions to retain the mask. Since masked regions contain the same pixel data as the region where the pixel map originated, this encoding technique shows a high level of robustness against geometric image transformations, as well as general color manipulations and even compression. However, one attack which destroys EM content is image cropping – such as when the image is subdivided into pieces, which are assembled for presentation.

Patchwork relies on the fact that, in probabilistic terms, the expected value of the difference between two randomly chosen points in an image is 0 since, given enough sample points, adding up successive differences eventually tends to zero as there are enough positive and negative components to cancel out in the limit (i.e. $E|a-b| = 0$). Patchwork uses a pseudo-random number generator with a seed point given by an encoding key which selects a pair of starting points in the pixel map. One of these points is darkened by an amount chosen from a Gaussian random distribution with a typical weight of 1-5 out of a possible 256 shades, while the other pixel is correspondingly brightened. This is repeated about 10,000 times for a given image. The end result is that the expectation of the of difference in pixel value (mentioned above) is artificially raised to a non-zero value. This does not convey inherent information, beyond merely showing the presence that patchwork was applied to the image. Hence, the bit rate is very low. However,

the encoding is robust against attacks such as cropping (where accuracy is reduced logarithmically, with decreased image size) or image tone manipulations. A known attack is geometric image transformation (e.g. Rotation or scaling), since the starting points derived through the secret key which initializes the random number generator no longer match the points in the transformed image. These attacks can be negated however, if original image size/orientation is encoded via a masking function. Other attacks include synchronization attack (where image content is shifted around, thus displacing the pixels marked by patchwork), or attack via StirMark (where the image is non-uniformly scaled, thus once again displacing the original pixels). Due to their non-linearity, such attacks are immune to any image restoration content mentioned above.

Digital Video

In contrast to still images, digital video typically entails causal transmission; that is, the signal is often broadcast in real time, while it is being generated. As such, a priori information of the data is not available and instead, hidden content must be encoded on the fly, as individual frames pass through the encoder. Because of this, video frames must be processed independently and temporal relationships between frames can thus only be taken into account if a sufficient delay between frame generation and broadcast time is introduced.

Discrete Cosine Transform (DCT) Coefficient Manipulation is the technique most commonly used for encoding video. In video communications, efficient media compression is typically achieved through differential encoding – that is, by encoding the difference between adjacent video frames, since they typically exhibit a high level of correlation and hence the difference can be described using a lower level of information. During initial encoding, a DCT is used to transform the video image into the spectral domain – thus obtaining an invertible set of DCT coefficients. As described above in OP coefficient manipulation of still images, these spectral coefficients can be modulated using the EM such that the resulting output displays a slightly higher noise threshold, relative to the original input. However, since the frame content changes several times each second (even with no camera movement), this additional noise becomes analogous to static in content free video recording and thus can be negligible provided that the noise level is kept below a given threshold. This is in contrast to still image encoding, where, as an example, uniformly colored surfaces would much more readily give away image manipulation.

In differential coding, the differentials are transformed via DCT and are then subsequently modified by the EM signal. If frames were encoded on an individual basis, an eavesdropper seeking to extract the encoding key or establishing the encoding algorithm could use a statistical attack on images from the same scene that have a high correlation in order to attempt to determine the encoding key or algorithm. Since only the differentials of consecutive frames are

transmitted, this possibility is negated as differential information exhibits little to no correlation content. Therefore, adding information to differential information renders the output statistically equivalent to a set of output data with the same noise threshold. Consequently, differential coding is most appropriate when causal data is to be watermarked.

If the entire video stream is available a priori, watermarking can be performed over the entirety of the data set. In other words, rather than encoding the EM on a frame-by-frame basis, the stream can be processed as a whole. This would also entail a DCT, resulting in coefficients which would subsequently be modified and inverse-transformed to reproduce a copy of the original with additional noise content resulting from the pseudo-random EM encoding. With a large data set (e.g. a long duration video stream), it becomes possible to add a large amount of redundancy or to retain a sufficiently low noise floor. Alternatively, with a differentially encoded a priori video stream, the DCT coefficients of the differentials can still be modified as a whole, thus reducing overall added noise content.

Digital Audio

The human auditory system (HAS) exhibits greater performance when compared to the human visual system (HVS):

> The HAS perceives over a range of power greater than one billion to one and a range of frequencies greater than one thousand to one. Sensitivity to additive random noise is also acute. The perturbations in a sound file can be detected as low as one part in ten million (80 dB below ambient level).[10]

The most common techniques of data hiding in audio signals will be discussed next.

LSB Insertion is a technique that can be applied to audio data as well, to encode an EM. The same limitation apply as in still image and moving image coding: even slight data transformation (e.g. amplitude variation or linear filtering) can easily negate the embedded signal. The bit rate is high relative to other techniques and thus this method can be used for classical steganography, given additionally encrypted content. Compared to image processing, a comparatively large amount of noise is introduced, due to the HAS' greater noise sensitivity compared to that of the HVS.

Phase Coding takes advantage of the fact that the HAS is sensitive to changes in the phase of an audio system, but not to absolute phase. Hence, it is possible to take segments of an audio track and embed the EM within the initial phase of each of these segments, while retaining other phase relationships in each segment. These phase changes can sometimes be detected by experts in the professional audio industry, but are generally imperceptible by the average listener. As such, phase coding introduces relatively little signal distortion. In the

paper published by Bender et al., a useful bit rate of 8-32 bps was achieved, depending on the amount of noise in the CM as well as the size of the chosen audio segments (ranging from 32-128 frequency slots).

Spread Spectrum encoding can be used in audio coding as well, analogous to the still image case. In this case, a pseudo-random signal is generated which has a uniform frequency distribution and maximal key length (i.e. it has a long non-repetitive sequence length). This carrier is modulated by the EM and subsequently added as white additive noise to the CM to produce the SM. During decoding, the key is used to initialize the random vector that modulates the carrier wave. It is subsequently demodulated from the carrier and the input EM is retained from the phase offsets (where the initial phase given by the random vector may for example constitute a zero, while given by the RV plus 180 degrees may constitute a 1). Assuming that the noise threshold is kept to .5% of the overall CM amplitude, the noise content of the SM is kept negligible, relative to the CM. This method produces an average EM stream of 4 bps[10].

Echo Data Hiding inserts echo into the signal, using two different delays constituting on/off bits. It has been demonstrated that adding echo with a sufficiently low delay and amplitude is imperceptible to average listeners. This technique breaks the signal into segments, which are time- and amplitude-shifted, and added to the SM. The shifts in delay are based on the EM bits, while the amplitude is held constant. To extract the EM from the SM, the signal is again broken into segments and an autocorrelation is performed on the cepstrum of the segments. Since the signal is echoed once in the segment examined, the spectrum (which is the logarithm of the spectrum) will reflect this via a peak which corresponds to the delay of the echo. An autocorrelation of the cepstrum will therefore turn up the delay constant, which in turn is used to decode the EM. Since sufficiently short delays and amplitudes are inaudible, this is a useful technique for hiding data in audio content. Furthermore, this technique is the only one which can resist a jitter (synchronization) attack[11].

Conclusion

A number of techniques are available to code digital audio and video streams, as well as still images. The technique used is generally dependent on the carrier medium and application. Classical steganography places data rate over robustness and hence often a technique as simple as LSB insertion can be applied to encode a secret message within a data stream. On the other hand, for watermarking or fingerprinting, greater robustness is typically demanded, to reduce the possibility of attack of the embedded message (which constitutes the fingerprint or watermark) after initial distribution of the carrier medium. Simultaneously, causality vs. non-causality plays a role in determining how the EM is to be encoded, and audio data also demands a different level of encoding from visual data, due to the human auditory system's greater performance when

compared to the human visual system.

1. Matteo Fortini, Laboratory of Advanced Research on Computer Science, University of Bologna. "*Steganography and Digital Watermarking: a global view*", http://lia.deis.unibo.it/Courses/RetiDiCalcolatori/Progetti00/fortini/web/cover.html.
2. Merriam-Webster's collegiate dictionary (10th ed.). (1993). Springfield, MA: Merriam-Webster.
3. F. Petitcolas, R. Anderson, "*Weaknesses of copyright marking systems*", Proceedings of the Multimedia and Security Workshop at ACM Multimedia '98, pp. 55-62, Bristol, United Kingdom, September 1998.
4. F.A.P., R.J.Anderson and Markus G. Kuhn, University of Cambridge Computer Laboratory, Security Group, "*Information Hiding – A Survey*", http://www.cl.cam.ac.uk/~fapp2/publications/ ieee99-infohiding.pdf.
5. Marcia Ramos, Visual Communications Lab, Cornell University, "*Activity Selective Image and Video Coding*", http://foulard.ee.cornell.edu/marcia/research.html.
6. Neil F. Johnson, Sushil Jajodia, George Mason University, "*Exploring Steganography: Seeing the Unseen*", IEEE Computers, February 1998, pp. 26-34.
7. Neil F. Johnson, Sushil Jajodia, Center for Secure Information System, George Mason University, *"Steganalisys of Images Created Using Current Steganography Software"*, http://isse.gmu.edu/~csis.
8. F.A.P. Petitcolas and R.J. Anderson, "*Evaluation of Copyright Marking Systems*". In IEEE Multimedia Systems (ICMCS'99), vol. 1, pp. 574-579, (Florence, Italy), June 1999, http://www.cl.cam.ac.uk/~fapp2/publications/ ieeemm99-evaluation.pdf.
9. Andreas Westfeld, Gritta Wolf. "*Steganography in a Video Conferencing System*." pp. 32–47 in David Aucsmith (Ed.): Information Hiding. Second International Workshop, IH'98, Portland, Oregon, USA, April 1998, Proceedings, LNCS 1525, Springer-Verlag Berlin Heidelberg 1998, http://os.inf.tu-dresden.de/~westfeld/publikationen/ihw98slides.pdf.
10. S. K. (editor) and F. P. (editor). "*Information hiding techniques for steganography and digital watermarking*." In Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, 2001.
11. Fabien A. P. Petitcolas, Ross J. Anderson, and Markus J. Kuhn, "*Attacks on copyright marking systems*", 2nd Information Hiding Workshop, 1998