Combating Digital Document Forgery using New Secure Information Hiding Algorithm

Abbas Cheddad, Joan Condell, Kevin Curran and Paul Mc Kevitt
School of Computing and Intelligent Systems, Faculty of Computing and Engineering
Londonderry, BT48 7JL, Northern Ireland, United Kingdom
Emails: {cheddad-a, j.condell, kj.curran, p.mckevitt}@ulster.ac.uk

Abstract

The recent digital revolution has facilitated communication, data portability and on-the-fly manipulation. Unfortunately, this has brought along some critical security vulnerabilities that put digital documents at risk. The problem is in the security mechanism adopted to secure these documents by means of encrypted passwords; however, this security shield does not actually protect the documents which are stored intact. We propose a solution to this real world problem through a 1D hash algorithm coupled with 2D irFFT (irreversible Fast Fourier Transform) to encrypt digital documents in the 2D spatial domain. This method is described elsewhere in [1]. Further by applying an imperceptible information hiding technique we can add another security layer which is resistant to noise and to a certain extent JPEG compression. We support this assertion by showing a practical example which is drawn from our set of experiments. Our proposal not only points out forgery but also lets a jury or a forensics expert have access to the original document despite being manipulated. This would undoubtedly be very useful in cases of disputes or claims.

1. Introduction

Traditionally, document's forgery was carried out mechanically, however, since the recent boost in communication technology, the massive increase in databases storage and the introduction of the concept of e-Government, documents are more and more being stored in a digital form. This goes hand in hand with the aim of the paperless workspace, but it does come at the expense of security breaches especially if the document is transmitted digitally over a network. Document forgery is a worry for a range of organisations, i.e., Governments, Universities, Hospitals and Banks. The ease of digital document reproduction and manipulation has certainly attracted many eavesdroppers.

Relational Database Management Systems (RDBMS) secures scanned documents through the use of a password to the database. This means that scanned documents are stored with an encrypted string password. The main issue here is if a hacker is able to crack the password then he may be able to modify any document digitally and logs out as if nothing has happened. In July 2005, it was discovered that a number of Second World War files held at The National Archives contained forged documents. An internal investigation found that the forgery took place during or after the year 2000 [2].

In this paper we propose a highly robust protection scheme which protects scanned documents from forgery. The scheme is based on an information hiding technique, Steganography, which is the science that embeds data in a digital medium in an imperceptible manner. The advantage of this technology over the well known technique of Cryptography is that no one knows it is there, hence the name "hidden". A number of Steganographic methods have been introduced; however, very few authors have applied Steganography to real world problems. Hence, our objective is to put into context a practical application of our ongoing research on enhancing Steganography in digital images that could solve one of those problems. Our proposed algorithm is efficient, highly secure and robust against different image processing attacks.

2. Methodology

The fundamental concept of our proposal is to embed the secret message in the 1st-level 2D *Haar* DWT (Discrete Wavelet Transform) with the symmetric-padding mode. DWT is a well known transformation that gained popularity among the image processing community especially those who are dealing with image compression. Its applications in different areas is growing however (note that JPEG2000 uses DWT to compress images). 2D DWT provides a decomposition of the approximation, and the details in three orientations (horizontal, vertical, and diagonal) by means of a

978-1-4244-2917-2/08/$25.00 ©2008 IEEE

convolution-based algorithm using High and Low pass filters. In our case we compute four filters associated with the orthogonal or bi-orthogonal of the *Haar* wavelet.

We choose Wavelet over DCT (Discrete Cosine Transform) because [3]: the Wavelet transform understands the Human Vision System (HVS) more closely than does DCT; Visual artefacts introduced by wavelet coded images are less evident compared to DCT because the wavelet transform does not decompose the image into blocks for processing. DFT (Discrete Fourier Transform) and DCT are full frame transforms. Hence any change in the transform coefficients affects the entire image except if DCT is implemented using a block based approach. However DWT has spatial frequency locality, which means if the signal is embedded it will affect the image locally. Hence a wavelet transform provides both frequency and spatial description for an image. More helpful to information hiding, the wavelet transform clearly separates high-frequency and low-frequency information on a pixel-by-pixel basis [4].

We refer to data that we wish to embed as payload, herein the image itself. Since we need means for protecting scanned documents against forgery it is essential that the payload will carry as much information from the host (cover image) as possible. There is a trade-off between perceptual visualization and space demand for embedding (usually measured in bits). Without taking compression into account, the payload can be consistent with the cover signal; therefore, if the cover is stored as an 8-bit unsigned integer type then the payload will require 8 templates when applying the one bit substitution method. There is a high payload Steganography approach called *A Block Complexity Data Embedding* (ABCDE) [5], but it is prone to statistical attacks as it acts in the spatial domain; moreover it cannot resist any kind of manipulation to the Stego-image (image having embedding data).

An approximation of the cover document can be achieved through applying the gray threshold technique which results in a binary image demanding only 1 bit per pixel for storage. Some authors suggest using an edged image instead as it approximates the cover better. In the search for the best way to represent the cover image with the least bit requirement for embedding we identified dithering as our ultimate pre-processing step which is the foremost task in building our system. Dithering is a process by which a digital image with a finite number of gray levels is made to appear as a continuous-tone image [6].

Despite, in all versions, each pixel takes on only one bit, it is apparent that the way dithering quantizes image pixels contributes a lot to the final quality of data approximation. We observed that thresholding performs better in text based documents, while in capturing graphics it is proven to be a poor performer compared to dithering. Therefore, since our aim is to produce a general workable prototype we have to take into consideration the presence of both text and graphics; subsequently we opt to use dithering.

2.1 The embedding stage

Manipulating coefficients in the wavelet domain tends to be less sensitive unlike other transformations such as DCT and FFT. There are two methods to convert decimal integer to a binary string: one is to use the conventional decimal to binary conversion and the other is termed The *Binary Reflected Gray Code (BRGC)*[1]. This binary mapping is the key to the augmented embedding capacity introduced by *ABCDE* algorithm. There is a trade-off between robustness and distortion which is summarized in Figure 1.

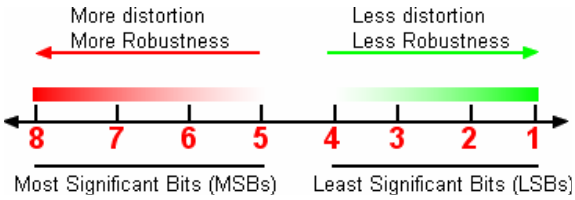

Figure 1. An 8-bit (1 byte) representation with the conventional integer to binary conversion. It is clear that choosing the right index for embedding is very crucial. This intricacy is less severe when using the RGC since it produces seemingly disordered decimal-to-binary representation.

A trade-off occurs during our algorithm's formulation which is due to the different levels wavelets can have. The lower we go the more robust we get but with less capacity for embedding. For example if the cover image is of size 255x255 (8-bit grayscale) we obtain 16384 bits to embed in the first level, the second level will reduce the image dimensions by a factor of two to yield 4096 bits and so on. In some cases the inverse transform in the wavelet domain truncates some values that fall larger or lower than the allowed limits in 8-bit type of images, the truncation occurs because of the introduction of "non-natural bits" coming from the secret message while embedding. To cope with this rare problem we choose to transform the RGB image into the YCbCr colour space prior to feeding it into DWT where we embed in the chrominance red channel (Cr). This step ensures that there will not be any data lost. Our proposed design is illustrated in Figure 2.

[1] http://mathworld.wolfram.com/GrayCode.html. Accessed on 10-06-08, at 11:42.

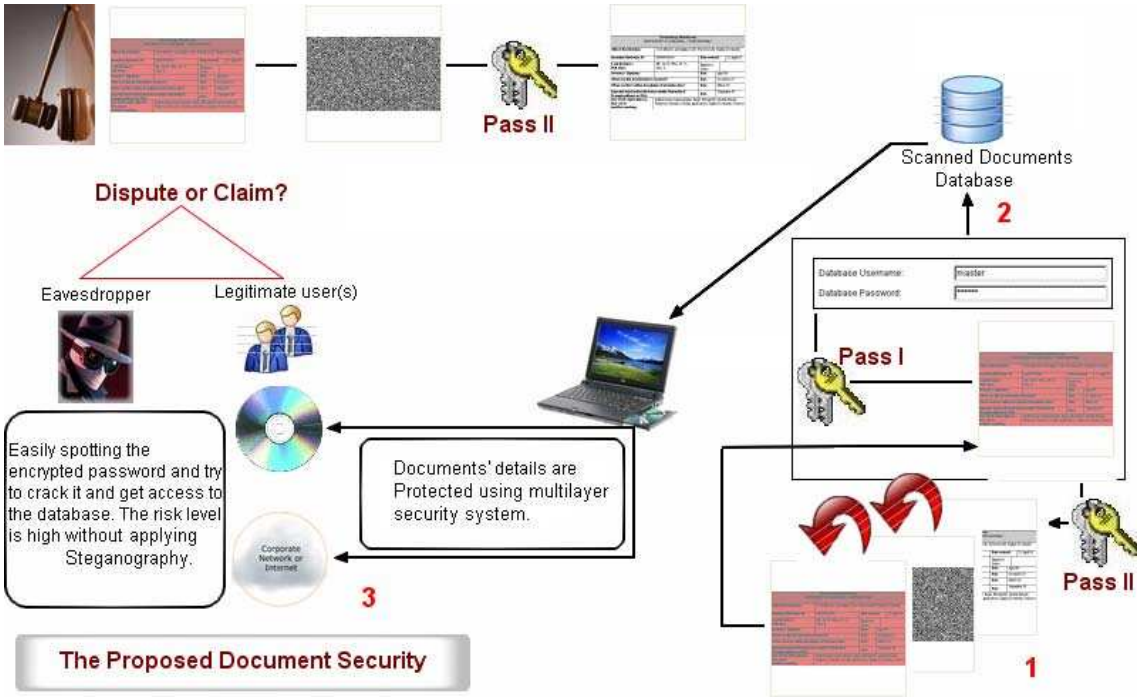

Figure 2. A general graphical scheme showing the advantage of adopting our algorithm for securing scanned documents archives.

3. Related Work

Popescu [7] shows a comprehensive investigation carried out on image forensics which aims to detect forgery by means of the preserved natural image statistics. Although, they seem to have successfully created a system whereby image forgery can be detected however our proposal goes beyond that by showing what the original 'non-forged' image looks like. We believe in some cases, for instance in court, it is not sufficient to tell just that the image/document has been modified (which can be caused by colour changes) without giving the jury a tool to actually extract the original document.

4. Conclusion

In this paper, we proposed a scanned document forgery detection method which uses an information hiding technique that is highly secure, efficient and robust to various image processing attacks. The experiments we carried out, of which a sample is reported in this paper, show promising results. The results show that the system can be relied on to combat scanned document forgery. Future work for this tool will involve tackling the problem of image compression (below 75%) and how to overcome it.

References

[1] Cheddad.A., Condell. J., Curran, K. and Mc Kevitt P. (2008). "Securing Information Content using New Encryption Method and Steganography". 3rd IEEE International Conference on Digital Information Management ICDIM08. London, UK.

[2] The national Archives (UK government's records and information management), available from: <http://www.nationalarchives.gov.uk/news/stories/195.htm?homepage=news>. Retrieved on 09-06-2008 at 11:23.

[3] Potdar V. M., Song Han and Chang E. (2005). "A Survey of Digital Image Watermarking Techniques". 3rd IEEE International Conference on Industrial Informatics (INDIN), pp: 709-716.

[4] Raja K. B., Vikas, Venugopal K. R., and Patnaik L.M., (2006). "High Capacity Lossless Secure Image Steganography using Wavelets". International Conference on Advanced Computing and Communications, ADCOM 2006, pp: 230-235.

[5] Hioki Hirohisa. (2002). "A Data Embedding Method Using BPCS Principle With New Complexity Measures". Proc. of Pacific Rim Workshop on Digital Steganography 2002 pp.30—47 Jul.

[6] Farid H. Fundamentals of Image Processing. Available from: <http://www.cs.dartmouth.edu/farid/tutorials/fip.pdf>. Tutorial, pp: 61. Accessed on 09-06-2008 at 10:00.

[7] Popescu A. C., (2005). "Statistical Tools for Digital Image Forensics". PhD thesis. Department of Computer Science, Dartmouth College Hanover, New Hampshire, USA.