Digital Investigation 9 (2013) 246–257

Contents lists available at SciVerse ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

JPEG steganography detection with Benford's Law

Panagiotis Andriotis*, George Oikonomou, Theo Tryfonas

Crypto Group, University of Bristol, Faculty of Engineering, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

ARTICLE INFO

Article history:
Received 29 October 2012
Received in revised form 23 January 2013
Accepted 25 January 2013

Keywords:
Steganalysis
Generalized Benford's Law
Steganography detection
Data hiding
Quantized DCT coefficients

ABSTRACT

In this paper we present a novel approach to the problem of steganography detection in JPEG images by applying a statistical attack. The method is based on the empirical Benford's Law and, more specifically, on its generalized form. We prove and extend the validity of the logarithmic rule in colour images and introduce a blind steganographic method which can flag a file as a suspicious stego-carrier. The proposed method achieves very high accuracy and speed and is based on the distributions of the first digits of the quantized Discrete Cosine Transform coefficients present in JPEGs. In order to validate and evaluate our algorithm, we developed steganographic tools which are able to analyse image files and we subsequently applied them on the popular Uncompressed Colour Image Database. Furthermore, we demonstrate that not only can our method detect steganography but, if certain criteria are met, it can also reveal which steganographic algorithm was used to embed data in a JPEG file.

1. Introduction

The use of several means of covert communication is appealing among individuals or groups that are interested in securing the content of an exchange concealing the act of their interactions. Steganography is one of the methods which have been introduced in order to hide information and covertly spread hidden data through public channels without causing suspicion. JPEG images constitute a widely used medium of secret communication, partially thanks to the fact that they can be produced by any camera, smartphone or image processing tool and can be easily exchanged between a variety of applications (McBride et al., 2005).

Steganography aims to transport a message in a hidden fashion by embedding it in a transport medium called a carrier (Fridrich et al., 2001). The grouping of the carrier with the secret message is known as a stego medium or stego cover. The detection of steganographic algorithms and techniques can be a hard task, even more so if the secret data are encrypted with a stego key. Steganalysis is the process of attacking and breaking steganographic methods, either by simply detecting the presence of a secret message or by extracting and potentially destroying it (Chandramouli et al., 2004). The success of a steganalytic method can be quantified either by the accuracy of the prediction of a secret message's presence in a stego object or by the extraction of the hidden information. Steganalysis methods can be further classified into two broad categories: *targeted* and *blind* (or *universal*). In *targeted* steganalysis the attack is mounted against an already known embedding technique. *Blind* steganalysis methods aim to determine whether an object is carrying a hidden message, without any a-priori knowledge.

When the stego carrier is a JPEG image steganalysis is prominently based on two approaches: *visual* and *statistical* attacks (Westfeld and Pfitzmann, 2000; Jolion, 2001). Visual attacks demand long training steps and a significant amount of resources. Statistical attacks are more resource-efficient and as a result, several can be found in the literature (Chandramouli and Subbalakshmi, 2004). These are

* Corresponding author. Tel.: +44 117 33 15740; fax: +44 117 33 15719.
E-mail addresses: p.andriotis@bristol.ac.uk (P. Andriotis), g.oikonomou@bristol.ac.uk (G. Oikonomou), theo.tryfonas@bristol.ac.uk (T. Tryfonas).

1742-2876/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.diin.2013.01.005

based on the fact that the images' histograms or high order statistics get modified after the steganographic techniques take place. Modern blind steganalytic schemes engage supervised learning to differentiate between the plain media and stego images and also distinguish the data hiding algorithm used for steganography (Solanki et al., 2007).

Benford's empirical law of anomalous numbers (Benford, 1938) has been successfully used in the past for fraud detection in the accountancy sector. It has also been demonstrated that the law in a generalized form can be employed to perform a series of forensic tasks on JPEG images, such as the detection of double compression (Fu et al., 2007). This work was limited to grey scale images however. The generalized Benford's Law has been employed for steganalysis elsewhere (Zaharis et al., 2011), but there it was applied on raw byte values and not from an image analysis perspective.

In this context, this paper's contribution is two-fold:

- We adopt the generalized Benford's Law as the basis of a novel statistical attack for blind steganalysis and we provide evidence of its applicability on colour JPEG images.
- We demonstrate that the attack can perform steganalysis very quickly and achieves a satisfactory detection rate.

The proposed attack is based on an analysis of the quantized coefficients of a large amount of colour images. Our method indicates that it is possible to predict the behaviour of the distributions of their significant digits and any disturbances of these distributions can then be considered an indication of the presence of steganography. By studying the deviations of their distributions, we propose a decision making model based on our findings related to the behaviour of digit 2. Moreover, we developed a set of automated tools which implement the attack and can be used to conduct blind steganalysis and thus help forensic analysts to identify suspicious colour JPEG images. In order to validate the method and assess its performance, we used it to analyse files taken from a widely-used database of approximately 1340 images, enriched by our own set created by the use of a smartphone. Our analysis includes comparative evaluation with the open source steganalysis software Stegdetect.

The rest of this paper is organized as follows. In Section 2, we highlight our main motivations and discuss some theoretical background. In Section 3 we present our new detection algorithm. The experimental results are provided in Section 4 and the discussion on the results can be found in Section 5. Finally, in Section 6 we present results from testing our method in various steganalytic tasks. The conclusion is drawn in Section 7.

2. Theoretical background and motivation

The term JPEG comes from the consortium that created the standard (Joint Photographic Experts Group). It is one of the most common formats and it is widely used by all the manufacturers of digital consuming products such as digital cameras. It comes from the need to exchange images through different platforms and applications. The main goal of the JPEG compression is to discard information which is imperceptible to the human eye while leaving unchanged the aesthetic details of the image. Simultaneously, the JPEG compression reduces image data size. A detailed presentation of the procedure followed in order to compress a data stream with the JPEG standard can be found in Wallace (1992). Usually the Discrete Cosine Transform (DCT) encoding procedure consists of six basic steps: Conversion of the representation of colours from RGB (Red, Green, Blue) to YC_bC_r, downsampling of the chrominance values (usually by a factor of two), transformation of values to frequencies (using 8×8 pixel blocks), quantization process, zigzag ordering, lossless compression using a variant of Huffman encoding.

In more detail, an image consists of pixels and each pixel usually has three bytes that represent its three basic colour components: Red, Green and Blue. The first step to the JPEG encoding procedure is to convert these pixel values from RGB to YC_bC_r which is another colour space that has three components. Y represents the brightness of an image and is called luminance while C_b and C_r represent colours and they are called chrominance. It is known that the human eye can recognize the difference in the luminance of an image more easily than the chrominance coefficients (Lee et al., 2006). The type II DCT is responsible for the quantization process. DCT is a mathematical transformation (uses cosine functions) that converts the pixel values of 8×8 blocks to blocks of 64 frequency coefficients. These numbers are critical for our method.

A digital image and especially a JPEG image can be a perfect cover medium because it usually has large amounts of space where one can embed information. There are numerous factors that result in a successful embedding procedure such as the embedding technique and the cover image characteristics (McBride et al., 2005). A general assumption is that the image should be busy, meaning that it should lack large areas of similarities. Popular techniques used to hide information in images are the Least Significant Bit (LSB) and the DCT encoding. Embedding techniques focus on the quantized DCT coefficients and they usually embed data by applying LSB encoding in those coefficients that are not equal to zero. In McBride et al. (2005) we can find a list of tools that use the quantized DCT coefficients to embed data in JPEG images. They rely on the fact that the procedures which follow the quantization phase are lossless and the hidden information can then be obtained. Indicative algorithms from this category are Jsteg, Outguess, JPHide and F5. Those techniques introduce irregularities in the statistics of the quantized DCT coefficients of a colour JPEG image. Our goal is to reliably detect such irregularities.

Statistical attacks aim to determine whether the examined data comply with specific statistical rules that normal image files would follow. A very popular attack is the Chi-squared test which compares the statistical behaviour of a suspected image with the theoretically expected properties of its carrier (Westfeld and Pfitzmann, 2000). Histogram attacks, which can also be classified as statistical, depict disturbances in the distribution of the frequencies of DCT coefficients of a JPEG image. These figures can reveal the