Information Sciences 281 (2014) 211–224

Contents lists available at ScienceDirect

Information Sciences

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

Steganalysis classifier training via minimizing sensitivity for different imaging sources

Wing W.Y. Ng, Zhi-Min He *, Daniel S. Yeung, Patrick P.K. Chan

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

A R T I C L E I N F O

Article history:
Received 7 January 2014
Received in revised form 8 May 2014
Accepted 19 May 2014
Available online 28 May 2014

Keywords:
Steganalysis
Neural network
Quantization table
Robustness

A B S T R A C T

Owing to the ever proliferation of digital cameras and image editing software, a large variety of JPEG quantization tables are used to compress JPEG images. As a result, learning-based steganalysis methods using a pre-selected quantization table for training images degrade significantly when the quantization table of testing images is different from the one used for training. Recognizing that it would be undesirable and not practical to train a steganalysis classifier with all possible quantization tables, we propose an approach that the differences in features extracted from images with different quantization tables are formulated as perturbations of those features. Then we define a stochastic sensitivity by the expected square of classifier output changes with respect to these feature perturbations to compute the robustness of classifiers with respect to perturbations. A Radial Basis Function Neural Network based steganalysis classifier trained by minimizing the sensitivity is proposed. Experimental results show that the proposed method outperforms learning methods such as Support Vector Machine and Radial Basis Function Neural Network without considering feature perturbations.

1. Introduction

Steganography presents a potential security threat to society, in general and corporations, in particular. Embedded messages, named stego messages, are hidden in digital media such as images [1–3], audio [4] and video [5] for secret communication. Steganalysis is a technique used to determine whether a digital media has a stego message or not. Current learning based steganalysis methods consist of two major components: a feature extractor and a classifier. Steganalysis classifiers are trained by a set of images which consist of both clean and stego images. Among different types of digital media, JPEG image is the most widely used digital media on the Internet. Therefore, JPEG is a favorable carrier of steganography. In particular, for most JPEG steganalysis, both training and testing datasets use JPEG images compressed by the same quantization table. When different compression quantization tables are used for training and testing image sets, the performance of the steganalysis classifier degrades significantly [6,7].

With the ever proliferation of digital cameras and image editing software available today, JPEG images on the Internet are compressed by many different quantization tables [8]. Moreover, a growing number of digital camera manufacturers, such as Sony, Nikon and Pentax, adopt variable quantization tables which are computed based on the image content dynamically. By

* Corresponding author. Tel.: +86 20 39380285.
E-mail addresses: wingng@ieee.org (W.W.Y. Ng), zhimin.he87@gmail.com (Z.-M. He), danyeung@ieee.org (D.S. Yeung), patrickchan@ieee.org (P.P.K. Chan).

http://dx.doi.org/10.1016/j.ins.2014.05.028
0020-0255/

using the quantization table for training to re-compressing an image, which is compressed by an unseen quantization table, cannot recover an image directly. Extra quantization step will change the steganalysis features of the JPEG image [9]. Therefore it would be unreasonable or impracticable to assume prior knowledge of compression quantization table of an unseen image examined for stego message.

In the real world situation, both images and quantization tables could be different from those used for training the classifier. In addition to quantization table difference, the difference in image content will also affect the performance of steganalysis [10]. The extracted steganalysis feature values will be different from those of the training images [7]. These differences could be treated as perturbation in features and are unavoidable. So, the robustness of steganalysis classifier with respect to feature perturbations is essential to its performance.

Current steganalysis methods make use of off-the-shelf classification methods such as neural network [11,12], Support Vector Machine (SVM) [13,14], dynamic evolving neural fuzzy inference system [15] and ensemble of classifiers [16,17]. However, none of them addresses the issue of perturbation between training and testing images.

In this work, we propose a Localized Generalization Error Steganalysis classifier (LG-Steganalyzer), which is robust to images compressed by quantization tables different from that of training images. Such a situation is unavoidable in real-world applications. A Radial Basis Function Neural Network (RBFNN) is trained via a minimization of a training error and a stochastic sensitivity. RBFNN is selected because of its fast training speed in the presence of large data which is important for dealing with network security problems. The stochastic sensitivity is proposed to capture the influence of feature perturbation created by changes in JPEG quantization table of testing images with respect to the RBFNN classification. With the proposed steganalyzer training method, the LG-Steganalyzer provides a better robustness of steganalysis in real-world applications. Major contributions of the LG-Steganalyzer to steganalysis are:

1. RBFNN trained by the LG-Steganalyzer is robust to real-world situations, e.g. difference in quantization tables and difference in content between training and testing images.
2. The proposed LG-Steganalyzer could be used with any compression quantization table, and any steganalysis feature extraction technique.

This paper is organized as follows: Section 2 provides a brief introduction on steganalysis and JPEG quantization tables. The LG-Steganalyzer is described in Section 3. Experimental results are presented in Section 4. Section 5 concludes this work.

2. Steganalysis and quantization table

We first provide a brief introduction to current steganalysis methods in Section 2.1. Section 2.2 discusses the importance of quantization tables in steganalysis. Section 2.3 demonstrates input perturbations caused by changes in quantization tables and image contents.

2.1. Steganalysis

In [18] at least one hundred and ten steganographic and steganalysis tools were reported in the year of 2007. Today the number is still growing. Therefore it is desirable to train a steganalysis classifier to work with multiple types of steganographic methods. The current trend is to adopt off-the-shelf classifiers with different feature extraction methods proposed for JPEG steganalysis. Shi et al. [19] suggested extracting Markov features based on differences among neighboring Discrete Cosine Transform (DCT) coefficients using transition probability matrices. Steganalysis based on features extracted by Markov approach using the DCT and the Discrete Wavelet Transform (DWT) coefficients was proposed in [20]. Pevny et al. [14] proposed to extract features based on the average of Markov feature in four directions and combine them with DCT features for multiclass JPEG steganalysis. They also proposed an ensemble method to deal with compressions of different quantization tables. Each base classifier is trained for a particular quantization table. Image is classified by the base classifier trained by the quantization table being most similar to the one compressing the image. Its performance depends on both selection of the base classifiers and the robustness of the base classifiers with respect to unseen quantization tables. In [21] Kodovsky et al. showed that the calibration method in [14] yielded negative effects for several steganographic schemes and thus proposed a Cartesian Calibration method to extract CC-Pevny features. Chen and Shi [13] proposed steganalysis features based on both intra-block and inter-block correlations among DCT coefficient difference matrix of images.

2.2. Quantization table in steganalysis

JPEG compresses a raw image by dividing each DCT transformed 8x8 block of pixels by a quantization table (the Quantization box in Fig. 1). Hidden message is then embedded into a JPEG image using a steganography method (the steganography box in Fig. 1). There are 100 quantization tables provided by the Independent JPEG Group and each corresponds to a distinct compression quality, a.k.a. a standard quantization table. We denote quantization tables other than these 100's as non-standard quantization tables.