J. Vis. Commun. Image R. 24 (2013) 846–856

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Block-based image steganalysis: Algorithm and performance evaluation

Seongho Cho [a], Byung-Ho Cha [b,*], Martin Gawecki [a], C.-C. Jay Kuo [a]

[a] Ming Hsieh Department of Electrical Engineering and Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089-2546, USA
[b] Convergence S/W Lab, Samsung Electronics Co. LTD., Suwon-City, Gyeonggi-Do 443-742, Republic of Korea

ARTICLE INFO

Article history:
Received 22 June 2012
Accepted 16 May 2013
Available online 5 June 2013

Keywords:
Steganalysis
Steganography
Block-based image steganalysis
Decision fusion
Stego image
Block decomposition
Fisher linear discriminant classifier
Dempster-Shafer theory

ABSTRACT

Traditional image steganalysis is conducted with respect to the entire image frame. In this work, we differentiate a stego image from its cover image based on steganalysis of decomposed image blocks. After image decomposition into smaller blocks, we classify image blocks into multiple classes and find a classifier for each class. Then, steganalysis of the whole image can be obtained by integrating results of all image blocks via decision fusion. Extensive performance evaluation of block-based image steganalysis is conducted. For a given test image, there exists a trade-off between the block size and the block number. We propose to use overlapping blocks to improve the steganalysis performance. Additional performance improvement can be achieved using different decision fusion schemes and different classifiers. Besides the block-decomposition framework, we point out that the choice of a proper classifier plays an important role in improving detection accuracy, and show that both the logistic classifier and the Fisher linear discriminant classifier outperforms the linear Bayes classifier by a significant margin.

1. Introduction

The goal of image steganography is to embed secret messages in an image so that no one except the intended recipients can detect presence of secret messages. It has many applications such as embedding the copyright information into professional images, personal information into photographs in smart IDs (identity cards), and patient information into medical images [1]. Using image steganalysis, one attempts to detect the presence of secret messages hidden in such images.

With the advance of image steganography, many steganalysis methods have been developed to deal with new breakthroughs in image steganography. In the early stage, it is assumed that some prior information about steganographic algorithms that embeds a secret message into images is available. This is called targeted steganalysis. However, more attention has been paid to a more realistic situation in recent years. That is, no information about steganographic algorithms is available. This is known as blind steganalysis, which attempts to differentiate stego images from cover images without the knowledge of steganographic embedding algorithms [2]. Using features extracted from cover and stego images in a training set, we may design a classifier that separates cover and stego images in the feature space.

Most previous work on image steganalysis focused on extracting features from images and used a binary classifier to differentiate stego images from cover images. The research objective was to find a better feature set to improve the steganalysis performance. Fridrich [3] proposed the use of DCT features for steganalysis since inter-block dependency between neighboring blocks is often affected by steganographic algorithms. Shi et al. [4] proposed to use Markov features since the differences between absolute values of neighboring DCT coefficients can be modeled as a Markov process. This feature set is useful because intra-block correlations among DCT coefficients within the same block can be affected by steganographic embedding. Pevnỳ and Fridrich [5] proposed a set of 274 merged features by combining DCT and Markov features together.

So far, little attention has been paid to the characteristics of cover images to design content-adaptive classifiers in steganalysis. An input image typically consists of heterogeneous regions. We may decompose an image frame into smaller blocks and use each block as a basic unit for steganalysis. The effect of steganographic embedding on similar image blocks is known to have a stronger correlation [6]. As a result, the characteristics of smaller blocks can be used to design content-adaptive classifiers.

The frame-based steganalysis, which extracts a set of features from the whole image, was reported in almost all previous work [3–5]. In contrast, the block-based steganalysis, which extracts features from each individual block, was proposed by the authors in [7]. Based on the block features, a tree-structured vector quantization (TSVQ) scheme can be adopted to classify blocks into multiple

* Corresponding author.
E-mail addresses: seonghoc@usc.edu (S. Cho), bhpaul.cha@samsung.com (B.-H. Cha), gawecki@usc.edu (M. Gawecki), cckuo@sipi.usc.edu (C.-C. Jay Kuo).

1047-3203/$ - see front matter
http://dx.doi.org/10.1016/j.jvcir.2013.05.007

classes. For each class, a specific classifier can be trained using block features, which represent the characteristics of the block class. For a given test image, instead of making a single decision for the entire image, we repeat the block decomposition process and choose a classifier to make a cover/stego decision for each block depending on block features. Finally, a decision fusion technique can be used to fuse steganalysis results of all blocks so that one can decide whether an unknown image is a cover or stego image.

The rest of this paper is organized as follows. Related previous work is reviewed in Section 2. The proposed block-based image steganalysis system is presented in Section 3. Analysis of the performance of block-based image steganalysis by considering the effects of block sizes, block numbers and the block overlapping design is conducted in Section 4. Fusion of multiple block decisions into one final decision for a test image is examined in Sec. 5. Extensive experimental results are shown for thorough performance evaluation in Section 6. Finally, concluding remarks and future research directions are provided in Sec. 7.

2. Review of previous work

Previous research in blind steganalysis has focused on extracting features from the whole image [3–5]. The number of features was increased to achieve better steganalysis performance in recent years. Chen et al. [8] proposed a set of updated Markov features (486 features in total) by considering both intra-block and inter-block correlations among DCT coefficients of JPEG images. Kodovskỳ et al. [9] examined a set of updated merged features (548 features in total) using the concept of Cartesian calibration. Pevnỳ et al. [10] used higher order Markov models to capture the differences between neighboring pixels in the spatial domain and developed a subtractive pixel adjacency model feature set (686 features in total). This feature set is also known to be effective with the LSB matching algorithm. Note that LSB matching is similar to LSB replacement, but it differs in that LSB matching changes LSBs only when the LSB of the next pixel from the cover image is different from the next bit of the secret message. In general, the steganalysis of LSB matching is known to be much more challenging compared to that of LSB replacement. More recently, Kodovskỳ et al. [11] introduced the cross-domain feature set (1234 features in total), which considers features from the spatial domain and the DCT domain at the same time. This feature set is known to be effective for steganalysis of the YASS algorithm [12], which embeds secret messages into randomized locations to make the calibration process ineffective.

Many steganographic embedding algorithms are block-based; namely, embedding the secret message into each 8×8 DCT block separately. Yang et al. [13] performed an information-theoretic steganalysis on the block-structured stego image. They provided an approximation of the relative entropy between probability distributions of the cover and the stego images. The relative entropy increases linearly with $N/K - 1$, where N, K represent the total number of samples (pixels) and the block size, respectively. A larger relative entropy means a higher detection probability of the stego image. Although Yang et al. [13] studied block-structured stego images, their work is still a frame-based approach from our viewpoint since only one set of features is extracted from an image.

The block-based image steganalysis was first introduced in [7], which extracted features from smaller blocks for image steganalysis. While the frame-based approach extracts a set of features from the whole image, the block-based approach takes advantage of the rich information of images by extracting a set of features from each individual image block. The characteristics of smaller image blocks were also exploited in [7] to design a content-adaptive classifier for steganalysis. It was shown by experimental results that the performance of blind steganalysis with merged features is significantly improved using the block-based approach. In this work, we will review results in [7] and add more discussion.

3. Block-based image steganalysis

3.1. System overview

The block-diagram of a block-based image steganalysis system is shown in Fig. 1. It consists of the training process and the testing process, which will be detailed in the following two subsections, respectively.

- The training process. The system decomposes an image into smaller blocks and treats each block as a basic unit for steganalysis. A set of features is extracted from each individual image block and a tree-structured hierarchical clustering technique is used to classify blocks into multiple classes based on extracted features. For each class of blocks, a specific classifier can be trained using extracted features which represent the characteristics of that block class. Note that if the number of training blocks is too large, a statistical sampling method can be used to reduce the number of training blocks.
- The testing process. The system performs the same block decomposition and feature extraction tasks on the test image. Then, it classifies each image block into one specific block class, and uses its associated classifier to make a decision whether the underlying block is a cover/stego block. Finally, there is a decision fusion step that integrates the decisions of multiple blocks into a single decision for the test image is conducted.

For block-based image steganalysis in [7,14], the merged feature set as proposed in [5] was extracted from image blocks, random sampling was adopted as the statistical sampling method in the training process, and the majority voting rule was used to fuse decision results from all the image blocks. For the classification task, a binary classifier was proposed in [7] and a multi-classifier was considered in [14]. It was shown by experimental results in [7,14] that the block-based approach offers better blind steganalysis performance than the frame-based approach.

There are two main advantages with the block-based steganalysis. First, it can offer better steganalysis performance without increasing the number of features. It provides a methodology to complement traditional frame-based steganalysis research that has focused on the search for more effective features. Second, the

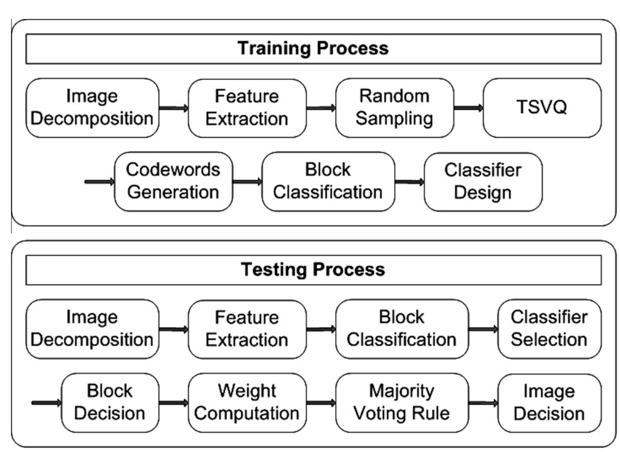

Fig. 1. The block-based image steganalysis system.