

# Automatic Detection of the Presence of Stego-signals and Watermarks in Images

Bülent Sankur<sup>a</sup>, Nasir Memon<sup>b</sup>, \*İsmail Avcıbaşı<sup>a</sup>

<sup>a</sup>Department of Electrical and Electronics Engineering, Boğaziçi University, Bebek, İstanbul, Turkey

<sup>b</sup>Department of Computer and Information Science, Polytechnic University, Brooklyn, NY, USA

[avcibas@hotmail.com](mailto:avcibas@hotmail.com), [memon@poly.edu](mailto:memon@poly.edu), [sankur@boun.edu.tr](mailto:sankur@boun.edu.tr)

## ABSTRACT

In this study we present techniques for steganalysis of images that have been potentially subjected to a watermarking algorithm. Our hypothesis is that a particular watermarking scheme leaves statistical evidence or structure that can be exploited for detection with the aid of proper selection of image features and multivariate regression analysis. We use some sophisticated image quality metrics as the feature set to distinguish between watermarked and unwatermarked images. To identify specific quality measures, which provide the best discriminative power, we use analysis of variance (ANOVA) techniques. The multivariate regression analysis is used on the selected quality metrics to build the optimal classifier using images and their blurred versions. The idea behind blurring is that the distance between an unwatermarked image and its blurred version is less than the distance between a watermarked image and its blurred version. Simulation results with a specific feature set and a well-known and commercially available watermarking technique indicates that our approach is able to accurately distinguish between watermarked and unwatermarked images.

**Keywords:** Steganalysis, watermarking, image quality measures, multivariate regression analysis.

## 1. INTRODUCTION

In this study we present techniques for steganalysis of images that have been potentially subjected to a watermarking algorithm. Steganography refers to the science of “invisible” communication and its main goal is to communicate securely in a completely undetectable manner. Thus any third party should not be able to distinguish in any sense between cover-objects (objects not containing any secret message) and stego-objects (objects containing a secret message). In this context, “*steganalysis*” refers to the body of techniques that are designed to distinguish between cover-objects and stego-objects.

A digital watermark is an imperceptible signal added to digital content that can be later detected or extracted in order to make some assertion about the content. Given the proliferation of content in digital form, recent years have seen an increasing interest in digital watermarking. This research will also shed a light to the effectiveness of watermarking techniques for steganographic applications. In other words if digital watermarks are to be used in steganography applications, detection of their presence by an unauthorized agent defeats their very purpose. Even in applications that do not require hidden

---

\* The work was sponsored by NSF-INT 9996097 and The Scientific Council of Turkey TUBITAK, BDP Program.

communication, but only robustness, we note that it would be desirable to first detect the possible presence of a watermark before trying to remove or manipulate it. This means that a given signal would have to be first analyzed for the presence of a watermark. Based on this analysis there could then be attempts made to remove the watermark.

A general underlying idea behind watermarking is to create a watermarked signal that is *perceptually identical but statistically different* from the host signal. Our hypothesis is that a particular watermarking scheme leaves statistical evidence or structure that can be exploited for detection with the aid of proper selection of image features and multivariate regression analysis. In fact a decoder uses this statistical difference in order to detect the watermark. However, the very same statistical difference that is created could be potentially exploited to determine if a given image contains a watermark.

To this effect we have studied some 26 different image quality metrics and using statistical significance analysis we have gleaned out of them eight useful metrics that are instrumental in discriminating watermarked from the original images. To this effect we used ANOVA (Analysis of Variance) test to distinguish measures that are most consistent and accurate vis-à-vis the effects of watermarking and the effects of blurring. More specifically, we consider a set of training images, and the resulting quality degradation measures from watermarking and from blurring. The selected subset of image quality measures with respect to their discriminative power are the following: 1) Mean Square Error, 2) Multiresolution Distance Measure, 3) Structural Content, 4) Cross Correlation, 5) Weighted Spectral Distance, 6) Median Block Weighted Spectral Distance, 7) Normalized Absolute Error (HVS), 8) HVS Based L2, and 9) Gradient Measure.

In the design phase of the steganalyzer, we regressed the normalized quality measure scores to, respectively, -1 and 1, depending upon whether an image did not or did contain a watermark. Similarly, image quality scores were calculated between the original images and their blurred version. Once the prediction coefficients are obtained in the training phase, these coefficients can be used in the testing phase. Given an image in the test phase, first it is blurred and the  $q$  image quality measure scores are obtained using the image and its blurred version. Then using the prediction coefficients, these scores are regressed to the output value. If the output exceeds the threshold 0 then the decision is that the image contains watermark, otherwise the decision is that the image does not contain watermark.

The watermarking techniques we used were the following: 1) Photoshop plug-in Digimarc, Cox's technique and the technique from Swiss Federal Institute of Technology, PGS. One reason for the selection of these techniques was their free availability on the Internet and they were all popularly known algorithms. The other reason was that with these techniques it was possible to embed watermarks at different strengths.

The experimental results on a set of 30 images show that our method can distinguish the watermarked from non-watermarked images with 85 % accuracy. In addition it can identify with 75% correctness the origin of the algorithm. Thus both "Is it watermarked for steganography?" and "Whose watermark is it?" types of questions can be answered.