ICA based Steganography

By Stéphane Bounkong, David Saad and David Lowe

Neural Computing Research Group, Aston University, UK

Abstract

A domain independent ICA-based watermarking method is introduced and studied by numerical simulations. This approach can be used either on images, music or video to convey a hidden message. It relies on embedding the information in a set of statistically independent sources (the independent components) as the feature space. For the experiments the medium has arbitrarily chosen to be digital images.

1. Introduction

Interest in watermarking techniques has grown significantly in the past decade, mainly due to the need to protect intellectual property rights in products based on digitized electronic information. Watermarking schemes have not only to be imperceptible and convey a high amount of data, but also robust against attacks. These contradicting constraints are inherent to any watermarking framework. The present paper describes a principled domain independent watermarking framework suitable for music, video or images. The new approach is based on embedding the message in statistically independent sources of the covertext to minimize covertext distortion, maximize the information embedding rate and improve the method's robustness against various attacks.

2. Domain Independent Watermarking

In the past few years, significant attention has been drawn to blind source separation by Independent Component Analysis (ICA). The recent discovery of efficient algorithms and the increase in computational abilities, have made it easier to extract statistically independent sources from given data.

ICA is a general purpose statistical technique which, given a set of observed data, extracts a linear transformation such that the resulting variables are as statistically independent as possible. Such separation may be applied to audio signals or digitized images, see Hyvärinen & al (2001), assuming that they constitute a sufficiently uniform class of data so that a statistical model can be constructed on the basis of observations. Experiments that are conducted on a set of digitized images (Fig. 1 and 2), show that this hypothesis holds, providing the basis for a general domain independent framework †. The suggested framework can be based on various generative methods, although in this paper we will focus on ICA, for further details on the method see Hyvärinen & al (2001).

Basic watermarking schemes can be described in three steps. Firstly, a given message $\mathbf{m}$, also termed watermark, is embedded into the covertext $\mathbf{X}$ (e.g. a digitized image, audio or a transformed signal) providing a watermarked covertext $\hat{\mathbf{X}}$. Secondly, the watermarked text may be attacked either maliciously or non-maliciously, resulting in the

† In the case of multiple, significantly different, covertext groups, one may construct a different model for each group.

FIGURE 1. Example of natural scene image.

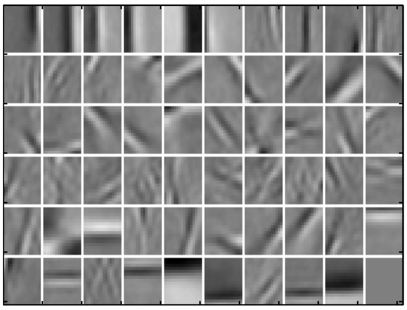

FIGURE 2. ICA basis obtained from images.

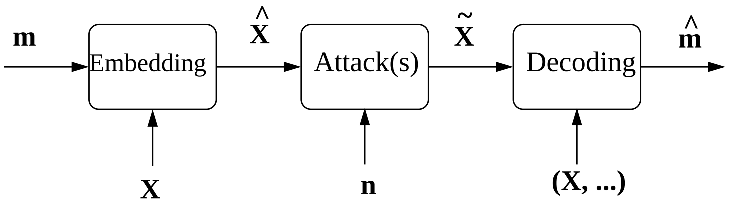

FIGURE 3. A general watermarking scheme where $\mathbf{m}$ is the embedded message, $\mathbf{X}$ is the covertext, $\hat{\mathbf{X}}$ the watermarked covertext, $\tilde{\mathbf{X}}$ the attacked covertext and $\hat{\mathbf{m}}$ an estimate of the original message $\mathbf{m}$.

attacked covertext $\tilde{\mathbf{X}}$. Finally, a decoder extracts the message estimate $\hat{\mathbf{m}}$ from the corrupted version of the watermarked covertext $\tilde{\mathbf{X}}$. The process is summarized in Fig. 3.

In the framework studied in this paper, $\mathbf{X}$ may be derived from any media, such as audio signals or digitized images. The demixing matrices W obtained by the ICA algorithm for the different domains may be significantly different but the principle remains the same. Indeed, the demixing process gives a set of independent sources, which share similar characteristics and have little correlation with the original domain.

The general ICA-based watermarking process comprises four stages as described in Fig. 4.

(a) The image is divided into contiguous image patches giving a set of mixed signals. Each patch is then demixed using a predetermined ICA demixing matrix W, prepared using an ensemble of typical images.

(b) For each patch, a set of relevant coefficients are selected according to the specific task.

(c) The selected coefficients are quantized and watermarked. The difference between the watermarked and original values is denoted by Δ.

(d) Δ is multiplied by the mixing matrix A to produce w which is then added to the original picture I.

Various efficient approaches have been suggested for hiding/embedding information. Here, we used the distortion-compensated Quantization Index Modulation (QIM) method studied by Chen & al (2000), that has been shown to be close to optimal in the case of additive Gaussian attacks and is easy to use. It is based on quantizing the covertext real-valued independent source to some central value followed by a quantized addition/subtraction representing the binary message bit. This may also be modified by a prescribed noise template making it difficult to identify the QIM embedding process and its parameters.

The decoding process is carried out in a similar way, by projecting the received (at-

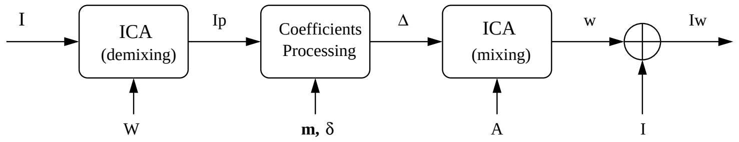

FIGURE 4. Domain independent watermarking scheme; where I is the original image, W the ICA de-mixing matrix, Ip is the set of de-mixed signals, $\mathbf{m}$ is the embedded message, δ is the quantization step, Δ the difference between the quantized and original values, A the ICA mixing matrix, w is the mixed signal and Iw the watermarked image.

tacked) value onto the nearest point on the quantization grid, which represents the embedding of a specific message bit (0/1). The description of the attacked text is computed from the attacked covertext by employing the demixing matrix W giving us the corrupted source $\tilde{\mathbf{X}}$. The estimation of $\mathbf{m}$ is computed from $\tilde{\mathbf{X}}$ in conjunction with other available information, such as attack characteristics, original covertext, cryptographic key, etc. (Fig. 4). More information about the method used can be found in Bounkong & al (2002).

3. Experiments

To test the performance of our watermarking scheme against existing state of the art methods, we carried out a set of experiments for each of the watermarking tasks. The maximum allowed distortion threshold δ_1, related to the embedding process, is set to 43 dB using the peak to signal noise ratio (PSNR) measure to ensure that all watermarking are imperceptible. The maximum allowed distortion threshold δ_1 defines the maximum allowed quantization step δ; the distortion level we use here is consistent with acceptable distortion levels used in the literature. All tests were carried out on a test set of 11 grey-scale natural scene images (Fig. 1) of 512×512 pixels. For any given attack various levels of corruption are tested. The embedded messages $\mathbf{m}$ are randomly generated binary sequences on $\{0, 1\}$. The latter is encoded using QIM for all tested methods. The decoding process is also common to all trials and is carried out by mapping the received (attacked) data to the nearest quantization grid point. For a given image and attack strength, the test is carried out 100 times.

For comparison purposes, two other watermarking schemes have been tested under the same attacks using the same embedding and decoding methods. Both methods operate in the discrete cosine transform (DCT) domain.

ICA - This is an ICA based algorithm, where we preselect a small subset of ICs that are particularly robust against a specific attack (i.e., that suffer very little distortion from the given attack); a single IC is then randomly selected from this subset to be watermarked in each patch.

DCT - A standard, commonly used, DCT based algorithm; QIM is used to embed the message in a DCT feature space, where the DCT is carried out over the entire image. Among the DCT coefficients which represent a signal with at least one cycle per image patch of 16×16, the 1024 lowest frequencies are chosen to convey the watermark. The watermarked picture is then obtained by the application of an inverse DCT.

DCTX - A local DCT based algorithm. It relies on a partitioning of the picture into contiguous patches of 16×16 pixels. The DCT is applied to each of them. For each patch, a single coefficient is randomly selected among the low frequencies to be quantized (QIM). An inverse DCT is then applied to obtain each watermarked patch.

We carried out four experiments where watermarked pictures are attacked by:

- White Noise (WN) of mean zero and of various standard deviation values,
- JPEG lossy compression with different quality levels,
- Set Partitioning In Hierarchical Trees (SPIHT) compression,
- Resizing with various factors.

4. Results

Figure 5a, shows that all schemes are quite robust against white noise attack considering the fact that the 43 dB attack distortion threshold is reached for a WN of standard deviation of about 2. In the case of ICA, it is easy to see a direct relation between the quantization step δ and the process robustness, since the noise in the feature space is also Gaussian. This may not be the case if other decoding methods, such as Maximum A Posteriori (MAP), are used. Moreover it also shows that one potential weakness of the this scheme, the ICA restriction on extracting only non-Gaussian sources, is not highly significant, even in the case of Gaussian noise attacks.

Figure 5b shows that ICA and DCTX are quite robust against JPEG compression, while DCT performs quite poorly. The distortion induced by the attack is below 38 dB for quality level under 90. Notice that JPEG Compression uses a similar transform to DCTX, which makes DCTX highly suitable for this type of attack.

Figure 5c shows excellent performances for all systems for bit rates above 1 bit per pixel. From this point and on, the error rate increases significantly while the bit rate decreases.

Figure 5d shows that ICA and DCTX perform better than DCT. However, the overall performance is quite poor; this is not surprising since resizing attack is quite lossy in terms of the information content, and for most of the resizing factor values tested here, the attacked image is significantly damaged. Note that for a 0.25 resizing factor, the picture size is reduced by more than 93% in storage.

5. Conclusion

In this paper, a new principled domain independent watermarking framework is presented and examined. Experiments show highly promising performance in comparison to other state of the art methods on a limited set of attacks. The attacks include four of the most common attacks: white noise attack, JPEG lossy compression, SPIHT lossy compression and resizing.

A novel approach to watermarking, using ICA as the feature space in which watermarks are embedded, is presented. The new approach, being based on embedding information in statistically independent sources, shows high information embedding rate and minimal distortion as proved analytically in the literature Moulin and O'Sullivan (2001). Its performance is examined numerically on a set representative images, random messages and various attacks.

The main advance is that, being based on embedding information using statistically independent sources, the same watermarking method can be easily applied across different media. Based on local information and a linear transform, our method is computationally efficient, offering additional security in the use of *specific* mixing/de-mixing matrices that are not easy to obtain.

Providing statistical models for both sources and attack will facilitate the use of a Bayesian decoding methods that will potentially provide an optimal decoding scheme. Further research may improve the performance by using such statistical models and focus

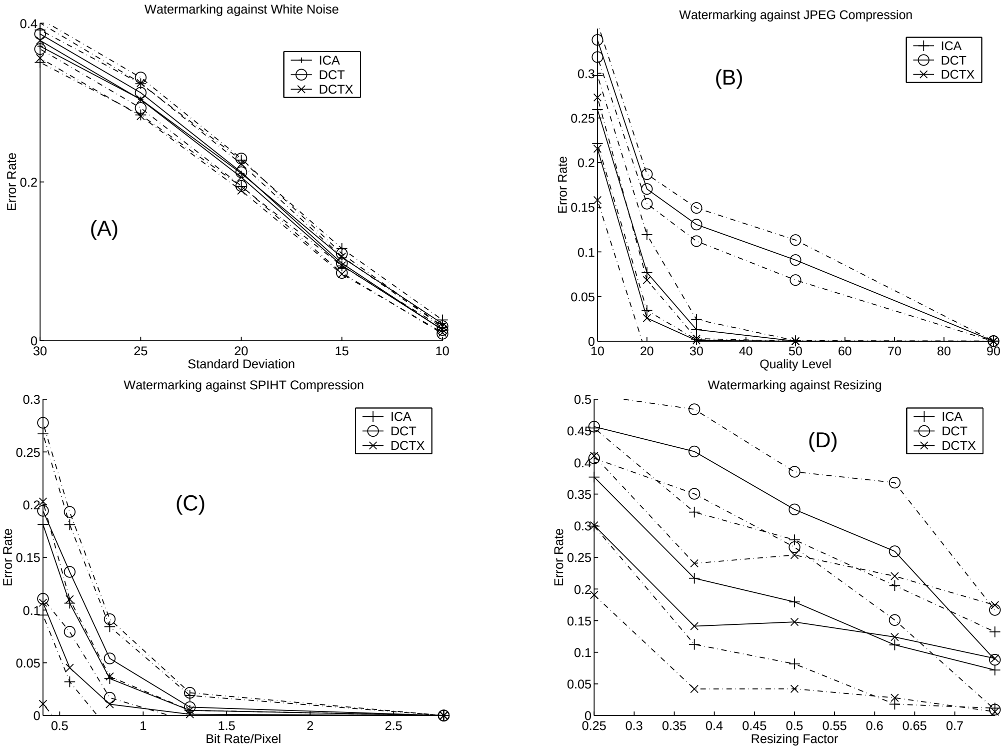

FIGURE 5. The performance of the three watermarking tested: ICA, DCT and DCTX, against various attacks; solid lines and symbols represent the mean values; dashed lines denote error bars. (A) White noise of different standard deviation values. (B) JPEG lossy compression for different quality levels. (C) SPIHT lossy compression for different bit rate. (D) Resizing with different resizing factors.

on a new approach to select the IC's to be watermarked based on a human visual system distortion measure.

Support by EPSRC-GR/N63178 (DS) is acknowledged.

REFERENCES

BOUNKONG, S., SAAD, D., LOWE, D. 2002 ICA for Domain Independent Watermarking *Conference Proceedings, International Conference on Artificial Neural Network*, LNCS 2415.

HYVÄRINEN, A., KARHUNEN, J., OJA, E. 2001 Independent component analysis. *Wiley-Interscience, New York.*

COX, I., MILLER, M. L., BLOOM, J. A. 2001 Digital Watermarking. *Principles and Practice, Morgan Kaufmann, San Francisco.*

MOULIN, P., O'SULLIVAN, J. A. (2001) Information-Theoretic Analysis of Information Hiding `http://www.ifp.uiuc.edu/~moulin/` .

PETITCOLAS, F. A. P., ANDERSON, R. J., KUHN, M. G. 1999 *Proceeding of IEEE Multimedia Systems.* **87**, 1062

B. CHEN AND G.W. WORNELL 2000 Quantization Index Modulation : A Class of Provably Good Methods for Digital Watermarking and Information Embedding *IEEE Trans. Inform. Theory*, in press.