IEICE TRANS. INF. & SYST., VOL.E91–D, NO.7 JULY 2008

LETTER

Intelligent Extraction of a Digital Watermark from a Distorted Image

Asifullah KHAN[†], **Syed Fahad TAHIR**[†], *Nonmembers, and* **Tae-Sun CHOI**[†a)], *Member*

SUMMARY We present a novel approach to developing Machine Learning (ML) based decoding models for extracting a watermark in the presence of attacks. Statistical characterization of the components of various frequency bands is exploited to allow blind extraction of the watermark. Experimental results show that the proposed ML based decoding scheme can adapt to suit the watermark application by learning the alterations in the feature space incurred by the attack employed.

key words: *watermark, decoding, transform, copyright protection, robust, blind, Machine Learning*

1. Introduction

Due to the rapid development of multimedia technologies, digital content is easy to create, duplicate and distribute through the internet. In these circumstances, implementing digital right management has now become an urgent goal. Watermarking is considered to be the most prospective technology for answering issues related to digital right management. The watermark, after being embedded in a digital medium and transmitted, is supposed to be extracted at the receiving end. However, due to normal image processing and intentional attacks on the watermarked image, the accurate extraction of the watermark has become a challenging problem [1].

There is no such watermark decoding scheme that can perform well under all hostile attacks. However, with the growing need of sophisticated watermarking applications, we require a decoding scheme that can adapt well towards a specific application. Generally, regular signal processing, channel noise, and JPEG compression are the most common attacks. Blind extraction of the hidden information becomes complicated, after distortions are incurred to the watermarked image. For example, in case of telesurgery; due to possible attacks, blind extraction of valuable embedded information about the identity of the patient, hospital, medical instruments, etc. is demanding. Medical data networks are now widely used in countries such as Japan. Even in transform domain watermarking approaches, for instance, Discrete Cosine Transform (DCT), attacks can appreciably change the underlying distribution of the coefficients. During watermark extraction phase, it is usually assumed that the distribution of DCT coefficients is not heavily altered. Therefore, watermark extraction performance degrades under attacks. As against the conventional correlation based extraction of watermark, some researchers have developed new decoder structures [2]. In addition, Machine Learning (*ML*) models for watermark detection/decoding are also effectively employed [3]–[6]. For instance, Fu et al. [3] utilize the learning capabilities of Support Vector Machine (*SVM*) for optimal detection of a watermark. On the other hand, Bounkong et al. [4] have proposed independent component analysis based watermarking. Fridrich et al. [5] exploit the learning capabilities of *SVM* for improving blind steganalysis. In a recent work, Khan et al. [6] have proposed the modification of decoder structure using Genetic Programming in accordance to both the cover image and conceivable attacks.

Nonetheless, most of these approaches do not consider the presence of attacks during the training phase and thus are not adaptive. Similarly, watermarking approaches [2] that do not exploit *ML* techniques, generally, use simple Threshold Decoding (*TD*) and thus, are also not adaptive towards the attack on the watermark. These approaches neither consider the alterations that may incur to the features and nor exploit the individual frequency bands; rather treat all the frequency bands collectively. In contrast, we present an innovative scheme of exploiting the selected frequency bands individually. Our proposed technique is adaptive towards a new hostile application of the watermarking scheme, as we exploit the learning capabilities of *ML* models to gain knowledge of the distortion that might have incurred varyingly on the different frequency bands due to the attack.

2. Proposed Watermark Extraction

Simple TD model can accurately classify bits if the distribution of the features does not overlap. This is because using a threshold, only linear bifurcation could be possible. However, in case of an attack on the watermarked image, the distributions of the features of a decoding model overlap as shown in Fig. 1. Consequently, linear bifurcation is not possible and thus a simple *TD* model is unable to decode the message bits efficiently. For this purpose, we assume that a non-separable message in lower dimensional space might be separable if it is mapped to higher dimensional space. This mapping to higher dimensional space is what the hidden layers in case of Artificial Neural Networks (*ANN*) and the kernel functions in case of *SVM* perform. As shown in Fig. 2, our decoding scheme mainly consists of following two modules; Dataset Generation and *ML* based Decoding.

Manuscript received August 31, 2007.
Manuscript revised March 4, 2008.
[†]The authors are with GIST, Gwangju, 500–712 South Korea.
a) E-mail: tschoi@gist.ac.kr
DOI: 10.1093/ietisy/e91–d.7.2072

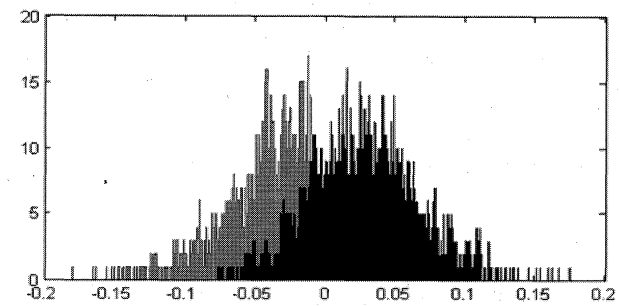

Fig. 1 Distribution of sufficient statistics of the maximum likelihood based decoding system after Gaussian attack (σ_{attack} = 10).

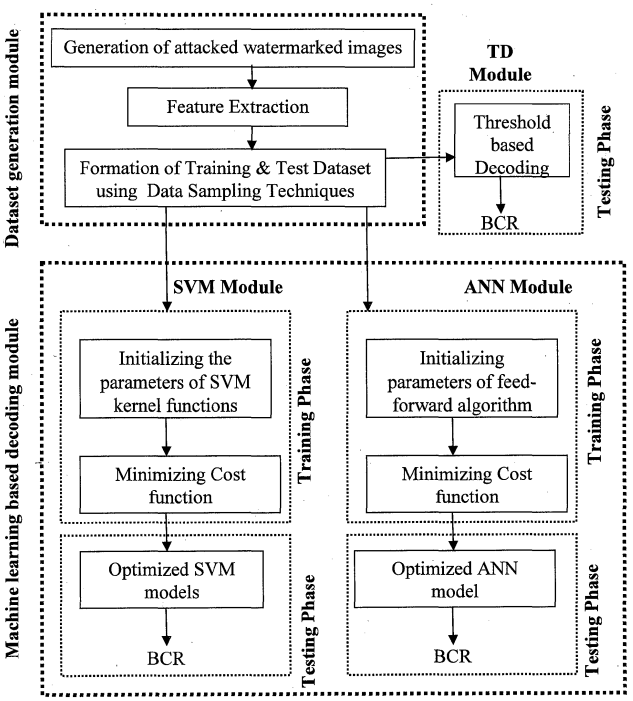

Fig. 2 Basic block diagram of watermark extraction.

2.1 Generating Attacked Watermarked Images

In order to analyze the performance of the proposed adaptive *ML* based extraction of watermark, we have generated a dataset of 16000 bits by considering five different images, each of size 256 × 256. Next, in each image, a message of size 128 bits is embedded [2]. The whole process is repeated 25 times by changing the secret key used to generate the spread spectrum sequence. The anticipated attack is performed on each image, which corrupts the embedded message as well. The product of the spread-spectrum sequence $S[\mathbf{k}]$ and expanded code vector $b[\mathbf{k}]$; corresponding to the message to be embedded, is multiplied with a perceptual mask $\alpha[\mathbf{k}]$ to obtain the watermark. $[\mathbf{k}]$ denote the 2-D discrete indices in DCT domain. The 2-D watermark signal $W[\mathbf{k}]$ is as:

$$W[\mathbf{k}] = S[\mathbf{k}] \cdot \mathrm{b}[\mathbf{k}] \cdot \alpha[\mathbf{k}] \tag{1}$$

Adding this watermark to the original image in transformed domain performs the embedding:

$$Y[\mathbf{k}] = X[\mathbf{k}] + W[\mathbf{k}] \tag{2}$$

where $X[\mathbf{k}]$, and $Y[\mathbf{k}]$ represents the original and watermarked images in DCT domain respectively.

The embedded message is supposed to be blindly extracted by modeling the coefficients of each frequency band of an 8×8 block DCT domain and applying maximum likelihood estimation. We consider the sufficient statistics for the watermark decoding as feature for the hidden bit classification. A statistic $T(Z)$ is sufficient for a parameter θ, if the conditional probability distribution of the data Z, given the statistic $T(Z)$, is independent of the parameter θ. However, as opposed to [2] that utilize a single feature for the *TD* model, we do not compute sufficient statistics collectively across all the frequency bands; rather we compute it across each frequency band individually. This is because we assume that the frequency channels are independent, but not identical; a single attack might have different effects on the different frequency channels. Additionally, this allows us to keep the sufficient statistics across each frequency band as a separate feature itself. Thus, for a single bit, the number of features is equal to the number of selected frequency bands. Considering all the selected frequency channels to be identical, the sufficient statistics r_i corresponding to each embedded bit is given as:

$$r_i \triangleq \sum_{\mathbf{k} \in G_i} \frac{\left|Y[\mathbf{k}] + \alpha[\mathbf{k}]s[\mathbf{k}]\right|^{c[\mathbf{k}]} - \left|Y[\mathbf{k}] - \alpha[\mathbf{k}]s[\mathbf{k}]\right|^{c[\mathbf{k}]}}{\sigma[\mathbf{k}]^{c[\mathbf{k}]}} \tag{3}$$

where G_i denotes the sample vector of all DCT coefficients in different 8×8 blocks that correspond to a single bit i. σ represents the standard deviation of the distribution, while c dictates the shape of generalized Gaussian distribution.

We modify this model by assuming that each frequency band is distorted differently by an attack. Thus, the sufficient statistics corresponding to a single bit are computed separately for each frequency band:

$$r_i = \sum_j r_i^j \qquad j = 1, 2, \ldots, J_{\max} \tag{4}$$

where $J_{\max}$ is the maximum number of selected frequency bands, and r_i^j is defined as:

$$r_i^j \triangleq \sum_{\mathbf{k} \in Q_i^j} \frac{\left|Y[\mathbf{k}] + \alpha[\mathbf{k}]s[\mathbf{k}]\right|^{c[\mathbf{k}]} - \left|Y[\mathbf{k}] - \alpha[\mathbf{k}]s[\mathbf{k}]\right|^{c[\mathbf{k}]}}{\sigma[\mathbf{k}]^{c[\mathbf{k}]}} \tag{5}$$

where Q_i^j, is defined as the sample vector of all DCT coefficients in different 8×8 blocks that correspond to a single bit i and the jth frequency band. c and σ are estimated from the received watermarked image.

2.2 Proposed ML Based Decoding

For bipolar signal; $b[\mathbf{k}] \in [-1, 1]$, the estimated bit $\hat{b}_i$ in *TD* model is computed as $\hat{b}_i = \text{sgn}(r_i) \quad \forall\, i \in \{1, 2, \cdots, N\}$.

In contrast, we treat the sufficient statistics as features for the *ML* based decoding scheme. In order to make these features linearly separable, they are first mapped to a higher dimensional space. Consider n training pairs (r_i, q_i), where $r_i \in R^N$ and $q_i \in [1, -1]$, then, in case of *SVM*, we have the nonlinear mapping:

$$f(r) = \sum_{i=1}^{N_S} \alpha_i q_i K(r_i, r) + b = \sum_{i=1}^{N_S} \alpha_i q_i \Phi(r_i) \cdot \Phi(r_v) + b \qquad (6)$$

where $\Phi(r)$ is nonlinear mapping function and b is bias. The attractiveness of this approach is its ability to develop an optimal hyper-plane in view of a new attack by learning the distortion in the features. We use three different kernels; Linear, Radial basis function (Rbf), and Polynomial. Similarly, r_i are provided as features to the multi-layered neural Network. Back-propagation learning algorithm is employed during training phase and weights are updated according to *Levenberg-Marquardt* Algorithm. This algorithm computes the error e, for output neuron m, as:

$$e_m(t) = z_m(t) - p_m(t) \qquad (7)$$

where $z_m(t)$ and $p_m(t)$ are the actual and target output for neuron m for iteration t.

3. Simulations

Figure 3 (a) and 3 (b) show the original and watermarked images. Figure 3 (c) shows the Gaussian noise attacked watermarked image, where the imperceptibility is affected strongly by the resultant distortions, as is evident from the difference image (d). Figure 4 shows the 4-fold cross-validation comparison in terms of Bit Correct Ratio (*BCR*); ratio of correct bits to that of total bits. The order of performance of decoding models on training data is: *PolySVM* > *ANN* > *RbfSVM* > *LinearSVM* > *TD*. On the other hand, for test data, we have *PolySVM* > *LinearSVM* > *RbfSVM* > *TD* > *ANN*. In case of test data, *PolySVM* is able to correctly decode all the message bits from a distorted image. This shows that in view of conceivable attacks on a watermarked image, which are most common in real world applications of watermarking, it is far better intelligently employing an *ML* technique for learning the distortion introduced by the attack.

To analyze the adaptability of the proposed *scheme*, we change the conceivable attack and retrain the *ML* model accordingly. In this case, the attack is JPEG compression (QF = 80). The sufficient statistics (Eq. (5)) are computed in the same way and the *ML* model is able to learn the new distortion. Consequently, it is able to blindly extract all the message bits (Table 1). We then change the conceivable attack from JPEG compression to Wiener estimation. The

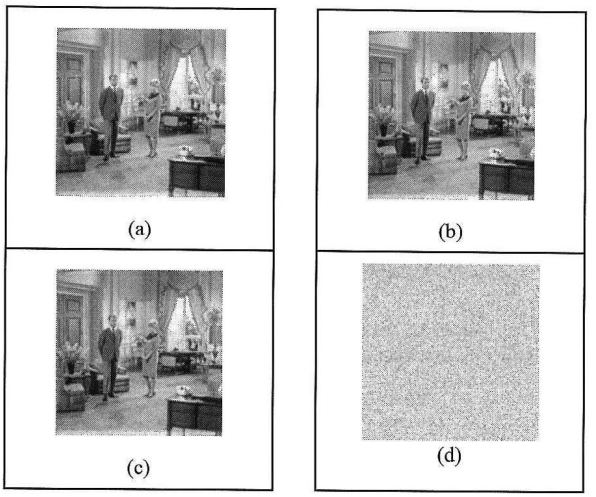

Fig. 3 Analyzing distortion. (a) Original, (b) Watermarked, (c) Gaussian noisc attacked, and (d) Difference of (a), and (c) images.

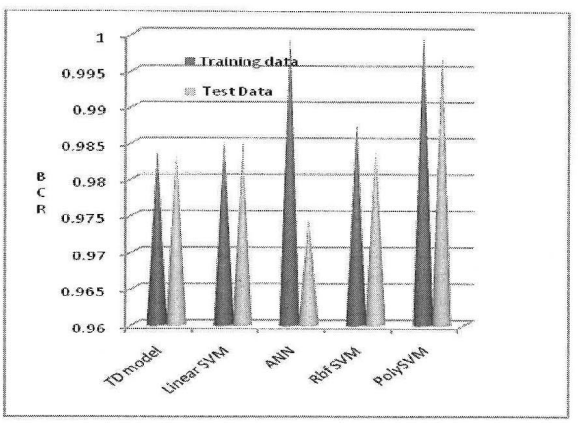

Fig. 4 *BCR* comparision using 4-fold cross-validation.

Table 1 BCR performance against different attacks. Notc: data size = 16 K (bits), and feature set = 22.

Type Of Attack	*ML Model*	*Parameters*		*Time (sec)*	*BCR*
		C	γ		
JPEG Comp. QF –80	*Linear SVM*	2	128	40	0.9266
	Poly SVM	2	128	6895	0.9942
	Rbf SVM	**2**	**128**	**697**	**1.0**
	ANN	Hidden layers =3 (8,4,2)		160	0.9431
	TD	-		-	0.9119
Wienr. Filter 3x 3	*Linear SVM*	2		49	0.9191
	Poly SVM	2	1.4	573	0.9490
	Rbf SVM	**2**	**16**	**713**	**1.0**
	ANN	Hidden layers =3 (8,4,2)		284	0.9324
	TD	-		-	0.8316

ML decoding schemes during their training phase are able to learn the novel distortion being introduced. It can be observed from Table 1 that *RbfSVM* is able to cope with such

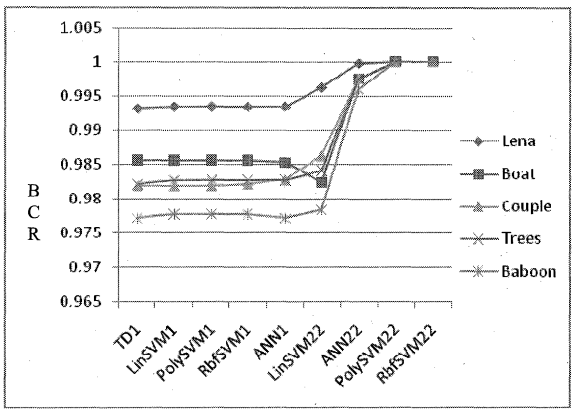

Fig. 5 BCR performance using different feature subsets.

change in distortion and offers highest performance, achieving $BCR = 1.0$ as compared to a $BCR = 0.8316$ for *TD* model. Figure 5 demonstrates that using a single feature, the performance of all the models deteriorates as we move from a relatively smooth image towards a textured image. In contrast, by exploiting 22 features, *ML* models are able to cope with severity of attack by learning the distortion. Specifically, *PolySVM, and RbfSVM* show $BCR = 1$, even across highly textured image of Baboon.

4. Conclusions

As regards blind watermark extraction in presence of attacks, both *SVM* and *ANN* decoding models are able to adopt according to the hostile environment. Our proposed intelligent decoding scheme has blindly extracted message bits from a distorted image and is a generic one—not limited to a specific set of watermarking schemes. The proposed approach could be highly effective in dynamic applications of watermarking, where varying attacks are expected at different times.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation, grant funded by the Korea government (MOST) (No.R01-2007-000-0227-0).

References

[1] I.J. Cox, M.L. Miller, and J.A. Bloom, Digital Watermarking and Fundamentals, Morgan Kaufmann, San Francisco, 2002.
[2] J.R. Hernandez, M. Amado, and F. Perez-Gonzalez, "DCT domain watermarking techniques for still images: Detector performance analysis and a new structure," IEEE Trans. Image Process., vol.9, no.1, pp.55–68, 2000.
[3] Y. Fu, R. Shen, and H. Lu, "Optimal watermark detection based on support vector machines," Lecture Notes in Computer Science, vol.3137, pp.552–557, 2004.
[4] S. Bounkong, B. Toch, D. Saad, and D. Lowe, "ICA for watermarking digital images," J. Machine Learning Research, vol.4, pp.1471–1498, 2003.
[5] J. Fridrich and T. Penvy, "Merging Markov and DCT features for multi-class JPEG steganalysis," Proc. SPIE Electronic Imaging, Photonics West, Jan. 2007.
[6] A. Khan, "A novel approach to decoding: Exploiting anticipated attack information using genetic programming," Int. J. Knowl. Based Intell. Eng. Syst., vol.10, no.5, pp.337–347, 2006.