A Robust Image Zero-watermarking using Convolutional Neural Networks

Atoany Fierro-Radilla
ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
afierror@hotmail.com

Mariko Nakano-Miyatake
ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
mnakano@ipn.mx
ORCID:0000-0003-1346-7825

Manuel Cedillo-Hernandez
ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
mcedillohdz@hotmail.com
ORCID: 0000-0002-9149-9841

Laura Cleofas-Sanchez
ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
laura18cs@hotmail.com

Hector Perez-Meana
ESIME Culhuacan
Instituto Politecnico Nacional
Mexico City, Mexico
hmperezm@ipn.mx
ORCID: 0000-0002-7786-2050

***Abstract*— In the image zero-watermarking techniques, a watermark sequence is not physically embedded into the host image but has a logical linkage with the host image. This property of zero-watermarking is desirable for some kinds of images in which a minimum distortion may cause serious detection or diagnostic errors, such as medical images and remote sensing images. In this paper, we propose a robust zero-watermarking algorithm based on the Convolutional Neural Networks (CNN) and deep learning algorithm, in which robust inherent features of image is generated by the CNN, and it is combined with the owner's watermark sequence using XOR operation. The experimental results show the watermark robustness against several attacks and common image processing.**

Keywords—Zero-watermarking, Deep Learning, Convolutional Neural Networks, Robust Features

I. Introduction

During last two decades, several types of image watermarking algorithms have been proposed for different purposes, such as the copyright protection [1, 2], owner identification [3, 4] and content authentication [5, 6]. Almost all of them, a watermark sequence is physically embedded into the host image, causing some distortion in it. In the invisible watermarking, the distortion caused by the embedded watermark is almost imperceptible by the Human Visual System (HVS). However, in some kind of images, such as medical images and remote sensing images, a small distortion may causes serious diagnostic or detection errors.

Until now two kinds of distortion-free watermarking techniques have been proposed in the literature. The first one is reversible watermarking, in which once watermark sequence is detected or extracted from the watermarked image, the original undistorted host image can be recovered [7, 8]. Generally, the reversible watermarking techniques cannot provide sufficient robustness of watermark sequence to common signal processing, such as JPEG compression and noise contamination. The second distortion-free watermarking technique is zero-watermarking, in which the watermark sequence is not physically embedded into the host image, but logically linked with the host image, keeping the host image intact.

In the zero-watermarking, instead of embedding a watermark sequence or watermark pattern, some inherent features are extracted from the host image. These inherent features are linked with an owner's watermark sequence to generate a master share, which is stored in a secure manner [9-13]. The protected images by the zero-watermarking are transmitted through any insecure public communication channel and the owner can verify his ownership using the master share and the inherent features extracted from the image under analysis. In the zero-watermarking technique, the extraction of robust inherent features of the host image is the most important issue for their desirable performance.

In [9, 10], using Singular Value Decomposition (SVD), some largest singular values are obtained from the approximation sub-band (LL sub-band) after the 2D Discrete Wavelet Transform (DWT) decomposition is applied to the host image. The singular values are linked with the watermark sequence using XOR operation and stored as master share. In [11], first the host image is normalized using Hu's image normalization technique [14] and applied the SVD in the Contourlet domain of the normalized host image to extract robust features. In [12], the image normalization technique is used, and the Bessel-Fourier moment is obtained from the normalized image, which is used to generate a master share. In [13], some robust Quaternion Exponent moments are extracted from the host image to generate master share.

Recently, the Convolutional Neural Networks (CNN) together with deep learning algorithms are used to solve several computer vision problems, in which the CNN is trained to extract the useful features of image for a desired task. Unlike the conventional methods, in which the hand-crafted features are firstly extracted and then some classifiers are trained using pre-obtained hand-crafted features, in the CNN-based approach, the

CB2014-0237925, Mexican National Council of Science and Technology

feature extraction and classification or detection tasks are carried out at the same time through a deep neural network structure with an input layer, multiple hidden layers and an output layer.

In this paper, we propose a deep learning-based zero-watermarking scheme, in which the inherent features of image are constructed through the training process of the CNN. Once the CNN is trained, we obtain the output of the first fully connected layer (*fc_1*) and link it with the owner's watermark sequence to generate master share. In the image verification stage, the image under analysis is introduced to the pre-trained CNN to obtain its inherent features, which are used to obtain the watermark sequence. Although the training of the CNN is time-consuming, once the CNN is trained, the feature extraction of given image is carried out automatically within a second.

The rest of the paper is organized as follows. In Section II, we provide the proposed zero-watermarking scheme in detail and experimental results are provided in Section III. Finally, we conclude this paper in Section IV.

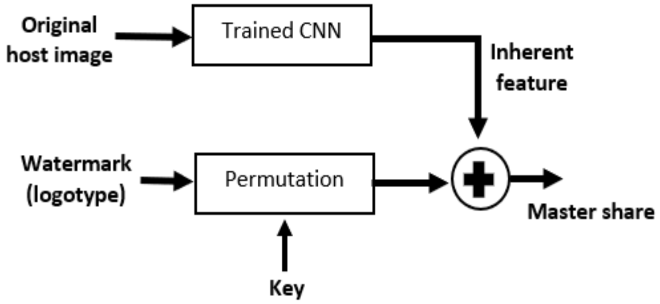


Fig. 1 Master share generation scheme

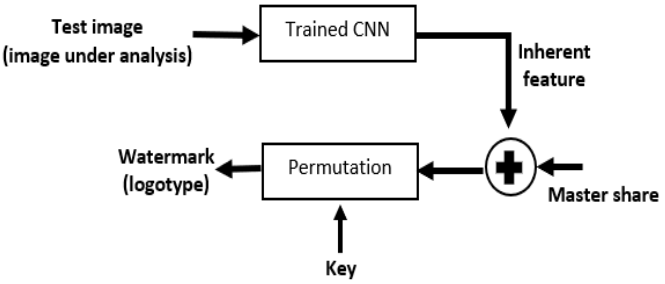


Fig. 2 Image verification stage

II. Proposed zero-watermarking scheme

The proposed zero-watermarking scheme is composed of the master share generation stage and the image verification stage, as other common zero-watermarking scheme. The master share generation stage is shown by Fig. 1, while the image verification stage is shown by Fig. 2. In both stages, the previously trained CNN is employed as an important part of our scheme. In this section, first we describe the CNN architecture used in the proposed scheme and some hyper-parameters employed in the training stage of the CNN. Next, we describe both stages comprising the proposed scheme.

A. Convolutional Neural Networks

The CNN architecture is presented in Fig.3, which is composed of 13 convolutional layers and two fully-connected layers. Since max-pooling operation decreases the spatial information in the network, we implemented this operation every three convolutional layers (After *Conv3, Conv6,* and *Conv9*). After each convolutional layer, the neural responses are normalized using *Batch Normalization* technique in order to adapt them through the next layer. The activation function implemented in all layers is ReLU, except in the classification layer, where we used Softmax operation. The architecture of the CNN used is given by Fig. 3, which consists of 13 convolutional layers and two fully-connected layer mentioned above. Through these convolutional layers, the inherent image features are extracted. To encode these image feature, we use two fully-connected layers, which are the responsible to generate robust features of input image, specifically, the 100-neurons information contained in the *fc_1* layer. The configuration of the CNN is given by Table I. It is worth noting that for each convolution process, nonlinear function ReLU and Batch normalization operation are carried out.

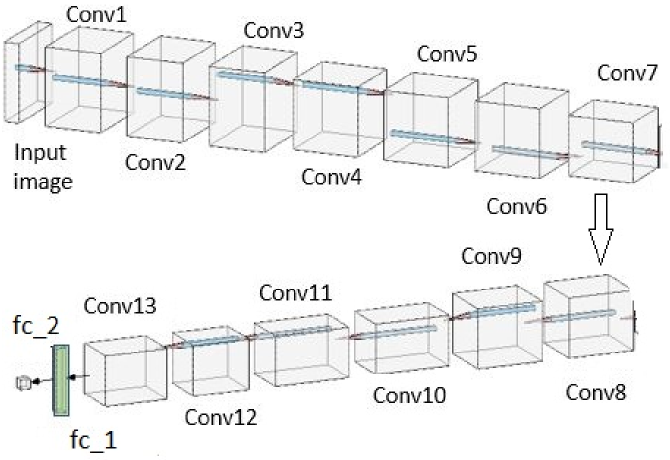


Fig. 3 The architecture of the CNN used in the proposed scheme

To train the CNN from scratch, we used 10,558 images, divided into 2 sets, one for training (7,390 images) and another for validation (3,168 images). We train our CNN in order to recognize two classes (image target and no image target). The training dataset was created applying several image processing, such as compression with different quality factors, filter operations, noise contamination and geometric operations. For evaluation purposes, we created a test dataset attacking the original image Lena using compression, filtering, rotations and noise addition. This test dataset is composed by 49 images.

After several experimentations, the best hyper-parameters were selected for the proposed CNN structure, which is given by Table II. The training was done using a Graphics Processing Units (GPU), NVIDIA GeForce GTX1080, and MATLAB Deep Neural Networks Toolbox.

TABLE I. THE CNN CONFIGURATION

Layer	Configuration
Input	$\mathbf{300} \times \mathbf{300} \times \mathbf{3}$ images
Conv1, Conv2, Conv3	64 $\mathbf{3} \times \mathbf{3}$ convolutions, Stride: 1 Padding: 1
Max-pooling	$\mathbf{2} \times \mathbf{2}$ pooling, Stride:2, Padding:0
Conv4, Conv5, Conv6	128 $\mathbf{3} \times \mathbf{3}$ convolutions, Stride: 1 Padding: 1
Max-pooling	$\mathbf{2} \times \mathbf{2}$ pooling, Stride:2, Padding:0
Conv7, Conv8, Cnv9	128 $\mathbf{3} \times \mathbf{3}$ convolutions, Stride: 1 Padding: 1
Max-pooling	$\mathbf{2} \times \mathbf{2}$ pooling, Stride:2, Padding:0
Conv10, Conv11	256 $\mathbf{3} \times \mathbf{3}$ convolutions, Stride: 1 Padding: 1
Conv12	64 $\mathbf{3} \times \mathbf{3}$ convolutions, Stride: 1, Padding: 1
Conv13	100 $\mathbf{1} \times \mathbf{1}$ convolutions, Stride: 1, Padding: 1
fc_1	100-neuron FC layer
fc_2	2-neuron FC layer
Softmax	

TABLE II. HYPER-PARAMETERS FOR THE CNN

Optimizer	Stochastic gradient descent algorithm
Momentum	0.9
Number of epochs	8
Laerning late	0.0001
Minibatch size	16 images

B. *Master share generation*

Once the CNN is trained, the inherent features of image is extracted from the Fully Connected Layer 1 (*fc_1*) of the CNN, which contains 100 real output data. This output data is converted in binary data using threshold value 0, being positive value equal 1, otherwise 0. The watermark pattern is 10×10 binary matrix, which is given by Fig. 4. The watermark pattern is permuted using owner's secret key to random-like binary image.

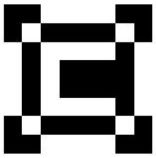


Fig. 4 Binary watermark pattern

The master share is generated by XOR operation between the binary sequence of the inherent image feature and permutated binary watermark pattern.

$$MS = IF \otimes \widehat{W} \tag{1}$$

where IF is binary sequence extracted from fully connected layer *fc_1* layer of the trained CNN, $\widehat{W}$ is the permutated binary watermark sequence, *MS* is the master share and $\otimes$ is XOR operation. The master share is stored by owner in a secure manner.

C. *Image verification stage*

In this stage, the ownership of some images is verified using the image features extracted by the trained CNN and the stored master share *MS*. The inherent feature of the image under analysis can be rapidly obtained, introducing the image into the trained CNN and obtaining 100 real output data from fully connected layer *fc_1*. Then, these data are binarized using the same manner in the Master share generation stage. Using the binary feature and master share, we obtained the permutated watermark sequence, which is given by

$$\widetilde{W} = \widetilde{IF} \otimes MS \tag{2}$$

where $\widetilde{IF}$ is the feature obtained from *fc_1* layer of the CNN, *MS* is the master share and $\widetilde{W}$ is the extracted permutated watermark pattern. Using the same owner's key, the watermark pattern $\dot{W}$ is obtained from $\widetilde{W}$.

III. EXPERIMENTAL RESULTS

To evaluate the robustness of the proposed zero-watermarking scheme, several distorted images are generated, which are not used in the training phase. The distortions or attacks considered in the watermark robustness evaluation are JPEG compression with different quality factor, mean filter, Gaussian smoothing filter, noise contamination.

As evaluation metrics, we used Bit Error Rate (BER) and Normalized Cross Correlation (NCC) to evaluate the fidelity of the recovered watermark sequence $\dot{W}$ respect to its original version *W*. These metrics are given by

$$BER = \frac{Number\ of\ error\ bits}{Total\ bits\ of\ W} \tag{3}$$

$$NCC(W, \dot{W}) = \frac{\sum_{i=1}^{M} \sum_{i=1}^{N} W(i,j) \times \dot{W}(i,j)}{\sum_{i=1}^{M} \sum_{i=1}^{N} W(i,j)^2} \tag{4}$$

where W is the original watermark pattern, $\dot{W}$ is the recovered watermark pattern, *M* and *N* are dimension of the watermark pattern, in our case, *N*=*M*=10. Although the BER is enough to evaluate the recovered watermark fidelity respect to the original one in the binary watermark case, the NCC metric is used commonly in zero-watermarking scheme, so values of this metric are added in robustness evaluation.

Tables III, IV show the evaluation results of the robustness to the JPEG compression with different quality factors, filter operation by the mean filter with different window sizes and the Gaussian smoothing filter with different variances. The Peak Signal Noise Ratio (PSNR) values indicate the distortion level of the attacked image respect to the original one. We can observe from these tables, the proposed scheme provides watermark robustness after JPEG compression and filtering process. Table V shows some attacked image, recovered watermark pattern after the corresponding attacks and NCC values.

TABLE III. ROBUSTNESS OF THE WATERMARK PATTERN TO COMPRESSION

Quality Factor	PSNR (dB)	BER	NCC
0	22.33	2%	0.9696
10	29.29	2%	0.9696
20	31.43	2%	0.9696
30	32.17	2%	0.9696
40	32.47	2%	0.9696
50	32.68	2%	0.9696
60	32.80	2%	0.9696
70	32.94	2%	0.9696
80	33.02	1%	0.9848
90	33.13	1%	0.9848
100	33.15	0%	1

TABLE IV. ROBUSTNESS OF THE WATERMARK PATTERN AGAINST FILTERING

Filtering	PSNR (dB)	BER	NCC
Mean $\mathbf{3 \times 3}$	29.37	0%	1
Mean $\mathbf{5 \times 5}$	26.52	0%	1
Mean $\mathbf{7 \times 7}$	24.77	1%	0.9848
Mean $\mathbf{9 \times 9}$	32.49	1%	0.9848
Gaussian $\boldsymbol{\sigma} = \mathbf{1}$	31.27	0%	1
Gaussian $\boldsymbol{\sigma} = \mathbf{2}$	27.87	0%	1
Gaussian $\boldsymbol{\sigma} = \mathbf{3}$	25.90	1%	0.9848
Gaussian $\boldsymbol{\sigma} = \mathbf{4}$	24.61	2%	0.9696

TABLE V. ATTACKED IMAGE AND RECOVEED WATERMARK PATTERN

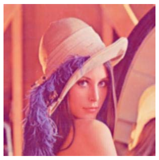
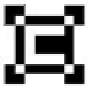
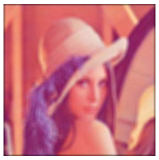
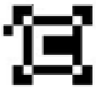
Attack	Attacked image	Recovered watermark	NCC
JPEG FC=0 (PSNR=22.33dB)			0.9696
Gaussian σ=1 (PSNR=31.27 dB)			1.0
Mean 9×9 (PSNR=32.49 dB)			0.9848
3° Rotation (PSNR=13.15 dB)			0.9545

Table VI and VII show the robustness of watermark pattern against noise contamination by Gaussian noise and rotation with different rotation angles.

TABLE VI. ROBUSTNESS OF THE WATERMARK PATTRN AGAINST NOISE CONTAMINATION

Noise	PSNR	BER	NCC
Gaussian 0.001	31.39	0%	1
Gaussian 0.002	29.55	3%	0.9545
Gaussian 0.01	22.68	6%	0.9242
Impulsive 0.01	27.27	2%	0.9848
Impulsive 0.02	24.70	2%	0.9848
Impulsive 0.03	23.02	4%	0.9394

TABLE VII. ROBUSTNESS OF THE WATERMARK PATTERN AGAINST TO ROTATION

Angle	PSNR	BER	NCC
1°	17.14	4%	0.9394
2°	14.58	6%	0.9242
3°	13.15	7%	0.9091
4°	12.12	7%	0.9091
5°	11.35	7%	0.9091

From these tables, the proposed scheme provides a sufficient robustness to noise contamination and rotation attacks. In almost all watermarking algorithms, the rotation attacks are considered as one of the most aggressive attacks due to the loss of synchronization between the watermark embedding stage and the watermark extraction stage. We consider that the proposed scheme can provide a higher robustness to the geometrical distortion such as rotation, scaling and any affine transformation, if the image normalization technique [14] can be applied as pre-processing of the CNN process.

IV. CONCLUSIONS

In this paper, we proposed a CNN-based zero-watermarking scheme, in which inherent features of host image are extracted from a first fully-connected layer of the trained CNN. The extracted features are binarized and logically linked with the owner's watermark pattern by XOR operation to generate master share, which is stored in secure manner. When an image is in the ownership dispute, the image is introduced to the trained CNN as input data, and its inherent features are extracted. The XOR operation is applied to the extracted image features and the master share to recover the watermark pattern. Although the training process of the CNN is time-consuming process, once the CNN is trained, the feature extraction of an image is instantaneous, it takes approximately 0.5 second, under GeForce GTX1080 environment.

The watermark robustness against several attacks and common image processing, such as JPEG compression, filtering operations, noise contamination and rotation, is evaluated. The experimental results show the sufficient robustness of the proposed scheme, even though attacked images are highly distorted with 22dB of the PSNR respect to its original version.

Recently, zero-watermarking is used efficiently in telemedicine or teleradiology application, taking advantage of the distortion-free embedding property of this watermarking technique. The principal inconvenience of the conventional zero-watermarking is its high computational complexity in the verification stage, in which robust inherent features must be extracted from each input medical image. In the proposed CNN-based zero-watermarking, the extraction of the inherent features is performed through the trained CNN, which performs instantaneously. So, we consider that the proposed zero-watermarking scheme will be suitable for above mentioned applications. Evaluation and possible improvement of the proposed zero-watermarking scheme is considered as a future work.

REFERENCE

[1] V. Y. Wang, J. F. Doherty and R. E. Van Dyck, “A Wavelet-based Watermarking Algorithm for Ownership Verification on Digital Images,” IEEE Trans. on Image Processing, vol. 11, no. 2, pp. 77-88, August 2002.

[2] S. D. Lin, C. F. Chen, “A robust dct-based watermarking for copyright protection”, IEEE Trans. on Consumer Electronics, vol. 46, no. 3, pp. 415-421, August 2000.

[3] Y. H. Chen and H. C. Huang, “Coevolutionary genetic watermarking for owner identification”, Neural Computing and Applications, vol. 26, no. 2, pp. 291-298, February 2015.

[4] K. Rangel, E. Fragoso, C. Cruz, R. Reyes, M. Nakano and H. Perez, “Adaptive removable visible watermarking technique using dual watermarking for digital color images”, Multimedia Tools and Applications, vol. 77, no. 11, pp. 1347-1374, June 2018.

[5] L. Rosales, M. Cedillo, M. Nakano, H. Perez and B. Kurkoski, “Watermarking-based image authentication with recovery capability using halftoning techniques”, Signal Processing:Image communication, vol. 28, no. 1, pp. 69-83, January 2013.

[6] X. Qi and X. Xia, “A Singular-value-based semi-fragile watermarking scheme for image content authentication with tamper localization”, Journal of Visual Communication and Image Representation, vol. 30, pp. 312-327, July 2015.

[7] A. M. Alattar, “Reversible Watermarking using the Difference Expansion of a Generalized Integer Transform”, IEEE Trans. on Image Processing, vol. 13, no. 8, pp. 1147-1156, July 2004.

[8] I. C. Dragoi, D. Coltuc, “Local-prediction-based Difference Expansion Reversible Watermarking”, IEEE Trans. on Image Processing vol. 23, no. 4, pp. 1779-1790, February 2014.

[9] Y. Zhou and W. Jin, “A Novel Image Zero-watermarking Scheme based on DWT-SVD”, Proc. on Int. Conf. on Multimedia Technology (ICMT 2011), pp. 2873-2876, July 2011.

[10] A. Sinfh and M. K. Duta, “Lossless and Robust Digital Watermarking Scheme for Retinal Images”, Proc. on Int. Conf. on Computational Intelligence and Communication Technology (CICT 2018), pp 1-5, February 2018.

[11] V. Seenivasagam and R. Velumani, “A QR Code Based Zero-Watermarking Scheme for Autehtication of Medical Image in Teleradiorogy Cloud”, Computational and Mathematical Methods in Medicine, vol. 2013, ID 516465, 2013.

[12] G. Gao and G. Jiang, “Bessel-Fourier Moment-based Robust Image Zero-Watermarking”, Multimedia Tools and Applications, vol. 74, no. 3, pp. 841-858, January 2015.

[13] C. P. Wang, X. Y. Wang, Z. Q. Xia, C. Zhang, X. J. Chen, “Geometrically Resilient Color Image Zero-watermarking Algorithm Based on Quaternion Exponent”, Journal of Visual Communication and Image Representation, vol. 41, pp. 247-259, July 2016.

[14] M. K. Hu, “Visual Pattern Recognition by Moment Invariants”, IRE Trans. on Information Theory, vol. 8, no. 2, pp. 179-187, February 1962.