



Research article

A batch copyright scheme for digital image based on deep neural network

Haoyu Lu¹, Daofu Gong^{1,*}, Fenlin Liu¹, Hui Liu² and Jinghua Qu¹

¹ Zhengzhou Science and Technology Institute, Zhengzhou, 450001, China

² University of Surrey, Guildford, Surrey, GU2 7XH, UK

* **Correspondence:** Email: gongdf@aliyun.com.

Abstract: Digital signature and watermarking are effective image copyright protection techniques. However, these methods come with some inherent drawbacks, including the incapacity of carrying information and inevitable fidelity loss, respectively. To improve this situation, this paper proposes a neural network-based image batch copyright protection scheme, with which a copyright message bitstream can be extracted from each registered image while no modifications are introduced. Taking advantage of the pattern extraction capability and the error tolerance of the neural network, the proposed scheme achieves perfect imperceptibility and superior robustness. Moreover, the network's preference for diverse data content makes it especially appropriate for multiple images copyright verification. These claims will be further supported by the experimental results in this paper.

Keywords: digital image; copyright protection; deep neural network; robust feature extraction; digital watermarking

1. Introduction

Copyright protection of multimedia contents have become a great concern over the decades and it also developed a series of techniques, the most notable of which are the cryptographic schemes and digital watermarking methods. The cryptographic schemes, also called digital signature methods, are usually based on a hashing function [1] or image feature extraction [2–5] followed by a public key encryption system. Unfortunately, the digital signature-based methods are mainly used for integrity authentication rather than ownership protection, because no explicit copyright information can be extracted from digital products. On the other hand, digital watermarking methods embed a watermark into the image without causing significant damage to its usage, and later it can be extracted from the watermarked image for ownership verification. Generally speaking, an effective watermarking scheme should be perceptually invisible and robust to possible attacks including signal processing, geometric distortion, and intentional manipulations.

In recent years, robust watermarking systems has become an intriguing research topic, among them are the spatial domain methods [6–15] and the transform domain methods [16–22]. Either way, though, in order to gain robustness, as Cox’s prominent argument in [23], the destination of the watermark should be the perceptually most significant components of the carrier signal, despite the risk of potential fidelity distortions. Obviously, this criterion would lead to a counter-balance paradox between robustness and imperceptibility, since improving the robustness will have a notable impact on the visual quality of the image. To alleviate this problem, many adaptive watermarking embedding algorithms emerged, including watermark embedding strength adaption [24–26] and embedding position adaption [27], etc. However, neither of these methods has solved this issue with finality. Additionally, most of the existing robust watermarking methods essentially design or utilize some stable features of the image to embed the watermark. For example, transform domain methods typically embed the watermark in the DCT coefficients or the wavelet subbands of the image because they are relatively stable. The transform process can be regarded as the robustness feature extraction from the original image. Nevertheless, artificially designing and extracting robust features is no easy task, especially when the method is supposed to resist one or several particular attacks.

Based on these considerations, we propose a new scheme for multiple digital image copyright protection to displace the traditional image watermarking architecture by resorting to the deep neural network technique, which has achieved superior performance on feature extraction and representation [28–30]. In order to get around the paradox between robustness and imperceptibility, we no longer modify the image but represent the copyright message by exploring the patterns of image data itself through a neural network. To be specific, a large number of selected image blocks along with the message bits as their corresponding labels will be feed into the network, enabling the network to extract copyright message from images by classifying the image blocks to target labels (bits). Moreover, to enhance the robustness, we provide a preventive strategy to further train the network, namely, feeding image blocks from the post-attacked images to the network. By doing this, the neural network could automatically extract those classification features that are robust to certain attacks. The trained neural network can then be stored in a database (or added to the image header) for future use in ownership verification. Finally, we found the neural network-based scheme is particularly suitable for multiple images copyright protection. Based on experiments, a well-designed neural network is capable of representing the copyright message for a considerable number of images under various intentional attacks. Superior robustness and perfect imperceptibility are achieved in the proposed scheme.

This paper is organized as follows. Section 2 outlines the proposed system for image copyright protection. Some discussion and analysis for the scheme are described in Section 3. The experiments in Section 4 suggest our scheme is of considerable load capacity and strong robustness to a variety of attacks. Finally, Section 5 concludes the paper and discuss possible enhancements to the proposed scheme.

2. Proposed copyright protection scheme

This section will elaborate on the proposed scheme in detail. The image copyright protection scheme consists of two parts: copyright image registration where we train a neural network for a number of given images, and copyright verification where the stored neural network extracts

copyright message from the images to be verified.

2.1. Copyright images registration

The neural network in our scheme can be partly analogous to the hashing procedure in general digital signature systems, in the sense that the neural network serves as a mapping from images to discrete numeric values (digests). But we hope that our scheme could not only reduce the data but also name an explicit message bitstream for copyright verification. We then train a neural network to transform the image blocks into copyright message. The procedure for training is showed in Figure 1:

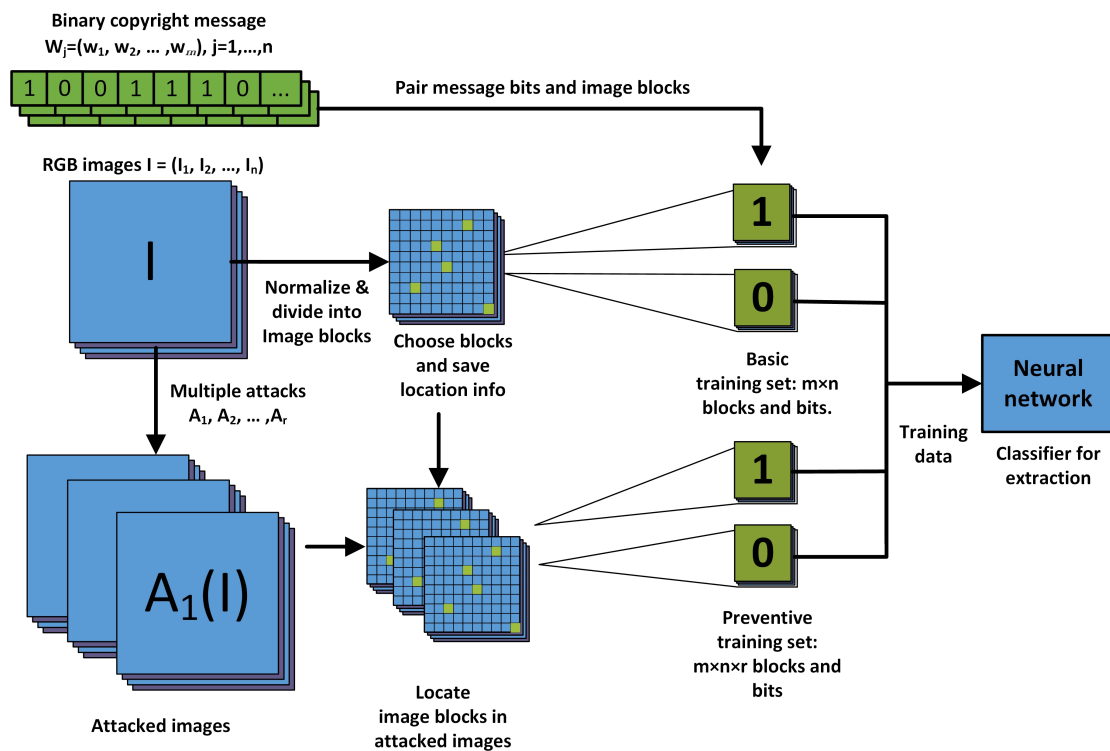


Figure 1. Network training procedure.

In a most common application scenario, our scheme starts with a series of RGB images $I = (I_1, I_2, \dots, I_n)$ to be registered to our verification system and binary sequences of copyright message for each image $W_j = (w_1, w_2, \dots, w_m)$, $w_i \in \{0, 1\}$, $j = 1, 2, \dots, n$. In practice, the choice of n and m depends on the capacity of the network. In general, as the number of images and the copyright message length are increased, the optimal network structure to which should be altered. In this paper, we adopted a 5-layer fully connected neural network structured in Figure 2. The choice of network is further elaborated in Section 3.4.

Before training the network, expanded training data is yield from an attack set $A = A_1, A_2, \dots, A_r$ including r types of attack, the proposed A is showed in Table 1. For each image I_k , the corresponding attacked image set is denoted by $A(I_k) = \{A_1(I_k), A_2(I_k), \dots, A_r(I_k)\}$.

To determine the location of the candidate training data, I is processed as follows: Normalize the

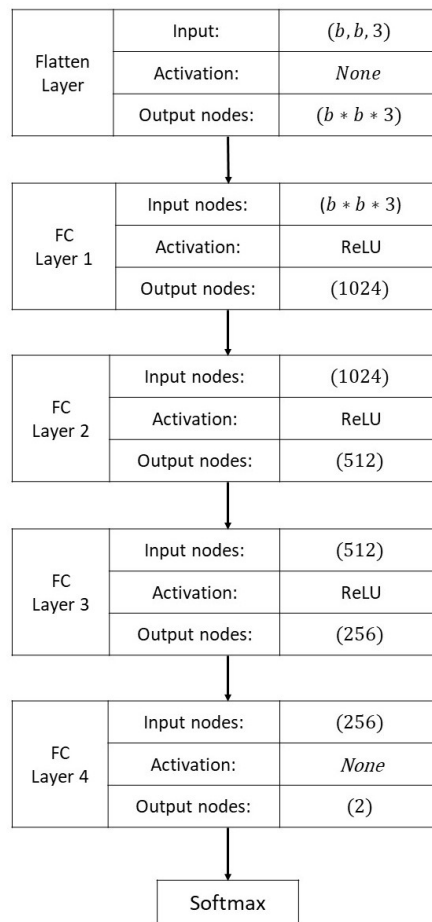


Figure 2. Proposed neural network structure.

size of all the images in I to $M \times M$, and then divide each image into non-overlapping blocks of size $b \times b$. For each image, calculate keypoints with diameter b of these normalized images, and locate m strongest keypoints at block scale. In other words, for each image, we choose the m image blocks containing the strongest keypoints and then save the location information of these chosen image blocks. These m image blocks are supposed to be the input of the network and the corresponding output should be the m message bits.

Then we normalize the images in the attacked image set $A(I_k), k = 1, \dots, n$ and choose the corresponding image blocks with the saved location information. Taking all n images in I into account, we train the neural network to map $nm(r + 1)$ image blocks to the n copyright message bitstreams for n images. The trained neural network can then be stored in a database (or added to the image header) for ownership verification.

2.2. Copyright verification

During the copyright verification stage, the copyright owner use the trained network to extract copyright messages from the registered images to declare the ownership.

Table 1. Attacks included in the training set.

Notation	Type of attack
A_1	Superposition Gaussian noise with $\mu = 0, \sigma = 0.2$
A_2	Superposition Gaussian noise with $\mu = 0, \sigma = 0.2$
A_3	Superposition Gaussian noise with $\mu = 0, \sigma = 0.2$
A_4	JPEG Compression with quality factor 50
A_5	JPEG Compression with quality factor 30
A_6	JPEG Compression with quality factor 10
A_7	Mean value filtering with kernel size 7×7
A_8	Median filtering with kernel size 5×5
A_9	Gaussian filtering with kernel size 7×7
A_{10}	Resize image to 64×64 (bilinear interpolation)

Continuing from the application scenario above, the image distributor registered an image on a neural network, let's say I_c , which has possibly been through some malicious or non-malicious manipulations. As Figure 3 shows, in order to feed the trained network N_{IW} , we should first normalize the image I_c to size $M \times M$ and then divide it into $b \times b$ non-overlapping blocks as the registration stage does. It is worth noting that during verification stage the distributor has already known the location information of the registered image blocks, and thus making it possible for distributor to locate the image blocks and feed them to the network sequentially. After this, the trained neural network N_{IW} , serving as a copyright message extractor, is able to yield a bit sequence W_c from the possibly distorted registered image I_c , and W_c is supposed to be the copyright message for image I_c to verify the copyright ownership.

3. Discussion and analysis

In this section, we give some explanation and discussion on the overall and details of the design scheme.

3.1. Feasibility

The universality theorem [31, 32] ensures that a neural network can fit any mapping from image blocks to the classification labels, as long as there is no such case where identical image blocks are mapped to different labels. Actually, this unexpected case is possible in our training set, in another word, it is possible that two identical image blocks (or highly similar image blocks) are chosen to participate in the training process and unfortunately they are paired with two different message bits. However, we have taken the following measures to avoid this block collision case as much as possible.

- Choosing the keypoint-included blocks reduces the probability of such coincidence. The more complex of an image block is, the less possible it coincides with another image block. What's more, there are much more similar simple areas than the complex areas in the natural image samples, thus, those areas containing complex contents are preferable in our scheme.
- Controlling the number of candidate image blocks also lessens the co-occurrence possibility of two similar image blocks. Our choices are based on the complexity of the image blocks, in this manner, as the number of blocks raises, the chance of simple blocks' co-occurrence raises.

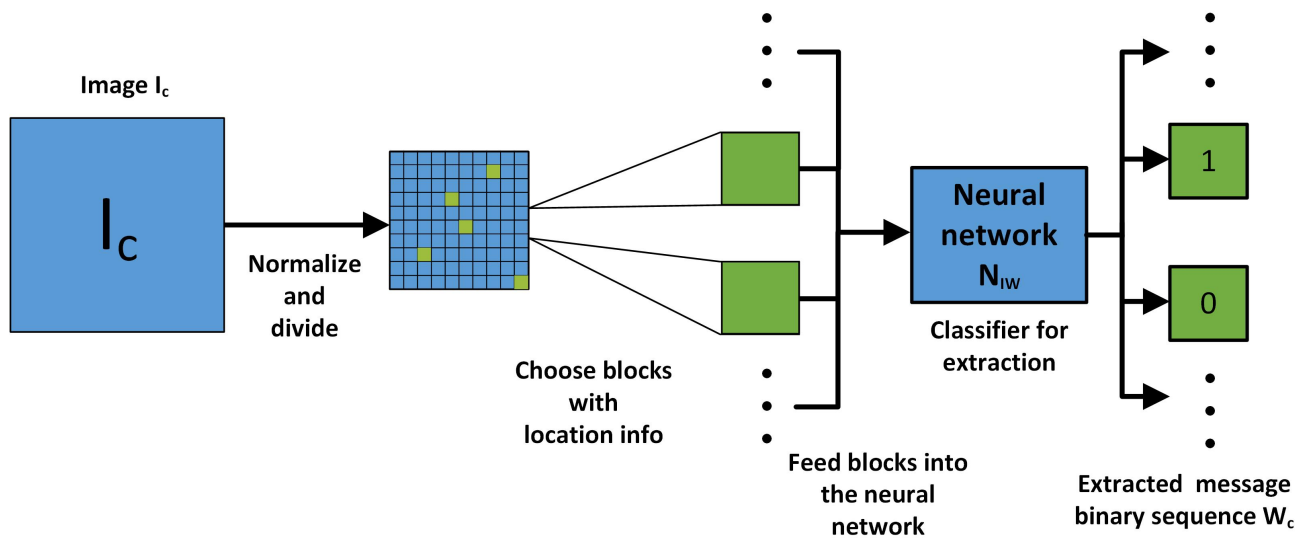


Figure 3. Copyright message extraction procedure.

- Likewise, an appropriate block size also suppresses the possibility of such coincidence. An undersized block may lead to more similarities in content. However, an oversized block may decrease the copyright message capacity of each image. For this reason, we determined an appropriate block size to further optimize the feasibility.

With the neural network's universality theorem and the collision avoidance rules listed above, we are able to guarantee the feasibility of the proposed scheme (See Section 4 for detailed settings). But still, we hope the trained network would be equipped with robustness for representing copyright message to adapt to lossy channel and resist malicious manipulation.

3.2. Robustness

The robustness of the scheme comes from three parts, the error tolerance of the neural network, the preventive training set, and the robust feature extraction. Firstly, error tolerance is the inherent nature of the proposed neural network. The neural network essentially works as a soft classifier which gives a probabilistic, rather than a hard yes/no classification result, allowing the disturbed signals to be classified to the expected class.

Secondly, since the expanded training set has already contained varied attacked images, a well trained neural network guarantees perfect extraction of those image blocks in the training set. So our scheme is capable of resisting certain attacks that contains in the preventive training set. Furthermore, the robustness of the proposed scheme can be enhanced by adding new attacked images to the preventive training set. As such, it enables a dynamic evolution of robustness according to practical demands.

Finally, our scheme also gets its robustness from the generalization ability of the neural network. Above all, the process that the neural network maps the original image block and the attacked image

block to the same label is exactly the process where the neural network extracts the robust features and then classifies them. In this way, even for those attacked images that are not included in the training set, the trained network is still able to perform a correct classification by recognizing the robust features.

3.3. Network capacity for multiple images

The capacity of our scheme is plainly defined as the number of “image block-message bit pair” that the neural network could effectively enroll. Due to transmission and storage considerations, we expect a single neural network could register as many images as possible. Interestingly, we have experimentally determined such a fact that choosing the same number of blocks from many disparate images instead of a small number of images, will significantly enlarge the network’s capacity. This is partly because choosing blocks from various images would provide more diverse pixel distributions, and then the image blocks probabilistically could have more identifiable features for the neural network. In conclusion, to train the network in an economical manner, we prefer a shorter copyright message for each image but a richer image set to be registered by the network.

Of particular note is that the robustness and capacity of our scheme also depend closely on the network structure and training algorithm, but confined to the length of the paper, we only focus on the network structure proposed above. All experiments in Section 4 that will further explore the robustness and capacity of our scheme are based on this structure.

3.4. Network choice

Our choice of fully connected neural network rather than the more effective CNN in image feature extraction is out of the resource utilization and classification optimization considerations. Firstly, as the proposed method is based on image blocks, to increase the length of the represented copyright message, the block size should not be too large otherwise there will not be sufficient room to cover copyright message length (expanding normalization size can also increase the number of image blocks, but we expect more ‘content’ in a block rather than simply more pixels). However, on a smaller scale we can no longer bring CNNs superiority into full play. To be specific, when the block size is 8×8 as Section 4 does, the convolutional kernel size of CNN can either be too small to capture the spatial correlation details among adjacent region or be too large to effectively reduce the number of parameters.

Moreover, our classification task might become rigorous at times. For example, sometimes we need the network be able to distinguish one image block from another highly similar one (See co-occurrence occasion described in Section 3.1) where leaving out any of the details can bring about classification failure, thus we dropped weight sharing and local connection to avoid any information loss and finally chose fully connected neural network to extract robust features and represent the copyright message.

4. Experiment

In this section, experiments are performed to evaluate the proposed scheme, which should be able to extract a copyright message from each registered image, even though the image could have suffered disturbance or damage. We take a number of RGB images from “Standard” test images and BossBase-1.0 as copyright images (Figure 4 shows some of them), and pseudo-random binary sequences as copyright messages. For each image we take 64 bits as its copyright message. All the experiments were

performed on Windows PC with Intel® Core™ I7-4720 CPU at 2.6GHz, 8GB RAM with Tensorflow version 1.6.0.

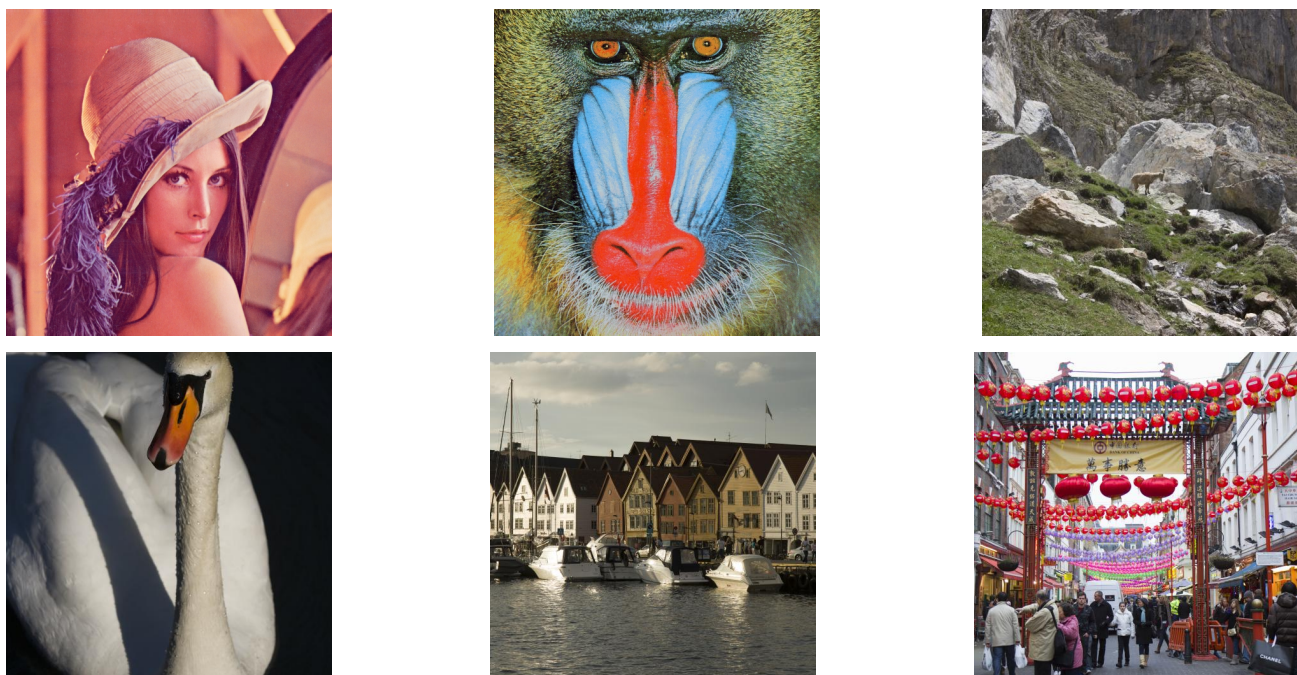


Figure 4. Some of the RGB images with size of 512×512 in our experiment.

The network architecture used in this paper is shown in Figure 2. The input image block is first flattened and then sequentially passed through four layers of gradually narrowed fully connected networks. Each layer is activated by the ReLU function, and the final output of the two nodes passes through the Softmax layer to obtain the probability of corresponding classification labels. The mini-batch size used for training is 32, and Adam is adopted to optimize the network for 600 training epochs on the training set.

Based on the analysis and discussion in Section 3, this part mainly shows the general resistance of the proposed scheme to various attacks, including the superposition Gaussian noise with zero mean and standard deviation ranging from 5% to 30% of the maximum pixel value 255, JPEG compression with quality factor 70, 50, 30 and 10, resizing and various filtering attacks such as Gaussian filtering and median filtering (some of the attacked sample images are illustrated in Figure 5). To further explore the capacity of our network, different numbers of images including 10, 20, 30, 40, 90 images are registered to 5 neural networks respectively in the following experiments, and then we evaluate the each network's extraction accuracy under different attacks. It should be noted that each image corresponds to an individual 64-bit copyright message. Therefore, we are actually testing the neural network's capability of 640, 1280, 1920, 2560 and 5760 bits copyright messages. In general, we take the ratio of the number of correctly extracted message bits to the total number of bits as extraction accuracy to evaluate the performance.

Table 2 shows the average extraction accuracy rate after superposition Gaussian additive noise

attacked the images, for each image, we register 64 bit copyright message for it. The noise is normally distributed with zero mean and standard deviation Std, which ranges from 0 to 76.5 in the experiment. At a larger Std, as the numbers of images participated in the training raises, the robustness performance presents a regular decline trend. Table 3 shows the robustness against JPEG compression. In the case of lower compression quality factors, the extraction accuracy rate reduced slightly. Even for those quality factors that are not involved in training, the trained network can still achieve a reasonably high extraction accuracy rate. Table 4 shows the extraction accuracy rate for three kinds of filtering attacks including mean value filtering, Gaussian filtering, and median filtering. With the same size of the filtering kernel, the extraction accuracy of the Gaussian filtering is relatively higher than the other two methods, since Gaussian filter impacts less to the image than the mean and median filters. Table 5 shows the robustness against rescaling attacks. The extraction is almost unaffected when the rescaling size is larger than the normalization size, which is by no means difficult to foretell. When the rescaling size is smaller than the normalization size, the extraction accuracy reduced considerably at 20% and 12.5% rescaling level. Still, expanding the training set can effectively improve the reduction.



Figure 5. Some of the attacked images in the test: (a) Original image (512×512), (b) JPEG compressed with $qf = 30$, (c) JPEG compressed with $qf = 10$, (d) Gaussian noise with $\mu = 76.5$, (e) Gaussian filtering with kernel size 11×11 , (f) Median filtering with kernel size 11×11 .

*The underlined items are included in the attack set during the training process.

Table 2. Extraction accuracy rate after adding Gaussian noise. *

Number of bits\Std.	0(no attack)	11.25(5%)	22.5(10%)	51(20%)	76.5(30%)
640 for 10 images	1	1	0.984	0.972	0.911
1280 for 20 images	0.989	0.989	0.984	0.950	0.884
1920 for 30 images	0.997	0.997	0.992	0.950	0.887
2560 for 40 images	0.998	0.997	0.991	0.953	0.889
5760 for 90 images	0.988	0.986	0.976	0.916	0.831

Table 3. Extraction accuracy rate after JPEG compression.

Number of bits\JPEG QF	100(no attack)	70	50	30	10
640 for 10 images	1	1	1	1	0.984
1280 for 20 images	1	1	0.990	0.990	0.978
1920 for 30 images	0.997	0.997	0.997	0.997	0.990
2560 for 40 images	0.997	0.997	0.997	0.997	0.987
5760 for 90 images	0.988	0.976	0.976	0.976	0.951

Table 4. Extraction accuracy rate after various filtering attacks.

Filter Kernel size	Mean value filtering		Gaussian filtering			Median filtering	
	7×7	9×9	7×7	9×9	11×11	3×3	5×5
640/10	0.997	0.969	1	0.997	0.964	1	0.986
1280/20	0.977	0.930	1	0.989	0.962	1	0.982
1920/30	0.983	0.945	0.997	0.993	0.969	0.997	0.985
2560/40	0.962	0.933	0.995	0.977	0.932	0.996	0.972
5760/90	0.956	0.893	0.986	0.972	0.923	0.990	0.968

Table 5. Extraction accuracy rate after resizing attacks.

Rescaling rate	50%	25%	20%	12.5%
Rescaled size	256×256	128×128	100×100	64×64
640/10	1	1	0.989	0.850
1280/20	0.989	0.989	0.978	0.741
1920/30	0.997	0.997	0.984	0.739
2560/40	0.997	0.997	0.967	0.672
5760/90	0.988	0.988	0.949	0.672

5. Conclusion

This paper proposes a deep neural network based large scale image copyright protection scheme. Instead of modifying the original image to embed the copyright message as the traditional watermarking system, the proposed scheme trains a neural network to register multiple images, and when copyright verification is required, the neural network could represent a copyright message for each image by classifying image blocks to message bits. With the error tolerance of the neural network and a preventive training strategy, the proposed scheme is remarkably robust to many attacks, including additive noise, JPEG compression, filtering and resizing. Moreover, this scheme is especially appropriate for multiple images' copyright verification as experimental results showed. Lastly, the preventive method can be modified for specific attacks in practice, thus a stronger robustness can be obtained by expanding the preventive training set. For now, our scheme is only based on a single network structure which clearly limited its robustness and capacity. To further improve its performance with alternative network structure will be a promising study in the future.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61772549, 61872448, U1736214, 61602508, and 61601517), and the National Key R&D Program of China(No. 2016YFB0801303 and 2016QY01W0105).

Conflict of interest

The authors declare no conflict of interest.

References

1. G. L. Friedman, The trustworthy digital camera: Restoring credibility to the photographic image, *IEEE T. Consum. Elec.*, **39** (1993), 905–910.
2. S. Bhattacharjee and M. Kutter, Compression tolerant image authentication, *International Conference on Image Processing*, (1998), 435–439.
3. J. Dittmann, A. Steinmetz and R. Steinmetz, Content-based digital signature for motion pictures authentication and content-fragile watermarking, *IEEE International Conference on Multimedia Computing and Systems*, (1999), 209–213.
4. C. Lin and S. Chang, Robust image authentication method surviving jpeg lossy compression, *Storage and Retrieval for Image and Video Databases VI. International Society for Optics and Photonics*, **3312** (1997), 296–308.
5. C. S. Lu and H. Liao, Structural digital signature for image authentication: an incidental distortion resistant scheme, *IEEE T. Multimedia*, **5** (2003), 161–173.
6. D. Gong, F. Liu and X. Luo, An image edge based robust watermarking algorithm, *Sci. Sin. Inform.*, **43** (2013), 1410–1430.

7. H. Qi, D. Zheng and J. Zhao, Human visual system based adaptive digital image watermarking, *Sig. Proc.*, **88** (2008), 174–188.
8. J. Seo and C. Yoo, Image watermarking based on invariant regions of scale-space representation, *IEEE T. Sig. Proc.*, **54** (2006), 1537–1549.
9. Y. Wang, R. Ni, Y. Zhao, et al., Watermark embedding for direct binary searched halftone images by adopting visual cryptography, *Comput. Mater. Continua*, **55** (2018), 255–265.
10. J. Tsai, W. Huang, C. Chen, et al., A feature-based digital image watermarking for copyright protection and content authentication, *IEEE International Conference on Image Processing*, **5** (2007), V–469.
11. Y. Zhang, D. Ye, J. Gan, et al., An image steganography algorithm based on quantization index modulation resisting scaling attacks and statistical detection, *Comput. Mater. Continua*, **56** (2018), 151–167.
12. Y. Lin, C. Huang and G. Lee, Rotation, scaling, and translation resilient watermarking for images, *IET Image Proc.*, **5** (2011), 328–340.
13. J. Tsai, W. Huang, Y. Kuo, et al., On the selection of optimal feature region set for robust digital image watermarking, *IEEE T. Image Proc.*, **20** (2011), 735–743.
14. M. Kutter, F. Jordan, F. Bossen, et al., Digital watermarking of color images using amplitude modulation, *J. Elec. Imaging*, **7** (1998), 326–333.
15. D. Gong, Y. Chen, H. Lu, et al., Self-embedding image watermarking based on combined decision using pre-offset and post-offset blocks, *Comput. Mater. Continua*, **57** (2018), 243–260.
16. T. Sridevi, T. Ramya, Y. Ramadevi, et al., A robust watermarking algorithm based on image normalization, in *2010 Second Vaagdevi International Conference on Information Technology for Real World Problems (VCON)*, (2010), 70–75.
17. K. Liu, Wavelet-based watermarking for color images through visual masking, *AEU-Int. J. Electron. C.*, **64** (2010), 112–124.
18. N. Kalantari, S. M. Ahadi and M. Vafadust, A robust image watermarking in the ridgelet domain using universally optimum decoder, *IEEE T. Circ. Syst. Vid.*, **20** (2010), 396–406.
19. C. Lin, M. Wu, J. Bloom, et al., Rotation, scale, and translation resilient watermarking for images, *IEEE T. Image Proc.*, **10** (2001), 767–782.
20. X. Kang, J. Huang, Jiwu, Y. Shi, et al., A dwt-dft composite watermarking scheme robust to both affine transform and jpeg compression, *IEEE T. Circ. Syst. Vid.*, **13** (2003), 776–786.
21. X. Kang, J. Huang and W. Zeng, Efficient general print-scanning resilient data hiding based on uniform log-polar mapping, *IEEE T. Inf. Foren. Sec.*, **5** (2010), 1–12.
22. D. Rosiyadi, S. Horng, P. Fan, et al., Copyright protection for e-government document images, *IEEE MultiMedia*, **19** (2012), 62–73.
23. I. Cox, M. Miller, J. Bloom, et al., in *Digital watermarking and steganography*, Morgan kaufmann, (2007).
24. H. Al-Otum and A. Al-Taba'a, Adaptive color image watermarking based on a modified improved pixel-wise masking technique, *Comput. Electr. Eng.*, **35** (2009), 673–695.

25. H. Song, S. Yu, X. Yang, et al., Contourlet-based image adaptive watermarking, *Sig. Proc. Image Comm.*, **23** (2008), 162–178.
26. L. Ma, D. Yu, G. Wei, et al., Adaptive spread-transform dither modulation using a new perceptual model for color image watermarking, *IEICE T. Inf. Sys.*, **93** (2010), 843–857.
27. C. Song, S. Sudirman and M. Merabti, Robust digital image watermarking using region adaptive embedding technique, *2010 IEEE International Conference on Progress in Informatics and Computing (PIC)*, **1** (2010), 378–382.
28. L. Wu, Y. Wang, J. Gao, et al., Deep adaptive feature embedding with local sample distributions for person re-identification, *Pat. Recog.*, **73** (2018), 275–288.
29. L. Wu, Y. Wang, X. Li, et al., What-and-where to match: deep spatially multiplicative integration networks for person re-identification, *Pat. Recog.*, **76** (2018), 727–738.
30. L. Wu, Y. Wang, J. Gao, et al., Where-and-when to look: Deep siamese attention networks for video-based person re-identification, *IEEE T. Multimedia*, **21** (2019), 1412–1424.
31. G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signal.*, **2** (1989), 303–314.
32. K. Hornik, M. Stinchcombe and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks*, **2** (1989), 359–366.



AIMS Press

©2019 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)