Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09555-5

A robust hybrid digital watermarking technique against a powerful CNN-based adversarial attack

Sai Shyam Sharma[1] · **V. Chandrasekaran**[1]

Received: 2 September 2019 / Revised: 24 June 2020 / Accepted: 6 August 2020 /
Published online: 29 August 2020

Abstract
Digital watermarking techniques are valuable tools to embed digital signatures on multimedia content to establish the legal ownership and authenticity claims by the owners. Firstly this paper investigates the robustness of popular transform domain-based digital image watermarking schemes such as DCT, SVD, DWT, and their hybrid combinations against known image processing type attacks such as image blurring, compression, noise addition, rotation and cropping. Then, an enhanced hybrid scheme using DWT and SVD methods is proposed and its improved performance is demonstrated in terms of the quality of the extracted watermarks measured in terms of PSNR, SSIM and NCC values. This paper then proposes a novel adversarial attack based on a powerful Deep Convolutional Neural Network based Autoencoder(CAE) scheme. The CAE is specifically chosen to exploit its intrinsic capability to represent the image content (spatial and structural) through lower dimensional projections in the intermediate layers. The CAE is trained and tested on the entire image repository of the CIFAR10 data set. Once CAE is trained on a class of images and the parameters are frozen, it will serve as a system to produce a perceptually close image for any unseen input image belonging to the same class. The power of the proposed adversarial attack scheme is shown in terms of the quality of extracted watermarks against popular water mark embedding schemes. Finally the proposed enhanced hybrid strategy of DWT+SVD is shown to be robust against the new form of attack and outperforms all other techniques measured in terms of its high quality watermark extraction.

Keywords Digital watermarking · Convolutional autoencoder · Copyright protection · Adversarial attacks · Hybrid transforms

1 Introduction

The proliferation of multimedia data in the form of images and videos are a boon and a curse. WhatsApp, Facebook, Instagram, Twitter and Snapchat are some of the most popular platforms that have millions of active users contributing petabytes of image data on a

✉ Sai Shyam Sharma
saishyamsharma@gmail.com

Extended author information available on the last page of the article.

daily basis. Easy misuse of this data emphasizes the relevance of image forensics today. The technologies used for representation, storage and transmission of multimedia files are a major challenge for media forensics [8]. Digital watermarking and steganography methods have been used for multimedia security. Both steganography and digital watermarking hide information in the cover data. Steganography techniques are applied in covert communication or message passing while watermarking methods are used for copyright protection, authentication, counterfeit protection, traitor tracing etc. [30].

Protection of intellectual property is a necessity in the multimedia industry. Digimarc Corporation is one of the leading companies that applies digital watermarking technology to various media formats. Most of their innovations are modeled on the aspects of human cognition and sensory perception. Anti-piracy is a major concern for digital cinema and robust watermarking strategies developed at Digimarc Corporation provide a way to trace the source of such piracy [31].

Three important characteristics which form a golden triangle for watermarking strategies are their imperceptibility, robustness and the payload. Spatial domain watermarking strategies are able to achieve good imperceptibility in a simple manner but fail against simple image manipulations [15]. In [15], Cox et al. propose spread spectrum watermark embedding. They suggest embedding the watermark in the perceptually significant regions of the image by choosing the low frequencies of the image in the transform domain. The watermark used is a random Gaussian sequence. The watermarking strategy "detects" the presence of the watermark after undergoing attacks. In [22], Hsu et al. introduce embedding visually recognizable patterns as watermarks and also aim to "extract" these patterns and not just detect them. In order to embed and extract watermarks which are images, their study proves that embedding in the mid-frequencies of the image achieves a good balance of the golden triangle of watermarking strategies.

The impact of deep learning methods in computer vision in the last two decades has posed new challenges and solutions equally. Kandi et al. explore the possibility of CNNs for robust image watermarking [26]. Shumeet Baluja proposed a method for hiding colour images in colour images of same size using autoencoders ensuring minimum quality loss of both images [4]. Hayes et al. use the powerful architecture of Generative Adversarial Networks(GANs) for generating steganographic images [20]. Based on these works, we have explored the idea of using convolutional autoencoders (CAE) to act as a new form of attack on popular digital watermarking schemes and found at least one strategy that resists such an attack.

In this paper:

1. A novel attack based on autoencoders is proposed. This attack exploits the representation learning ability of autoencoders.
2. An enhanced hybrid DWT+SVD watermarking strategy which is robust against all known image processing type attacks and the autoencoder attack is proposed.
3. Under the scenario that the adversary is unaware of the watermarking strategy of the owner, cross-validation performance of the CAE is done by providing an input of images watermarked using a technique other than the technique on which the network was trained.
4. The proposed DWT+SVD E (Enhanced) scheme is investigated whether or not the choice of a sub-band has a major role to play in its robustness against various attacks.

In this paper, only grayscale watermarking strategies are considered in which both the watermark and the original image are in grayscale format. The robustness of the chosen strategy is evaluated based on the quality of extracted watermark using measures such as PSNR, normalized correlation coefficient (NCC) and structural similarity index measure(SSIM).

Section 2 analyses the performance of transform domain techniques that use DCT, DWT and SVD as a stand alone or their hybrid combinations as seen in the literature under various image processing attacks. The advantages of hybrid combinations over the stand alone approach are clearly brought out for consideration.

Having established the above, a novel adversarial attack that uses autoencoders is proposed in Section 3. The novel attacking strategies are based on building autoencoders that use a simple feed-forward neural network and autoencoders that exploit the power of deep learning convolutional neural network architecture (CAE).

Section 4 demonstrates the failure of the strategies covered in Section 2 against the novel autoencoder based adversarial attacks specifically. A hybrid watermarking technique based on the enhanced version of a hybrid DWT-SVD technique is shown to outperform other techniques in terms of the quality of extracted watermarks under different scenarios viz. i) adversary having the knowledge of watermark embedding technique by the owner and ii) adversary having no clue about the watermarking strategy adopted by the owner. In conclusion section, the scope for further study into deep learning based digital watermarking strategies is elucidated.

2 Related work

JPEG and JPEG2000 are famous compression standards for images which are based on Cosine Transform and Wavelet transform of the image respectively. Frequency domain representation of an image provides a lot of flexibility to watermarking developers. For example, contourlet transform was first applied for digital watermarking by Jayalakshmi et al. in [24]. In their study they use directional sub-bands of a contourlet transform for watermarking of maps images. Subsequently contourlet domain methods have been studied in [14, 34, 50] and [43] for different kinds of images. There are a few techniques that use the Karuhen-Louve Transform (KLT) for watermarking [6, 9], and [27] to increase the robustness of watermarks by embedding them in a secret hidden space which is dependent on the basis chosen for the KLT. Using 1D-DFT along the temporal dimension for video watermarking enhances the robustness against video compression and Radon transform is applied on the frames with highest temporal frequencies to embed the watermark pattern providing robustness against geometric transformations as explained in [29]. In this manner, there are numerous transform based methods in literature which take advantage of the powers of the chosen transform. The study undertaken in this paper has considered three most popular transforms used, i.e, DCT, DWT and SVD and their dual hybrid combinations.

In this section we show how these transforms perform individually and in a hybrid fashion against image processing attacks like JPEG compression, Gaussian blurring, salt & pepper noise and geometrical distortions like cropping and rotation. For all the methods, the original image used is a grayscale Lena image of size 512×512 and the watermark image is also a grayscale logo image of size 128×128 as shown in Fig. 1. A very broad framework of single transform based watermarking strategy is shown in Fig. 2. Most of the existing techniques in the literature combine the frequency coefficients of the watermark with the frequency coefficients of the original image using a scaling factor α which

(a) Original Image (b) Watermark Image

Fig. 1 Original image and watermark image

lies in the range (0, 1). The scaling factor is chosen to achieve good imperceptibility and robustness simultaneously.

$$I_w = I_O + \alpha \times W \tag{1}$$

where I_O are the selected frequency coefficients of the original image, W are the frequency coefficients of the watermark image and I_w are the modified frequency coefficients of the watermarked original image.

2.1 Stand alone single transform based methods

Most of the watermarking strategies initially were based on using a single transformation. Even the literature for such methods is in the late 90's and early 2000's. This section provides an overview of the popular watermarking strategies that use DCT, DWT and SVD as stand alone and briefly describes the strategy used for comparison and evaluation purposes.

2.1.1 Discrete cosine transform (DCT)

Earliest papers in frequency domain watermarking are based on DCT transform [5, 15], and [38]. In these studies the watermark being used was either a Gaussian sequence or a bit

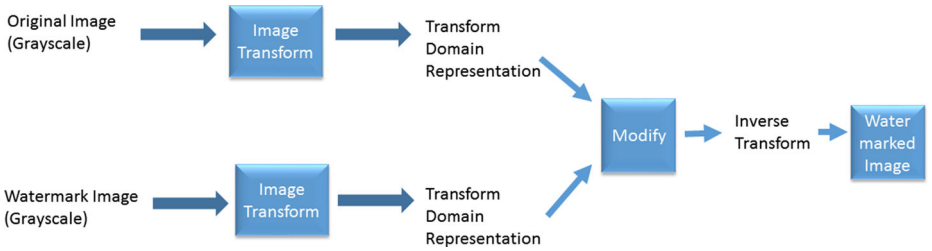
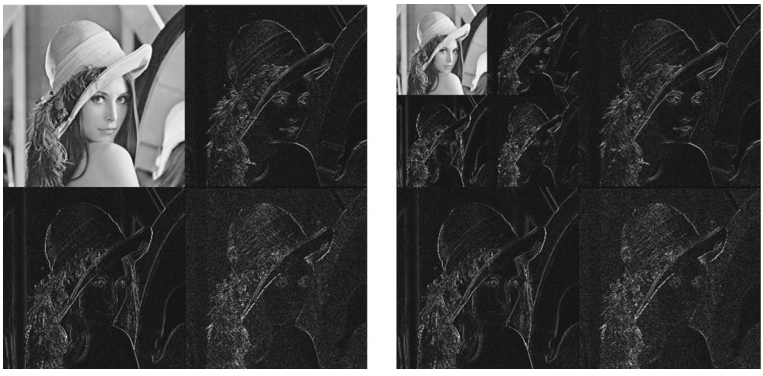


Fig. 2 Single transform watermarking strategy

stream. It was in the year 1999 that first study of embedding watermark images and extracting them to prove copyright and ownership of the original began [22]. Tao and Dickinson developed a regional classifier to identify regions with higher masking ability to embed a watermark. Block based DCT is then applied to such regions to insert the watermark. [45]. Huang and Shi, [23] embed a grayscale watermark image of size 64×64 into a grayscale original image of size 512×512. The bit rate of the watermark image is reduced by applying DCT globally and then quantizing the values. The compressed data is encoded using BCH error correcting code. Niu et al. [37] apply block based DCT to the original image, slice the grayscale watermark image into bit planes, these planes are then embedded into the DCT blocks. In order to make the watermarked image to be robust against image processing operations and malicious attacks, the mid frequencies were chosen for embedding the grayscale watermark image [46]. In our study, we transform the original image and the watermark image using DCT and then modify the mid-frequencies of the original image according to (1).

2.1.2 Discrete wavelet transform (DWT)

Xia et al. [49] proposed one of the first watermarking strategies which makes use of the multi-resolution nature of the DWT. Figure 3 shows how a 2D wavelet decomposes an image. An image is decomposed into 4 parts by sampling horizontal and vertical channels using sub-band filters. In Fig. 3a left-top quadrant contains the low frequencies of the image, also referred to as an approximate image. Other three quadrants denote the vertical, diagonal and horizontal details of the image (anti-clockwise). The second level of wavelet transform is applied on the approximate image, shown in Fig. 3b. The ability of time-frequency localization gives an edge to the wavelet transform over Fourier transform or cosine transform in image processing. Brannock et al. in [10] study eight family of wavelets consisting of bi-orthogonal and orthogonal wavelets for watermarking. They found that simpler wavelets like the Haar wavelet outperforms other complicated wavelets. High resolution detail bands, HL, LH are used for embedding the watermark. Similar technique is also used in [36], and [41]. For evaluation and comparison purposes, this technique is adopted in our study.



(a) Level 1 DWT (b) Level 2 DWT

Fig. 3 Levels of DWT transform on an image

2.1.3 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a very powerful tool used in linear algebra studies. In the last decade its usefulness in the field of signal processing in general and image processing in particular has been immense. Given a $n * n$ matrix A, the SVD of the matrix A will be

$$A = U * S * V^T \tag{2}$$

where U, V are orthogonal matrices and S is a diagonal matrix. The diagonal elements of S are called singular values and are arranged such that

$$S(1, 1) > S(2, 2) > ... > S(n, n) \tag{3}$$

SVD has been used extensively in digital watermarking due to the following reasons [11]

1. The first few singular values capture most of the energy information of the image, thereby also making it a good tool for compression.
2. The singular values of an image do not change drastically even if the image is manipulated with noise addition or compression etc.
3. SVD does not depend on the shape of the image.

SVD is a one-way decomposition algorithm and is an optimal matrix decomposition in a least squares sense. This helps it in resolving the issue of rightful ownership and robustness against various attacks. SVD transform based watermarking techniques can be very broadly categorized as:

1. Embed the watermark image values into the singular values of the orignal image using a scaling factor α [13, 32].
2. Apply SVD to the watermark image as well as the original image, modify the singular values of the original image by adding the singular values of the watermark image [12, 18, 40].
3. Modify the singular vectors of the original image with the watermark information [28].

In our study for comparison purposes, SVD based watermarking schemes of type 2 as mentioned above are considered. These SVD based schemes are most popular due to their robustness against image processing attacks.

2.2 Hybrid transform based methods

Hybrid transform methods combine the strengths of the individual transforms to improve on all three aspects of imperceptibility, robustness and capacity. Nandi et al. [35] use Particle Swarm Optimization (PSO) to find the scaling and embedding factor along with using DWT and SVD for watermarking. Arnold transform is used for scrambling the watermark image in order to improve its security along with existing strategies of DCT, DWT and SVD [1]. It is interesting to note that with the increase in watermark size, the quality of the extracted watermark which is scrambled using Arnold transform is degraded. Zhang et al. [51] process the original image using DCT, then 3-level DWT is applied on this image. SVD is then applied to the LL3 and HH3 sub-bands. The watermark image is scrambled using Arnold transform and 2-level DWT is applied to this image. SVD is applied to LL2 band of the resultant image. These values are embedded in the LL3 and HH3 bands of the original image, using a scaling matrix whose values are determined using PSO technique.

In this study, for evaluation purpose, only one way of combining DWT+DCT, DCT+SVD and DWT+SVD, amidst many possibilities is considered. Also, those hybrid techniques

having more than two transforms are not considered in this paper since the number of such combinations of transforms is unlimited and therefore does not fall in the scope of our paper.

This paper proposes a simple and effective modification to the hybridization scheme of DWT+SVD and compares its performance with the other combinations chosen as above. The advantages of the proposed modification over existing attacks such as JPEG compression, salt & pepper noise, Gaussian blurring, rotation and cropping and also the adversarial attack are established.

2.2.1 DCT+DWT

DCT and DWT are the most popular transforms used in frequency domain watermarking techniques. The energy compaction property of DCT is combined with excellent time-frequency localization property of DWT in hybrid techniques that involve DCT and DWT. The hybrid scheme provides robustness against the various attacks, when compared to the transforms used individually as can be seen in the PSNR, SSIM and NCC values in Table 1.

In most of the hybrid schemes that involve DWT and DCT, DWT is first applied to the original image and then DCT is applied on a chosen sub-band, [2, 3] and [16]. Watermark is then added to these modified DCT coefficients. Abdulrahman et al. provide a novel algorithm of combining DCT and DWT [1]. Unlike other methods where the original image is transformed first using DWT, here DCT is applied first to the original image and then DWT is applied on to the DCT transformed coefficients of the image. In our study, we have chosen this combination of DCT and DWT for evaluation and comparison purposes.

2.2.2 DCT+SVD

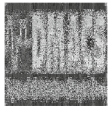
SVD transform is most often combined with other transforms to enhance the robustness of the technique. The energy compaction of DCT is combined with the robustness of the singular values. In most of the DCT+SVD based watermarking schemes, the original image is modified using DCT and then SVD is applied on the DCT coefficients either using the DC coefficients of the blocks in the image [42], and [33] or the whole image [44]. Hung-Vo et al. [48] developed an algorithm using DCT+SVD for copyright protection of stereo images.

In our study, the original image is transformed by applying DCT to the entire image and then the mid-frequencies are chosen. On these chosen block of frequencies, SVD is applied. These singular values are modified using the singular values of the watermark grayscale image according to (1).

2.2.3 DWT+SVD

Amongst the combination of hybrid transforms the most popular combination is of DWT and SVD. It combines the multi-resolution and time-frequency localization of DWT with the stability of the singular values using SVD. Ganic et al. were the first to use this combination of DWT+SVD [17]. The original image is first modified using DWT, then SVD is applied to each sub-band image. The watermark image is decomposed using SVD. The singular values of the watermark are then scaled and added to the singular values of the four sub-bands of the original image according to (1). Lai and Tsai choose the HL and LH sub-bands and modify the singular values of the sub-band adding the watermark values directly [28]. Bhatnagar et al. developed a reference image watermarking scheme using DWT+SVD [7]. One of the n-level sub-band is chosen to create a reference image. SVD is then applied

Table 1 Extracted watermark

Method	Cropping	Rotation	Blurring	S/P Noise	JPEG
DCT					
SVD					
DWT					
DCT+SVD					
DCT+DWT					
DWT+SVD					
DWT+SVD(E)					

on the reference image and watermark image respectively. The singular values are modified following (1). When SVD is applied even to the high frequency sub-band good robustness is achieved in general and especially to geometric attacks [25, 39]. Thakkar and Srivastava developed a blind and robust DWT+SVD based watermarking technique for medical images used in telemedicine [47]. To the knowledge of the authors, in all the popular hybrid techniques of DWT+SVD there is no algorithm which applies DWT and SVD on the grayscale watermark image also during embedding. This is a slight modification that is used in this study.

2.2.4 DWT+SVD enhanced

Embedding algorithm

1. 2-level DWT is applied on the original image I_O. The obtained sub-bands are denoted as $LL_2, HL_2, LH_2, HH_2, HL_1, LH_1, HH_1$
2. Choose either HL_2 or LH_2 and apply SVD on the chosen sub-band. Let's take HL_2 for consideration.
$$HL_2 = U \cdot S \cdot V' \tag{4}$$
3. 1-level DWT is applied on the watermark image W. The obtained sub-bands are denoted as $wLL_1, wHL_1, wLH_1, wHH_1$
4. The sub-band chosen in the watermark image to apply SVD is same as the sub-band chosen in the original image, which is the HL band, hence the chosen band is wHL_1. Thus
$$wHL_1 = U_w \cdot S_w \cdot V'_w \tag{5}$$
5. Singular values of the original image are now modified using a scaling factor α
$$S^{new} = S + \alpha \cdot S_w \tag{6}$$
6. The modified sub-band is obtained
$$HL_2^{new} = U \cdot S^{new} \cdot V' \tag{7}$$
7. IDWT is then applied on the the quadrants to get the watermarked image I_w
$$I_w = IDWT(LL_2, HL_2^{new}, LH_2, HH_2, HL_1, LH_1, HH_1) \tag{8}$$

Extraction algorithm

1. 2-level DWT is applied on the original image I_O. The obtained sub-bands are denoted as $LL_2, HL_2, LH_2, HH_2, HL_1, LH_1, HH_1$
2. Choose HL_2 since it was modified while embedding the watermark and apply SVD on the sub-band.
$$HL_2 = U \cdot S \cdot V' \tag{9}$$
3. 2-level DWT is applied on the watermarked image I_w. The obtained sub-bands are denoted as $cLL_2, cHL_2, cLH_2, cHH_2, cHL_1, cLH_1, cHH_1$
4. Choose cHL_2 and apply SVD on the chosen sub-band.
$$cHL_2 = U_c \cdot S_c \cdot V'_c \tag{10}$$
5. Find the difference between the singular values matrices
$$S_{mod} = \frac{S_c - S}{\alpha} \tag{11}$$

6. 1-level DWT is applied on the watermark image W. The obtained subbands are denoted as $wLL_1, wHL_1, wLH_1, wHH_1$
7. The chosen band wHL_1 is decomposed using SVD

$$wHL_1 = U_w \cdot S_w \cdot V'_w \tag{12}$$

8. The modified sub-band is

$$wHL_1^{new} = U_w \cdot S_{mod} \cdot V'_w \tag{13}$$

9. The watermark is obtained by applying IDWT as follows

$$W^* = IDWT(wLL_1, wHL_1^{new}, wLH_1, wHH_1) \tag{14}$$

2.3 Testing WM methods against image processing attacks

Having described various techniques that are considered for evaluation in this Section as above, we proceed to evaluate these watermarking techniques under various image processing attacks. Figure 4 depicts the various attacks performed on the watermarked image. Table 1 shows visually the quality of watermarks extracted after the attacks. Table 2 shows the quality of the extracted watermarks measured in terms of PSNR, SSIM, and NCC values. In each of the measure, higher the values mean, higher the degree of robustness of the watermarking embedding techniques against the chosen attacks. It is clear that the proposed DWT+SVD E (Enhanced) technique outperforms all others in terms of the quality of

(a) Gaussian Blur,(3, 3) filter (b) S&P noise, variance = 0.05

(c) 40 % JPEG compression (d) 50 % crop (e) 30° rotate

Fig. 4 Attacks performed on the watermarked image

Table 2 PSNR, SSIM and NCC values for extracted watermark

Method	Value	Cropping	Rotation	Blurring	S/P Noise	JPEG
DCT	PSNR	27.84	28.02	27.86	28.25	35.88
	SSIM	0.0010	0.3548	0.1861	0.1861	0.9605
	NCC	0.6883	0.8708	0.7947	0.7947	0.9976
SVD	PSNR	28.25	27.87	28.74	29.37	32.93
	SSIM	0.4110	0.0030	0.7138	0.7138	0.9461
	NCC	0.8107	0.7312	0.8671	0.8671	0.9596
DWT	PSNR	28.16	29.42	29.97	29.97	30.39
	SSIM	0.1888	0.6119	0.7138	0.7138	0.9324
	NCC	0.8327	0.8859	0.8308	0.8308	0.9427
DCT+SVD	PSNR	28.84	28.20	28.89	28.12	34.39
	SSIM	0.2660	0.6420	0.6972	0.6972	0.9537
	NCC	0.8829	0.9665	0.8632	0.8632	0.9941
DCT+DWT	PSNR	28.27	28.38	28.74	28.97	36.97
	SSIM	0.2127	0.3368	0.4151	0.4151	0.9651
	NCC	0.8665	0.9026	0.9026	0.9026	0.9927
DWT+SVD	PSNR	32.28	30.39	31.09	34.17	36.54
	SSIM	0.7344	0.6065	0.6965	0.6965	0.9584
	NCC	0.8405	0.9195	0.9355	0.9355	0.9985
DWT+SVD E	PSNR	**34.06**	**30.97**	**31.72**	**35.24**	**37.06**
	SSIM	**0.8175**	**0.6665**	**0.7404**	**0.7404**	**0.9626**
	NCC	**0.9889**	**0.9225**	**0.9873**	**0.9873**	**0.9989**

extracted watermarks. An analysis into which subband is robust for watermarking using the method of DWT+SVD E is provided next.

2.3.1 Sub-band selection for DWT+SVD E against conventional attacks

The sub-band selected for embedding the watermark is found to make a difference in the quality of the watermark extracted in the enhanced DWT+SVD algorithm. SSIM and NCC values are used to evaluate the quality of extraction as seen in Table 3. It is observed that the quality of the watermark extracted is not good when the LL band is used for embedding, especially against geometric attacks like rotation and cropping. Against the image processing attacks like Gaussian blurring and JPEG compression all the bands give almost equal performance. Against salt & pepper noise (S & P noise) the LL band has a slight advantage. However the PSNR of the watermarked image when the LL band is used is 37.65 DB when compared to the HL and LH band giving 49.85 DB and 47.17 DB and the HH band gives 57.24 DB. Keeping this in mind it is recommended not to choose the LL sub band as it affects the imperceptibility of the original image.

A comparison on popular and simple watermarking strategies against conventional geometric and image processing attacks was covered in this section. We will now propose in the next Section a novel adversarial attack based on autoencoders and investigate how the above watermarking strategies survive against this novel attack.

Table 3 Sub-band analysis for DWT+SVD E

Band	Value	Cropping	Rotation	Blurring	S/P Noise	JPEG
LL	SSIM	0.1747	0.0260	0.7134	0.9195	0.9321
	NCC	0.8081	0.7450	0.9893	0.9991	0.9976
HL	SSIM	0.8499	0.5490	0.7712	0.7738	0.9040
	NCC	0.9830	0.9470	0.9732	0.9737	0.9879
LH	SSIM	0.6560	0.6430	0.5620	0.7992	0.9521
	NCC	0.9668	0.9640	0.9556	0.7992	0.9954
HH	SSIM	0.8175	0.6330	0.6665	0.6528	0.9405
	NCC	0.9778	0.9580	0.9585	0.9597	0.9919

3 Autoencoder based adversarial attack

Autoencoder is an artificial neural network used for dimensionality reduction of data. There are two parts to an autoencoder i.e. the encoder and the decoder. The encoder reduces the input to a state space with fewer dimensions and the decoder reconstructs the input from that representation. It learns to ignore the latent noise in the higher dimensions of the data. autoencoders are often trained with only a single layer encoder and a single layer decoder, but usage of deep encoders and decoders offers many advantages. In our study, both architectures have been used to evaluate the aforementioned watermarking algorithms.

3.1 Autoencoder I

A simple autoencoder consisting of a single layer in the encoder and decoder. Architecture of this autoencoder is shown in Fig. 5.

The above simple autoencoder is trained on a single watermarked image of the owner as the adversary has access to this image only. The assumption is that the adversary knows the watermarking technique used by the owner. The trained autoencoder can not be used for the reconstruction of an unseen image of another owner's watermarked image as the resultant output trained on a different image input may not produce an image that will be perceptually

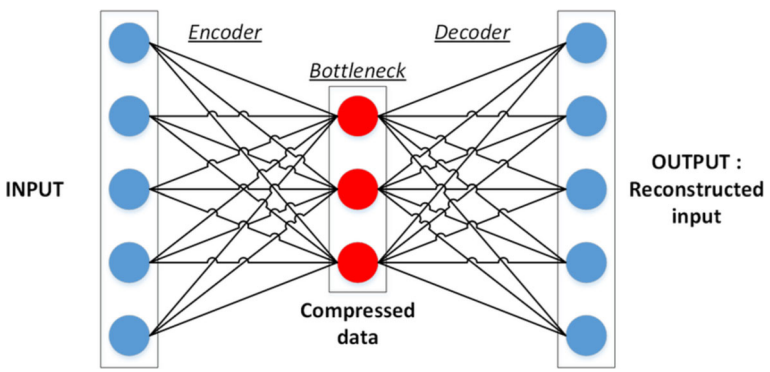


Fig. 5 Single layer autoencoder; Credit: ResearchGate

closer to the new input. This drawback can be overcome by training the network on a larger image data set like CIFAR 10.

3.2 Autoencoder II

In this subsection, a deep convolutional neural network based autoencoder (CAE) is proposed for the purpose of training on a large set of images in order that the trained CAE network has the ability to produce perceptually closer images at the output layer even for images that are unseen during training and testing phases.

In a CAE the convolution operation encodes the input in a form of simple signals and then tries to reconstruct the input from it [19]. Deep autoencoders are important because:

1. Depth can exponentially reduce the computational cost of representing some functions.
2. Depth can exponentially decrease the amount of training data needed to learn some functions.
3. Experimentally, deep autoencoders yield better compression compared to shallow or linear autoencoders [21]

The CAE architecture is shown in Fig. 6. The chosen network consists of 8 convolution layers.

4 Experimental setup

The adversarial attack based on an autoencoder has not been studied in digital watermarking methods until now. Here an in-depth study is presented comparing watermarking strategies

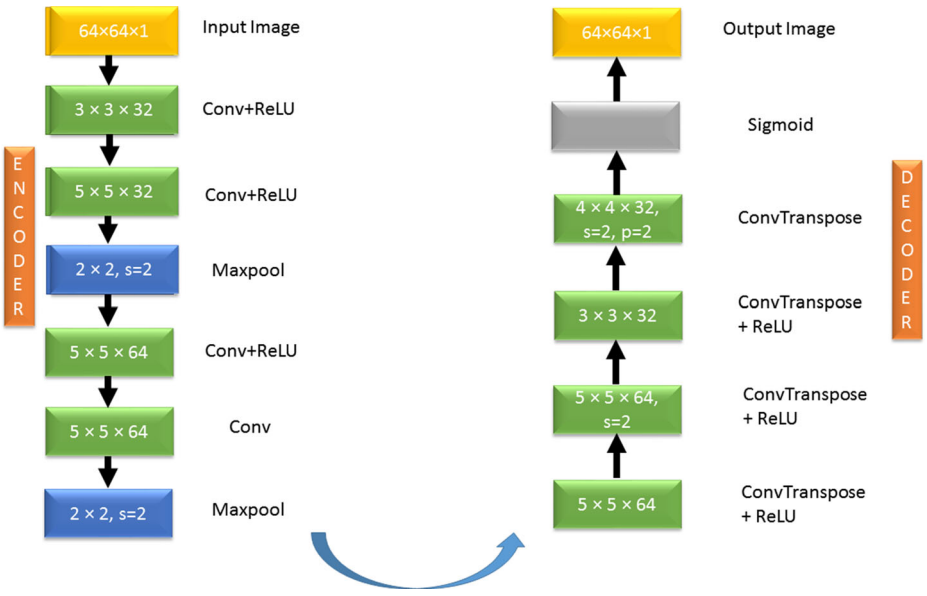


Fig. 6 CNN autoencoder architecture

against such an attack. DCT, DWT and SVD techniques dominate the literature of transform domain based methods for watermarking. There are numerous ways in which these techniques can be used individually as well as in a hybrid fashion. In this study, seven methods have been chosen for evaluation purposes as consideration of all possible watermarking methods is beyond the scope. In all these techniques the most popular way of embedding a watermark is described in (1).

4.1 Results for autoencoder I

Table 4 depicts the results for training Autoencoder I with watermarked Lena image. Input to the autoencoder is the grayscale watermarked Lena image of size 512×512 with watermark of size 128×128 Fig. 1b. In Table 4 the columns depict the PSNR of the reconstructed image (RI), the extracted watermark from the reconstructed image, the PSNR, the SSIM and the NCC of the extracted watermark respectively. Enhanced DWT+SVD method extracts the watermark image from the reconstructed image most efficiently. It is also observed that the watermarking strategies that contain the wavelet transform perform better than others without it.

4.2 Results for autoencoder II

The CIFAR 10 data set is used for training and testing the CAE architecture explained in Section 3.2. CIFAR 10 consists of 60000 images of size 32×32 in 10 classes. The data set is divided into 50,000 training images and 10,000 for testing. These colour images of the CIFAR 10 data set are converted to grayscale images of size 64×64. The data set is watermarked using one of the above mentioned strategies and these watermarked images are presented to the network. For every watermarking technique, the network is trained and tested on the above data set. The output of the CAE is a data set consisting of images which are visually same as the watermarked images. Mean square error is used as the loss function to be minimized while training the network. The objective of the study is to understand the power of CAE-based adversarial attack by extracting the embedded watermark from the reconstructed images of the watermarked data set and evaluating the quality. By doing so, we will understand the robustness of different watermarking strategies against this novel attack.

The results for the watermark extracted are shown in Fig. 7. Five images of different classes from the CIFAR 10 data set have been chosen for explanation. The results are similar over the entire data set.

First row contains sample images from the data set, the left column contains the watermarking technique chosen and the corresponding row depicts the watermark extracted from the reconstructed images. Individual transform methods based on DCT and SVD do not perform well, but the performance is better when they are combined together as DCT+SVD. Wavelet based methods, i.e, DWT, DCT+DWT, DWT+SVD and DWT+SVD E are seen performing better at the quality of the watermark being extracted.

The imperceptibility of the watermarking technique is evaluated by the average PSNR value over the entire data set. This is shown in Table 5 where the PSNR values for three different comparisons for all the seven strategies on the CIFAR10 data set.

1. Original Image and Watermarked Image (PSNR I)
2. Watermarked Image and Reconstructed Image (PSNR II)
3. Original Image and Reconstructed Image (PSNR III)

Table 4 Results for Autoencoder I

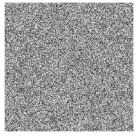
Method	PSNR (RI)	Extracted Watermark	PSNR	SSIM	NCC
DCT	30.23		27.13	0.0380	0.7334
SVD	35.05		28.02	0.2019	0.8571
DWT	36.57		30.44	0.5590	0.9298
DCT+SVD	35.16		27.81	0.2030	0.9122
DCT+DWT	33.43		29.62	0.4822	0.9385

Table 4 (continued)

Method	PSNR (RI)	Extracted Watermark	PSNR	SSIM	NCC
DWT+SVD	34.54		27.85	0.4932	0.9444
DWT+SVD(E)	36.36		**31.85**	**0.5444**	**0.9760**

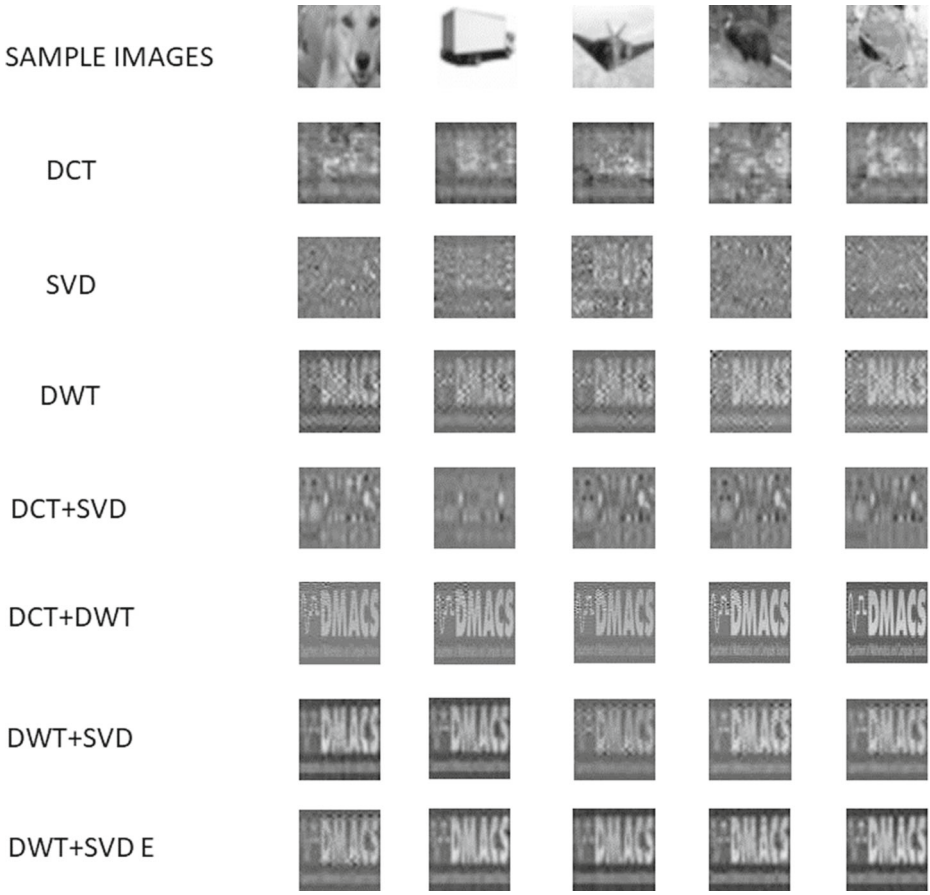


Fig. 7 Sample Results for CIFAR10 data set

4.2.1 Sub-band analysis for DWT+SVD E against CAE attack

In Section 2.3.1 it was shown that using the LL band in the DWT+SVD E strategy is not advised. A similar analysis is provided here too for the CAE attack. The same subband is chosen on the original image and the watermark for transforming and embedding. The

Table 5 PSNR values comparison

Method	PSNR I	PSNR II	PSNR III
DCT	40.52	36.50	30.55
SVD	38.12	35.08	31.67
DWT	50.05	35.01	34.99
DCT+SVD	40.59	35.77	29.71
DCT+DWT	47.35	36.28	35.9
DWT+SVD	40.81	35.58	34.17
DWT+SVD E	51.16	35.79	34.65

Table 6 Subband analysis for DWT+SVD E against CAE attack

Band	PSNR of RI	SSIM of EW	NCC of EW
LL	30.36	0.9418	0.9940
HL	35.58	0.9472	0.9944
LH	35.84	0.9466	0.9950
HH	37.58	0.9683	0.9961

Table 7 PSNR values for cross-validation testing

Trained on	Watermarking technique						
	DCT	SVD	DWT	DCT+DWT	DCT+SVD	DWT+SVD	DWT+SVD E
DCT	15.17	18.05	17.07	18.87	7.57	15.93	**25.05**
SVD	7.56	6.47	19.87	16.89	5.88	6.55	**24.32**
DWT	7.41	22.44	20.57	19.37	5.92	24.82	**25.62**
DCT+DWT	7.58	20.85	20.05	19.48	5.71	18.12	**25.64**
DCT+SVD	6.99	16.98	20.98	21.38	18.38	17.36	**26.54**
DWT+SVD	8.1	4.54	20.55	16.09	6.36	22.07	**26.36**
DWT+SVD E	8.11	20.10	21.11	20.95	6.11	21.09	**26.59**

Table 8 SSIM values for cross-validation testing

Trained on	Watermarking technique						
	DCT	SVD	DWT	DCT+DWT	DCT+SVD	DWT+SVD	DWT+SVD E
DCT	0.5227	0.7378	0.7557	0.8505	0.2197	0.6417	**0.9485**
SVD	0.0972	0.0747	0.8500	0.8785	0.0392	0.0292	**0.9379**
DWT	0.0745	0.9061	0.8795	0.8601	0.0354	0.9322	**0.9930**
DCT+DWT	0.0814	0.8785	0.8477	0.8587	0.0194	0.7829	**0.9569**
DCT+SVD	0.0617	0.8627	0.8966	0.8966	0.8474	0.8166	**0.9647**
DWT+SVD	0.1046	0.0408	0.8704	0.8824	0.0782	0.8846	**0.9638**
DWT+SVD E	0.1224	0.8356	0.8717	0.8837	0.0447	0.8756	**0.9504**

Table 9 NCC values for cross-validation testing

Trained on	Watermarking technique						
	DCT	SVD	DWT	DCT+DWT	DCT+SVD	DWT+SVD	DWT+SVD E
DCT	0.9494	0.9771	0.9671	0.9779	0.8702	0.9503	**0.9936**
SVD	0.7374	0.5827	0.9809	0.9676	0.3306	0.6186	**0.9930**
DWT	0.7261	0.9897	0.9847	0.9801	0.8955	0.9929	**0.9949**
DCT+DWT	0.7653	0.9874	0.9791	0.9502	0.6205	0.9786	**0.9945**
DCT+SVD	0.7216	0.9877	0.9858	0.9871	0.9733	0.9823	**0.9961**
DWT+SVD	0.8372	0.5814	0.9844	0.9621	0.8213	0.9869	**0.9958**
DWT+SVD E	0.8317	0.9819	**0**.9832	0.9861	0.8317	0.9859	**0.9939**

results of the average PSNR value of the reconstructed image (RI), along with the average SSIM and the average NCC values for the extracted watermark (EW) from these bands over the entire dataset is given in Table 6. The PSNR of the reconstructed image is low when the LL band is used for embedding, it is thus established that the LL band must not be chosen in the proposed methodology. Amongst the other sub-bands almost similar performance is found, thus we conclude that any of the mid-frequency sub-bands can be chosen.

4.3 Results for cross-validation testing

In Sections 4.1 and 4.2, the autoencoder was trained for reconstructing image(s) of the data set watermarked using a particular technique and the same technique was used to extract the watermark. For example, if the watermarked single image given as input to Autoencoder I or the watermarked CIFAR10 data set given as input to Autoencoder II used DCT technique, the watermark was extracted from the reconstructed image using DCT itself.

Consider the situation in which the autoencoder is trained for DCT method of watermarking and then its weights are frozen. To this network, if images watermarked using DWT are passed and the images are reconstructed, what will be the quality of the watermark extracted? Will it reflect the DCT quality of extraction or the DWT quality? This analysis is shown in three Tables 7, 8, and 9 that reflect the PSNR, SSIM and NCC values of the extracted watermark from the reconstructed images obtained over networks trained on image inputs whose watermarking methods are being different.

From Tables 7, 8 and 9 a few observations can be made as below:

1. DCT based methods, except DCT+DWT, fail to extract the watermark efficiently from the reconstructed image when tested against networks trained on other methods.
2. Wavelet transform based methods perform generally well against all strategies
3. DWT+SVD E performs best in all the scenarios considered.

5 Conclusion

Robust digital watermarking techniques are used for copyright protection, when an adversary attempts to destroy the watermark. High Quality of the extracted watermark gives a strong case in favour of the watermarking strategy to be used by the owner. An adversarial attack that has not been studied for digital watermarking techniques is proposed in this paper. A convolutional autoencoder has been used to reconstruct perceptually similar images when given an input of watermarked image using techniques like DCT, SVD and DWT and also a few of their hybrid combinations. It is observed that the wavelet based methods perform well against the CNN based autoencoder enabled adversarial attack. An enhanced hybrid combination of DWT+SVD is found to extract the watermark with best quality amongst other methods studied in this paper. Detailed analysis of sub-band selection using this method concludes that choosing the mid-frequency bands performs better against the proposed novel attack.

Acknowledgments We dedicate this paper to the founder Chancellor of Sri Sathya Institute of Higher Learning Bhagawan Sri Sathya Sai Baba under His dictum high quality value-based education is imparted without any cost over the past five decades. We take this opportunity to thank the support, innumerable discussions and guidance from our colleague V.Sai Raam. We would also like to thank Google for providing us a very useful resource of Google Colaboratory.

References

1. Abdulrahman AK, Ozturk S (2019) A novel hybrid dct and dwt based robust watermarking algorithm for color images. Multimed Tools Appl 78(12):17027–17049
2. Al-Haj A (2007) Combined dwt-dct digital image watermarking. J Comput Sci 3(9):740–746
3. Amirgholipour SK, Naghsh-Nilchi AR (2009) Robust digital image watermarking based on joint dwt-dct. Int J Digit Content Technol Appl 3(2):42–54
4. Baluja S (2020) Hiding images within images. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(7):1685–1697
5. Barni M, Bartolini F, Cappellini V, Piva A (1998) A dct-domain system for robust image watermarking. Signal Process 66(3):357–372
6. Barni M, Bartolini F, De Rosa A, Piva A (2002) Color image watermarking in the karhunen-loeve transform domain. J Electron Imaging 11(1):87–95
7. Bhatnagar G, Raman B (2009) A new robust reference watermarking scheme based on dwt-svd. Comput Stand Interfaces 31(5):1002–1013
8. Böhme R, Kirchner M, Katzenbeisser S, Petitcolas F (2016) Media forensics. In: Information hiding. Artech House, pp 231–259
9. Botta M, Cavagnino D, Pomponiu V (2015) Fragile watermarking using Karhunen–Loève transform: the klt-f approach. Soft Comput 19(7):1905–1919
10. Brannock E, Weeks M, Harrison R (2008) Watermarking with wavelets: simplicity leads to robustness. In: IEEE SoutheastCon 2008. IEEE, pp 587–592
11. Cao L (2006) Singular value decomposition applied to digital image processing. Division of Computing Studies, Arizona State University Polytechnic Campus, Mesa Arizona State University polytechnic Campus, pp 1–15
12. Chandra DS (2002) Digital image watermarking using singular value decomposition. In: The 2002 45th Midwest symposium on circuits and systems, 2002. MWSCAS-2002, vol 3. IEEE, pp III–III
13. Chang CC, Tsai P, Lin CC (2005) Svd-based digital image watermarking scheme. Pattern Recognit Lett 26(10):1577–1586
14. Chen L, Zhao J (2018) Contourlet-based image and video watermarking robust to geometric attacks and compressions. Multimed Tools Appl 77(6):7187–7204
15. Cox IJ, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
16. El Bireki MFM, Abdullah M, Ukasha AAM, Elrowayati AA (2016) Digital image watermarking based on joint (dct-dwt) and arnold transform. Int J Secur Appl 10(5):107–118
17. Ganic E, Eskicioglu AM (2004) Robust dwt-svd domain image watermarking: embedding data in all frequencies. In: Proceedings of the 2004 workshop on multimedia and security, pp 166–174
18. Ganic E, Zubair N, Eskicioglu AM (2003) An optimal watermarking scheme based on singular value decomposition. In: Proceedings of the IASTED international conference on communication, network, and information security, vol 85
19. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
20. Hayes J, Danezis G (2017) Generating steganographic images via adversarial training. In: Advances in neural information processing systems, pp 1954–1963
21. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
22. Hsu CT, Wu JL (1999) Hidden digital watermarks in images. IEEE Trans Image Process 8(1):58–68
23. Huang J, Shi YQ (2001) Embedding gray level images. In: ISCAS 2001. The 2001 IEEE international symposium on circuits and systems (Cat. No. 01CH37196), vol 5. IEEE, pp 239–242
24. Jayalakshmi M, Merchant SN, Desai UB (2006) Digital watermarking in contourlet domain. In: 18th International conference on pattern recognition (ICPR'06), vol 3. IEEE, pp 861–864
25. Joshi M et al (2010) Robust image watermarking based on singular value decomposition and discrete wavelet transform. In: 2010 3rd International conference on computer science and information technology, vol 5. IEEE, pp 337–341
26. Kandi H, Mishra D, Gorthi SRS (2017) Exploring the learning capabilities of convolutional neural networks for robust image watermarking. Comput Secur 65:247–268
27. Kasban H (2017) A spiral based image watermarking scheme using karhunen–loeve and discrete hartley transforms. Multidimens Syst Signal Process 28(2):573–595
28. Lai CC, Tsai CC (2010) Digital image watermarking using discrete wavelet transform and singular value decomposition. IEEE Trans Instrum Meas 59(11):3060–3063

29. Liu Y, Zhao J (2010) A new video watermarking algorithm based on 1d dft and radon transform. Signal Process 90(2):626–639
30. Lu CS (2004) Multimedia security: steganography and digital watermarking techniques for protection of intellectual property: steganography and digital watermarking techniques for protection of intellectual property. Igi Global
31. Milano D (2012) Content control: digital watermarking and fingerprinting. White Paper, Rhozet, a business unit of Harmonic Inc., http://www.rhozet.com/whitepapers/Fingerprinting-Watermarking.pdf, Last Accessed May 30
32. Mohammad AA, Alhaj A, Shaltaf S (2008) An improved svd-based watermarking scheme for protecting rightful ownership. Signal Process 88(9):2158–2180
33. Mukherjee S, Pal AK (2012) A dct-svd based robust watermarking scheme for grayscale image. In: Proceedings of the international conference on advances in computing, communications and informatics, pp 573–578
34. Najih A, Al-Haddad S, Ramli AR, Hashim S, Nematollahi MA (2017) Digital image watermarking based on angle quantization in discrete contourlet transform. J King Saud Univ-Comput Inf Sci 29(3):288–294
35. Nandi S, Santhi V (2016) Dwt–svd-based watermarking scheme using optimization technique. In: Artificial intelligence and evolutionary computations in engineering systems. Springer, pp 69–77
36. Narang M, Vashisth S (2013) Digital watermarking using discrete wavelet transform. International Journal of Computer Applications 74(20):34–38
37. Niu XM, Lu ZM, Sun SH (2000) Digital watermarking of still images with gray-level digital watermarks. IEEE Trans Consum Electron 46(1):137–145
38. O'Ruanaidh J, Dowling W, Boland F (1996) Watermarking digital images for copyright protection. IEE Proc-Vis Image Signal Process 143(4):250–256
39. Saxena P, Garg S, Srivastava A (2012) Dwt-svd semi-blind image watermarking using high frequency band. In: 2nd International conference on computer science and information technology (ICCSIT'2012) Singapore April, pp 28–29
40. Shieh JM, Lou DC, Chang MC (2006) A semi-blind digital watermarking scheme based on singular value decomposition. Comput Stand Interfaces 28(4):428–440
41. Singh A, Tayal A (2012) Choice of wavelet from wavelet families for dwt-dct-svd image watermarking
42. Singh P, Agarwal S (2013) A hybrid dct-svd based robust watermarking scheme for copyright protection. In: 2013 Africon. IEEE, pp 1–5
43. Su Q, Wang G, Lv G, Zhang X, Deng G, Chen B (2017) A novel blind color image watermarking based on contourlet transform and hessenberg decomposition. Multimed Tools Appl 76(6):8781–8801
44. Sverdlov A, Dexter S, Eskicioglu AM (2005) Robust dct-svd domain image watermarking for copyright protection: embedding data in all frequencies. In: 2005 13th European signal processing conference. IEEE, pp 1–4
45. Tao B, Dickinson B (1997) Adaptive watermarking in the dct domain. In: 1997 IEEE International conference on acoustics, speech, and signal processing, vol 4. IEEE, pp 2985–2988
46. Tewari TK, Saxena V (2010) An improved and robust dct based digital image watermarking scheme. Int J Comput Appl 3(1):28–32
47. Thakkar FN, Srivastava VK (2017) A blind medical image watermarking: Dwt-svd based robust and secure approach for telemedicine applications. Multimed Tools Appl 76(3):3669–3697
48. Vo PH, Nguyen TS, Huynh VT, Do TN (2017) A robust hybrid watermarking scheme based on dct and svd for copyright protection of stereo images. In: 2017 4th NAFOSTED conference on information and computer science. IEEE, pp 331–335
49. Xia XG, Boncelet CG, Arce GR (1997) A multiresolution watermark for digital images. In: Proceedings of international conference on image processing, vol 1. IEEE, pp 548–551
50. Zaboli S, Moin MS (2007) Cew: A non-blind adaptive image watermarking approach based on entropy in contourlet domain. In: 2007 IEEE international symposium on industrial electronics. IEEE, pp 1687–1692
51. Zhang L, Wei D (2019) Dual dct-dwt-svd digital watermarking algorithm based on particle swarm optimization. Multimed Tools Appl 78(19):28003–28023

Affiliations

Sai Shyam Sharma[1] · **V. Chandrasekaran**[1]

V. Chandrasekaran
vchandrasekaran@sssihl.edu.in

[1] Sri Sathya Sai Institute of Higher Learning, Anantapur, Andhra Pradesh, India