Mathematical Modelling of Engineering Problems
Vol. 11, No. 9, September, 2024, pp. 2483-2496
Journal homepage: http://iieta.org/journals/mmep

A New Face Image Manipulation Localization and Recovery Algorithm Using Image Watermarking and Integer Wavelet Transform

Asmaa Hatem Jawad[1], Rasha Thabit[1,2*], Muntadher H. Al-Hadaad[1], Khamis A. Zidan[3]

[1] Computer Engineering, College of Engineering, Al-Iraqia University, Baghdad 10053, Iraq
[2] Department of Computer Techniques Engineering, Dijlah University College, Baghdad 10022, Iraq
[3] Department of Scientific Affairs, Al-Iraqia University, Baghdad 10053, Iraq

Corresponding Author Email: rasha.thabit@aliraqia.edu.iq

https://doi.org/10.18280/mmep.110920

Received: 20 June 2023
Revised: 30 November 2023
Accepted: 15 December 2023
Available online: 29 September 2024

Keywords:
DeepFakes reveal, multimedia forensics, face manipulation reveal, face image recovery

ABSTRACT

Recognizing different kinds of modifications and identifying the altered portions of the face region have been the main focus of recent developments in face image manipulation detection. In actual applications, the ability to restore the facial region after modification localization would be highly helpful, but this was not addressed in earlier studies. This research utilizes integer wavelet transform (IWT) coefficients to produce recovery information from the face region and watermarking-based technique for incorporating the generated data into the cover face image. Three distinct algorithms have been proposed for producing the recovery data, and the one demonstrating superior performance, specifically IWT (cdf 3.5), is employed within main algorithms. The novelty of the suggested technique stems from its integration of an IWT-based recovery method, along with the manipulation detection process, which has not been showcased in prior research studies. The main contributions of the suggested algorithm include its efficiency to precisely identify altered blocks within the facial area and to reinstate the unaltered version when modifications are present. The advantage of the proposed algorithm is demonstrated through the comparisons with earlier methods where it can be used in digital art to ensure the originality, integrity, and security of facial images. The practical applications include various fields such as forensic investigations, digital image authentication, online safety, content moderation, medical imaging, security systems, entertainment, privacy protection, historical documentation, The limitation of the proposed algorithm is the restricted embedding capacity. The future researches can be conducted in different directions such as enhancing the embedding capacity, implementing a real-time detection system for live video streams, and investigating the main requirements for efficient algorithm's execution on hardware devices.

1. INTRODUCTION

Rapid technological advancements that make online photo sharing easier are said to be the main reason why so many people now rely on digital information exchange. Over the years, while the technology is improved, the digital images have been easily shared through internet for many purposes [1-3]. Nowadays, it is very convenient for the technology users to upload their images and personal information on the cloud to be available whenever they need it, however, the users' main anxieties revolve around the safety of the uploaded data [4, 5].

Various methods and security mechanisms have been introduced to guarantee safe sharing of sensitive data and digital images, including encryption, steganography, and watermarking [6-8]. The face image, which is a crucial kind of data in digital multimedia, has been shared through internet for different intends such as identity discrimination in biometric systems (i.e., for security and access control) [9], celebrity and fame intentions, social news, face recognition applications, and many others [9-12].

In recent years, due to the rapid advancements in technology, digital facial manipulation methods, algorithms, and applications have seen extensive dissemination [13-17]. Facial images manipulation techniques are either intentional attacks (i.e., there are some harmful intentions behind applying the manipulation process) or unintentional attacks (i.e., the manipulation process is applied for innocent intentions such as beautifying the face, changing the lighting, adding sum funny stickers, …) which in both cases are changing the contents of the facial image thus changing its features [18]. The field of data security has experienced a growing fascination with facial image manipulation techniques and their impact on data security systems. Over the past few years, the term 'DeepFakes' has captured the curiosity of the research community, paving the way for a burgeoning area of research [19-24]. Generally, the DeepFakes referred to the fake digital data that has been created using deep learning techniques. DeepFakes have been used maliciously for financial fraud, disinformation campaigns, and the production of phony pornography. Therefore, the researchers dedicated their efforts

for developing face image manipulation detection (FIMD) algorithms that can serve the general media forensics [25-30].

To current date, the FIMD algorithms have many limitations and there is no global FIMD algorithm, therefore, the doors of this research topic are opened to find new contributions in this field. Some FIMD restrictions and difficulties have been highlighted by Salih et al. [31] such as:

- **Swift advancements in manipulation tools:** Ongoing development and refinement of applications and tools for generating counterfeit images and videos pose an enduring challenge for detection techniques. As manipulators become more sophisticated, detection methods must adapt swiftly.
- **Restricted access to datasets:** The availability of extensive, top-quality datasets necessary for training and validating detection algorithms is often constrained. Gaining permissions from dataset owners can be cumbersome and limiting.
- **Generalization concerns in deep learning:** Methods rooted in deep learning hinge on specific training datasets, and their effectiveness diminishes when faced with input images significantly different from the training data. Achieving robust generalization remains a challenge.
- **Standardization deficiency:** The absence of standardized datasets, experimental procedures, and evaluation metrics makes it challenging to consistently compare and assess different detection studies and methods.
- **Resource-intensive supervised algorithms:** Many detection techniques rely on supervised learning, demanding substantial time and effort for data labeling and model training.
- **High-quality counterfeit images:** The process of producing high-quality fake photos has progressed, making it harder for detection algorithms to discern between altered and genuine contents.
- **Complexity of deep learning-based approaches:** Most available detection techniques are founded on deep learning, which involves considerable computational complexity, constraining their applicability in resource-constrained contexts.

One of the suggested solutions to these limitations is the use of digital image watermarking [31] to introduce new FIMD techniques that can detect different manipulations without the need for training and with high detection accuracy. Inspired by this suggestion, the work [32, 33] presents a new FIMD scheme using content-based image watermarking algorithm [34-36]. The scheme [32, 33] obtained promising results in detecting different types of manipulations with 100% accuracy, however, it cannot recover the original face region. In data security and forensics systems, it will be very useful if the FIMD scheme can detect the manipulations and recover the original face when manipulations exist. Face recovery after manipulation detection is a critical aspect of data forensics, serving multiple essential purposes. It preserves the integrity of digital evidence, ensuring that manipulated images don't compromise the accuracy and trustworthiness of evidence in forensic investigations. In legal proceedings, it bolsters the admissibility of evidence by verifying its unaltered state, enhancing its legal validity. Face recovery also plays a pivotal role in identifying and attributing individuals involved in cybercrimes, fraud, or identity theft. By restoring the original face, it aids in accurately connecting individuals to criminal activities.

Moreover, face recovery helps complete the forensic puzzle by addressing cases in which manipulated images obscure vital information. It assists in piecing together the full picture, providing essential details and clues for thorough analysis. Privacy protection is another key facet; when images are manipulated for privacy invasion or harassment, face recovery safeguards individuals' privacy and dignity. Additionally, it can prevent false accusations by exonerating innocent individuals wrongly incriminated through manipulated images. The face recovery offers a clear reference point and expedites the analysis process. In digital forensics, it contributes to comprehensive evidence examination, and in matters of public safety, it aids in accurate identification, ensuring the safety and security of the public.

Drawing inspiration from concepts related to tamper detection and image recovery in medical contexts [37-40], we propose incorporating a recovery algorithm into the FIMD scheme as part of the process [32, 33]. Based on previous studies of the medical images' authentication techniques [41-45] which have been applied to grayscale medical images, we suggest three different algorithms (i.e., average of each (4×4) pixels image's block, average of each (2×2) pixels image's block, and IWT based algorithms) for generating recovery data of the color face region. To test the suggested methods for producing recovery information for the face region, initial experiments have been carried out. The objective is to choose the approach that can minimize the payload length while restoring the facial region with good visual quality. Based on the preliminary study, we found that the IWT-based algorithm obtained the best results compared to average (4×4) and average (2×2) based algorithms, therefore, a new algorithm is proposed based on IWT and image watermarking technology. The presented work's innovations and contributions can be summed up as follows:

- The capacity of the suggested technique to restore the original face region in addition to precisely identifying the altered face region, a feature that has not been highlighted in earlier studies in this field.
- The preliminary algorithms provide the details of the three suggested recovery algorithms which are useful for future researches in this field.
- The inclusion of the proposed IWT (cdf 3.5)-based algorithm in face image authentication scheme to generate and recover the color face region with high visual fidelity.

The remaining portion of the paper is divided into four sections: an overview of relevant works, an explanation of the suggested algorithms for producing recovery data from the color face region, specifics of the extraction and embedding algorithms, the outcomes of the experiments carried out, and research conclusions.

2. RELATED WORKS

In recent years, there has been a proliferation of deep learning-based techniques for FIMD. The subsequent paragraphs briefly introduce some of these techniques and briefly illustrate the pros and cons of each technique, which served as the driving force for conducting this research.

In the study [46], a hybrid CNN-LSTM model was developed to distinguish between altered and unaltered regions within images. Notably, this approach focuses on the distinctive features found in the borders shared between manipulated and nearby unmanipulated pixels, emphasizing spatial structures as a key characteristic. A forgery detection method based on supervised learning was presented in another paper [47]. This technique made use of the Fruit Fly Optimization Algorithm (FOA) for optimization and a Support Vector Neural Network (SVNN) for supervised learning. The input data for the classifier was created by using texture operators, wavelet transforms, and Gabor filters to extract features from facial photographs. The fruit fly optimization method was then used to train the classifier to identify manipulation. It is worth noting that conventional methods mentioned by Bappy et al. [46] and Cristin et al. [47] are heavily relied on manually crafted features, which proved inefficient and time-consuming, as they necessitated extensive testing to select relevant features and classification algorithms.

To enhance accuracy and reduce complexity, an alternative FIMD technique has been introduced relying on a specialized Convolutional Neural Network (CNN) [48]. This approach avoids concentrating solely on specific manipulation attributes to achieve dependable detection results. The method employs a network model which incorporates multiple convolutional layers for effective feature extraction at various levels of abstraction within manipulated regions. Furthermore, to tackle the challenge of an unbalanced dataset, adaptive boosting (AdaBoost) and extreme gradient boosting were employed.

Other FIMD techniques have explored the utilization of capsule networks for face image forensics applications [49-52]. In the study [49], a capsule network was employed, offering performance comparable to traditional convolutional neural networks (CNNs) but with fewer parameters. This was done to mitigate domain-specific limitations and inefficiencies. The study advanced and improved the knowledge of this strategy within the field of picture forensics by introducing the use of capsule networks in forensics. Nguyen et al. [50] elaborated on how CNN-based detectors have historically sought performance improvements by increasing depth, width, internal connections, or merging characteristics and predicted probabilities from different CNNs. Consequently, CNN-based detectors grew in size, necessitated more training data, and demanded increased memory and processing resources. Their ability to generalize across various manipulation techniques was also a concern.

To address these challenges, a capsule network tailored for forensics, known as the "Capsule-Forensics" network, was proposed. To improve performance, the method used a dynamic routing and a pretrained feature extractor with pooling process. Another FIMD strategy employing capsule networks was used in the study by Khalil [51] with the goal of overcoming low-generalization concerns and thwarting face swap manipulations. Preprocessing was used in this method to reduce noise in the data and enhance input quality, which may have improved the detection model's accuracy and dependability. However, this enhancement came at the expense of increased processing time and complexity. Finally, Cao et al. [52] presented that shortcoming in the capsule network-based method from the study [49] were highlighted, particularly the use of a single network with only three layers in the feature extraction module, leading to inconsistencies between features and the human visual process. The supervised network is another FIMD technique that Cao et al. [52] designed to address this problem. These capsule-based strategies are innovative attempts to improve FIMD methods; each strategy focuses on particular difficulties and constraints associated with the detection of image alteration.

Furthermore, different FIMD techniques are categorized based on the domains that they target for image feature assessment, with two prominent categories being spatial-domain based references [26-30, 53-56] and transform-domain based references [57-62]. In the study [53], an FIMD technique was introduced that incorporates a preprocessing stage employing a modified Capsule Network (CapsNet) with an enhanced routing mechanism. This approach showcased superior performance when applied to the DFDC-P dataset. Moving to Hu et al.'s study [26], the algorithm detects corneal specular reflections by analyzing the disparity in light source reflections in the eyes. Handcrafted features extracted from the spatial domain are employed for manipulation detection. A drawback here is the reliance on high-quality images, which may not align with the typical quality of images found on most social networks. Addressing the limitations [26], an enhanced algorithm [27] incorporated a super-resolution module to improve image quality before utilizing a CNN and analyzing differences in handcrafted features for manipulation detection. Meanwhile, Yang et al. [28] relied on handcrafted features containing facial landmark locations, such as eye, nose, and mouth tips, to identify manipulation through the detection of unnatural feature placements. In the reference [29], another handcrafted feature-based algorithm detected manipulation by identifying deviations in natural correlations among color bands. In references [29, 30], an improved version was introduced using special features to identify tampering and applying filtering to the image's chrominance components.

Nataraj et al. [54] introduced yet another handcrafted feature-based algorithm that uncovered invisible artifacts in high-frequency image components. Co-occurrence matrices were extracted from image channels and used as input for manipulation detection via a CNN. To expedite processing, Barni et al. [55] utilized pre-trained CNN models [63] instead of traditional CNN modules. In the study [56], manipulation detection relied on network layer activity using the Deepxplore algorithm. These FIMD approaches span a spectrum of techniques within the spatial-domain category, each addressing distinct challenges and employing various feature extraction methods for image manipulation detection.

Frequency artifacts serve as input features for the networks [57-62]. In the study [57], the approach employs a k-nearest neighbors (KNN) classifier [64] with energy spectral distribution as an input feature. Mi et al. [58] used the frequency spectrum as input for a CNN, detecting pronounced peaks in the spectrum when manipulation occurs. Zhang et al. [59] explored various network architectures, datasets, and resolutions to identify artifacts used as input for a CNN classifier. It involves post-processing and spectral loss during GAN training to derive fitting parameters for manipulation detection [60-62].

To avoid the limitations of the deep learning-based methods, Salih et al. [32, 33] presented a watermarking-based FIMD technology. The scheme consists of two main stages that are face region detection using Multi-task Cascaded Neural Network (MTCNN) and Slantlet transform (SLT)-based algorithms. The process commences with the identification of the facial area through MTCNN, providing data about the boundaries of the face box. Subsequently, the output from MTCNN is fine-tuned to define the precise pixels within the

facial region. Based on the adjusted outcome, a mask image is then generated for the purpose of categorizing image blocks. The mask image is generated using zeros then the pixels related to the face region are set to ones. The initial facial image and the mask image are both segmented into blocks each measuring 16×16 pixels. These blocks are then categorized into two distinct groups: those included in the facial region and those located outside it. The manipulation localization data is derived from the blocks within the facial region and is subsequently incorporated into the blocks situated outside the facial region using the SLT-based watermarking method. The scheme obtained promising performance compared to the deep learning-based methods, however, the capability of restoring the face region after alterations is not available. Each FIMD technique has its pros and cons, and no method capable of addressing all limitations.

Based on the abovementioned review, this study proposes three distinct algorithms that are applied in order to enable the recovery of the original facial region. The best-performing method is then incorporated into the major algorithms of the proposed system. The suggested algorithms are explained in detail in the section that follows.

3. PROPOSED ALGORITHMS

The proposed approach comprises two primary algorithms known as the embedding and extraction algorithms. During the embedding phase, information related to manipulation detection and facial recovery is generated from the facial region and subsequently integrated into the non-facial area. In the extraction phase, this embedded information is retrieved from the non-facial area and employed for detecting manipulations and restoring the facial region in case manipulations are detected. The succeeding subsections elaborate on these proposed algorithms.

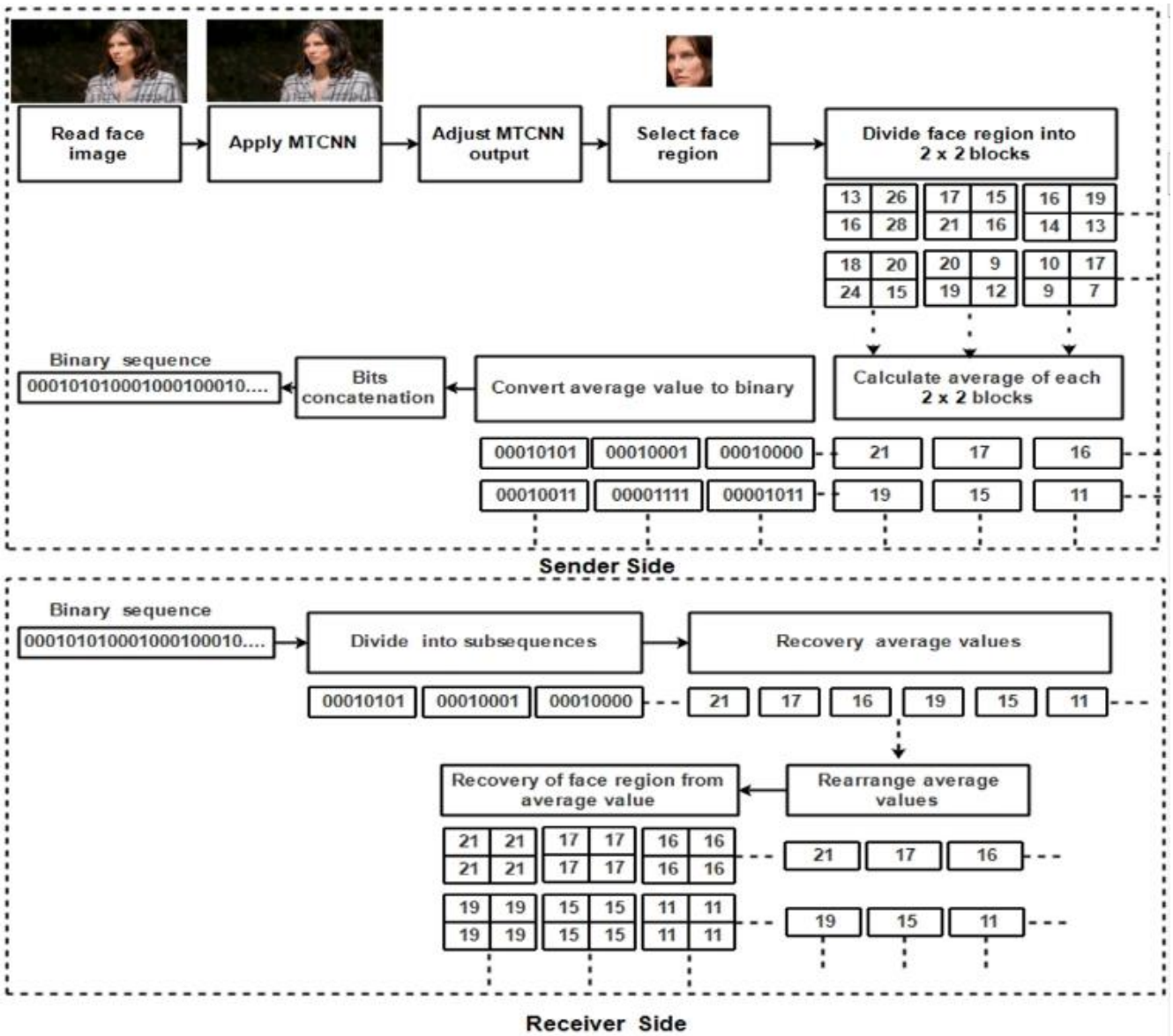

Figure 1. The algorithm suggested for creating recovery data through average (2×2)

3.1 Proposed algorithms for generating recovery information

Three suggested techniques are shown in this part for recovering the face region from the generated bits after obtaining recovery information from the face region and transforming it into a binary sequence. These techniques are based on the average of 2×2 and 4×4-pixel blocks, as well as

the Integer Wavelet Transform (IWT). To find the best algorithm for producing recovery information for the face region, a preliminary analysis was carried out on color face regions from different face images.

3.1.1 Recovery data generation using average (2×2) method

The first proposed recovery algorithm shown in Figure 1 starts at the sender side by reading the face image and detecting the face box using MTCNN-based algorithm [33]. The output parameters of the resultant face window are the width (w), height (h), top left corner (x_1, y_1), and bottom right corner (x_2, y_2). The parameters x_2 and y_2 are adjusted using the following equations to make the size of the window divisible by 16 (which is required at the embedding procedure):

$$x_{2\_new} = x_2 - Reminder\ (h, 16) \quad (1)$$

$$y_{2\_new} = y_2 - Reminder\ (w, 16) \quad (2)$$

The pixels of the selected face window are defined as $(x_1 : x_{2\_new}, y_1 : y_{2\_new})$. The chosen face window is split into 2×2-pixel blocks, and the average value of each block is computed and rounded to the closest integer number. After that, these average values are transformed into 8-bit binary sequences. The recovery data is created by concatenating these binary sequences to create a single binary sequence. To retrieve the average values, the binary sequence is split into 8-bit subsequences at the receiving end. Then, every 2×2-pixel block in the face region is rebuilt using these average values.

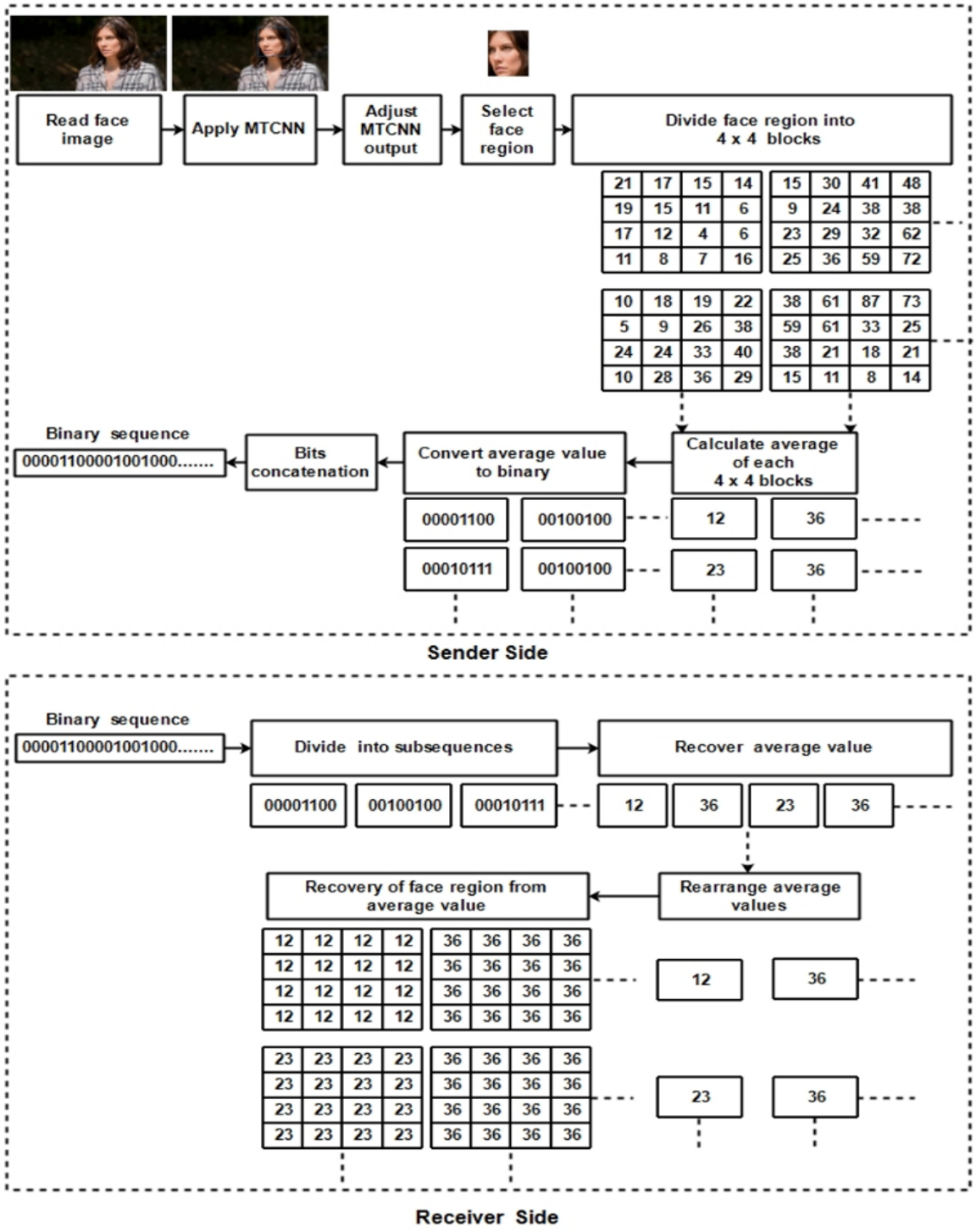

Figure 2. The algorithm suggested for creating recovery data through average (4×4)

3.1.2 Recovery data generation using average (4×4) method

The second proposed recovery algorithm shown in Figure 2 starts by reading the input image and identifying the face box. The size of the blocks, which are 4×4 pixels rather than 2×2 pixels, is the main distinction between this algorithm and the previous one. The generated recovery sequence has less length compared to the (2×2)-based algorithm which is at the cost of low visual quality of the recovered face region.

3.1.3 Recovery data generation using IWT method

The third proposed recovery algorithm shown in Figure 3 has the same starting procedure as explained in Section 3.1.1. After applying the parameters adjustment process using Eqs. (1) and (2), the selected face region is transformed using IWT. The resultant coefficients after transform are divided into four subbands called (approximation (*CA*), horizontal (*CH*), vertical (*CV*), and diagonal (*CD*)). The CA subbands is selected to generate the recovery data while the other subbands are ignored. The coefficients are adjusted using adjustment rules that have been presented by Tareef et al. [45]. Then, the resultant coefficients are converted to binary sequences each of length 8 bits and concatenated to form a single binary sequence which represents the recovery data.

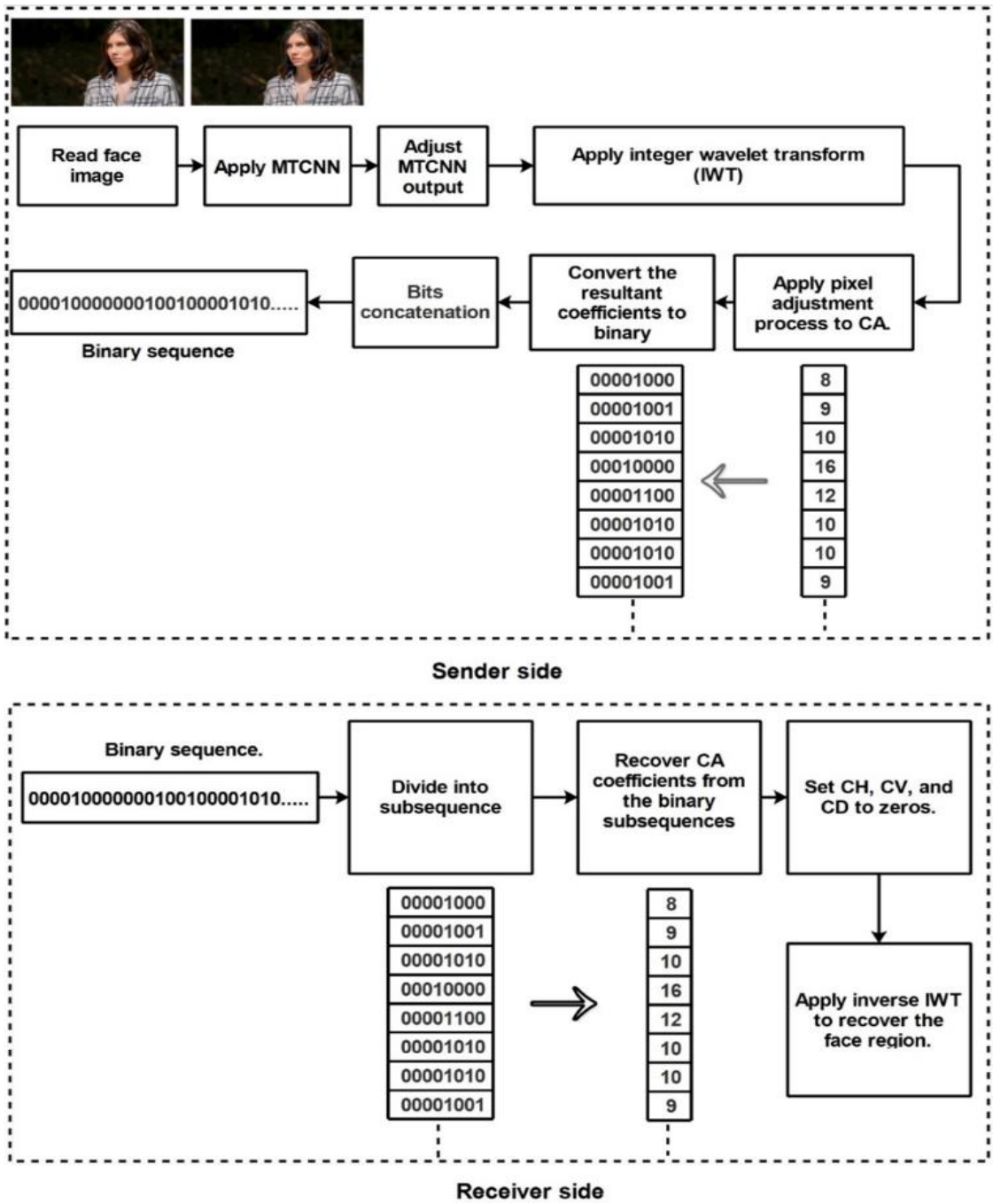

Figure 3. The algorithm suggested for creating recovery data through IWT

At the receiver side, the binary sequence is divided into subsequences and *CA* coefficients are recovered and rearranged to form the *CA* subband. The other three subbands (i.e., *CH*, *CV*, and *CD*) are set to zeros, then the face region is retrieved using inverse IWT.

The wavelet transform has different families, therefore, a test of visual quality using Peak Signal-to-Noise Ratio (*PSNR*) at different wavelet types has been performed to discover the

wavelet type that gives the highest PSNR values. Samples of experimental outcomes for testing various wavelet families are shown in Table 1 which proved that the IWT (cdf 3.5) obtained the best results.

Table 1. Comparison of PSNR for different wavelet families

Wavelet Type	PSNR (dB)				
	Image 1	**Image 2**	**Image 3**	**Image 4**	**Image 5**
bior5.5	24.3623	18.0065	18.8657	17.415	17.7289
cdf1.1	32.0368	37.178	33.5153	34.9857	30.2974
cdf1.3	31.9584	37.0745	33.4299	34.9069	30.2254
cdf1.5	31.8989	37.0062	33.3642	34.8596	30.1664
cdf2.2	32.0336	37.3676	33.2941	33.263	29.6802
cdf2.4	32.0741	37.4101	33.3312	33.2974	29.717
cdf2.6	32.079	37.4246	33.3373	33.3031	29.7214
cdf3.1	32.7623	38.1706	34.5836	35.4071	30.3429
cdf3.3	32.9486	38.3417	34.7419	35.5681	30.5224
cdf3.5	**32.9982**	**38.394**	**34.788**	**35.6123**	**30.5773**
cdf4.2	31.9031	37.0988	33.3159	33.397	29.5387
cdf4.4	32.1684	37.4017	33.5695	33.7094	29.8129
cdf4.6	32.2646	37.5202	33.6766	33.8286	29.9253
cdf5.1	31.1605	36.3035	33.2491	33.8112	28.7295
cdf5.3	30.5142	35.307	32.6954	33.0712	28.0457
cdf5.5	32.2911	37.0001	34.016	34.5233	29.7416
cdf6.2	28.9913	33.5113	30.3685	30.808	27.007
cdf6.4	30.1049	34.5366	31.3998	31.7553	28.0684
cdf6.6	30.5497	34.9254	31.7538	32.1625	28.4905
db2	32.7411	37.6525	34.4431	35.2105	29.9604
db3	32.5144	36.8807	34.222	33.9149	28.7998
db6	31.274	34.1777	32.7085	30.8518	26.2203
db8	30.1001	32.0581	30.9866	29.0037	24.9296
Haar	32.0368	37.178	33.5153	34.9857	30.2974
sym2	32.7411	37.6525	34.4431	35.2105	29.9604
sym3	14.6255	10.4511	10.2051	9.8947	11.1383
sym4	32.3373	37.7897	33.7536	34.0072	30.0528
sym5	32.856	37.8219	34.5598	35.3588	30.5005
sym6	24.8445	26.9725	24.1389	20.751	18.4486
sym7	25.0551	27.7253	24.8635	21.3963	19.0716
Max. PSNR	**32.9982**	**38.394**	**34.788**	**35.6123**	**30.5773**

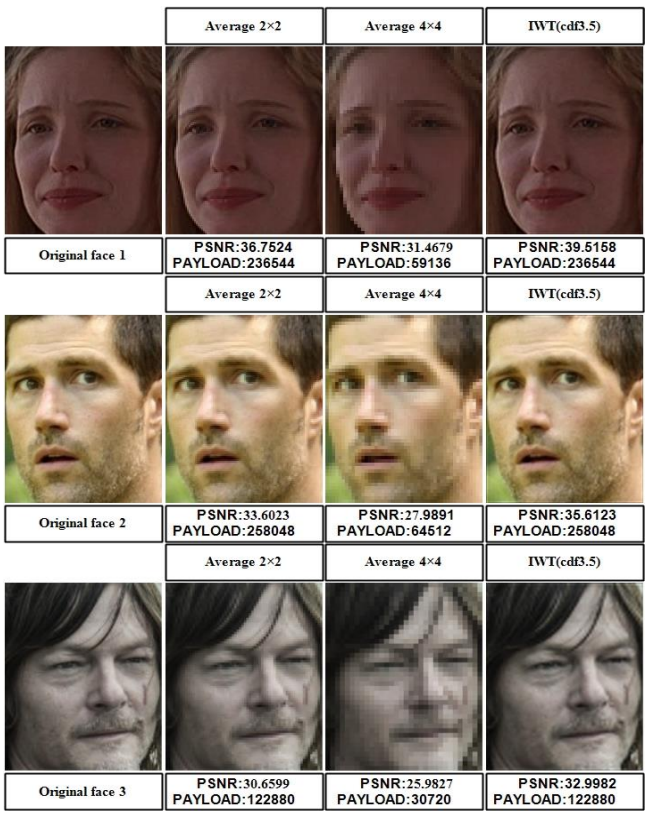

Figure 4. Comparison between the three recovery algorithms for sample test images

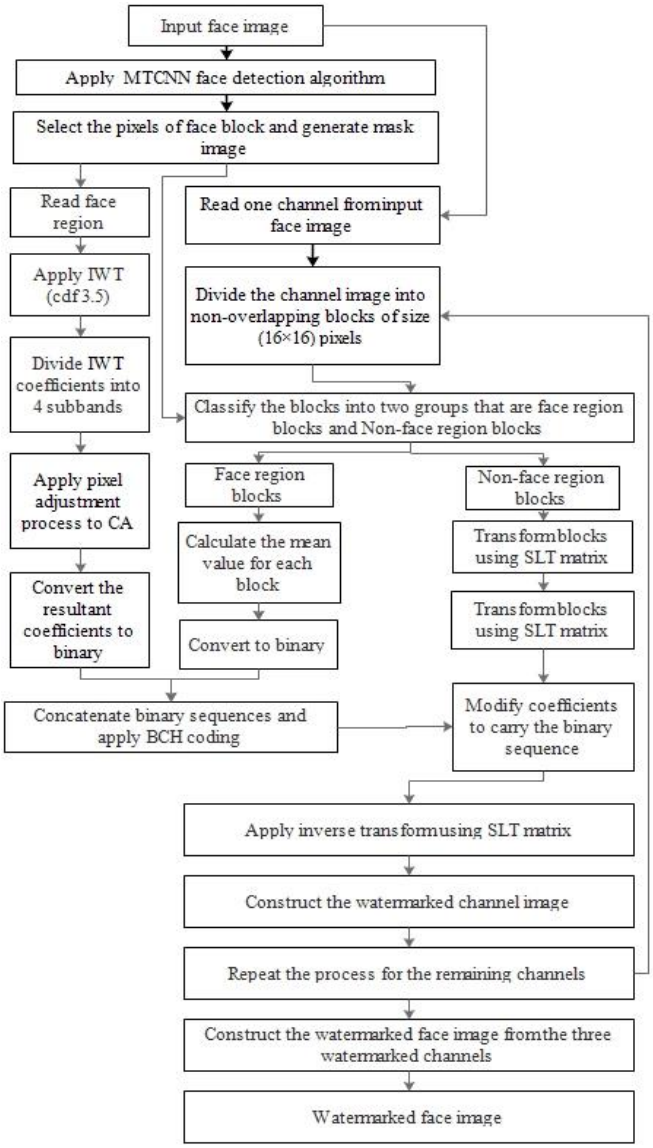

Figure 5. Block diagram for the proposed embedding algorithm

3.1.4 Comparison between the three suggested algorithms

The proposed work in this paper is intended to be applied in real-world applications, therefore, different face images have been collected from various websites [65-67] for experimental tests.

To evaluate the three recovery algorithms, sample test images have been used. The recovered face regions, their corresponding *PSNR* values, and the payload lengths are shown in Figure 4. The strength of the average (2×2) and the average (4×4) algorithms is their easy to implement procedure which requires direct processing of the image pixels. While the limitation of these two methods is the degradation in the *PSNR* values. The results proved that the payload of average (4×4) method is lower than the other methods, however, the *PSNR* values are very low, therefore, this method has not been adopted in the proposed scheme. The average (2×2) and IWT (cdf 3.5) obtained the same length of the binary sequence but the visual quality results using IWT (cdf 3.5) are higher, consequently, the IWT (cdf 3.5)-based algorithm has been adopted in the proposed scheme. Although the IWT-based algorithm is complex compared to the two other suggested algorithms, it can be recommended for practical applications because it can obtain a good compromise between the length of the generated data and the *PSNR* values of the retrieved face region.

3.2 Embedding algorithm

The embedding algorithm introduced here operates on the sender's end, taking the original facial image $I_f(M \times N \times 3)$ as input and producing the watermarked facial image $I_w(M \times N \times 3)$ as output, with M representing the height and N signifying the width of the original facial image. Figure 5 illustrates the block diagram of this embedding algorithm, while the specific steps of the algorithm are detailed below:

3.2.1 Face area detection and selection

MTCNN is utilized for face box detection within I_f, and the outcome of this process is adjusted to isolate the pixels within the facial area, following the procedure detailed by Salih et al. [32].

3.2.2 Mask image generation

Based on the outcome of the prior step, a binary mask image with dimensions (M×N) is produced, where pixels within the facial region are set to '1' while pixels outside the facial region are set to '0'.

3.2.3 Generation of recovery data

The algorithm starts by reading the face region and applying the IWT (cdf 3.5) to each channel from the face region. The output coefficients are divided into four subbands called (CA, CH, CV, and CD). Only CA coefficients are selected and adjusted to ensure their values are between (0 to 255) using the following:

$$CA_{new}(i,j) = \begin{cases} 0 & if\ CA(i,j) < 0 \\ 255 & if\ CA(i,j) > 255 \\ CA(i,j) & if\ 0 \le CA(i,j) \le 255 \end{cases} \quad (3)$$

where, $CA(i,j)$ and $CA_{new}(i,j)$ are the original and the adjusted approximation coefficients, respectively.

Each coefficient in $CA_{new}(i,j)$ is converted to 8 bits binary number thereafter one binary sequence called Bin_{CA} is generated from the binary representation of the coefficients. The process of generating the recovery information from one channel of the color face region is repeated to obtain the recovery data binary sequences for the three channels.

3.2.4 Dividing images and blocks classification

As depicted in Figure 5, a single channel from I_f is subdivided into blocks. The mask image is processed using the same process. The channel blocks are then divided into two groups based on the average of the mask image blocks: those that belong to the non-facial region and those that belong to the facial region. The equivalent channel block at the same place is classed as part of the face region if the average of a mask image block is not zero, and as part of the non-facial region otherwise.

3.2.5 Generation of localization data

The localization data is produced by deriving average values for each block within the facial region. These average values are transformed into binary format, and the resulting binary sequences are combined to form a single binary sequence referred to as Bin_{AVG}. This procedure of generating the localization data is replicated for all three channels of I_f.

3.2.6 Binary sequence embedding

The Bin_{CA} and Bin_{AVG} are concatenated to form Bin_{Seq} which must be embedded in the blocks belong to non-facial region. As explained by Salih et al. [32], the SLT-based watermark embedding procedure is applied to embed the Bin_{Seq} in the blocks and obtain the watermarked ones.

3.2.7 Watermarked image construction

The watermarked blocks are structured to create the watermarked channel, and this process is iterated to acquire the three watermarked channels. The final watermarked facial image is assembled by combining these watermarked channels.

3.3 Extraction algorithm

The extraction algorithm suggested here is implemented on the receiver's end, with the watermarked facial image as the input and the output representing the outcome of facial image authentication. The block diagram of this extraction technique is presented in Figure 6, and the individual steps of the algorithm are described below:

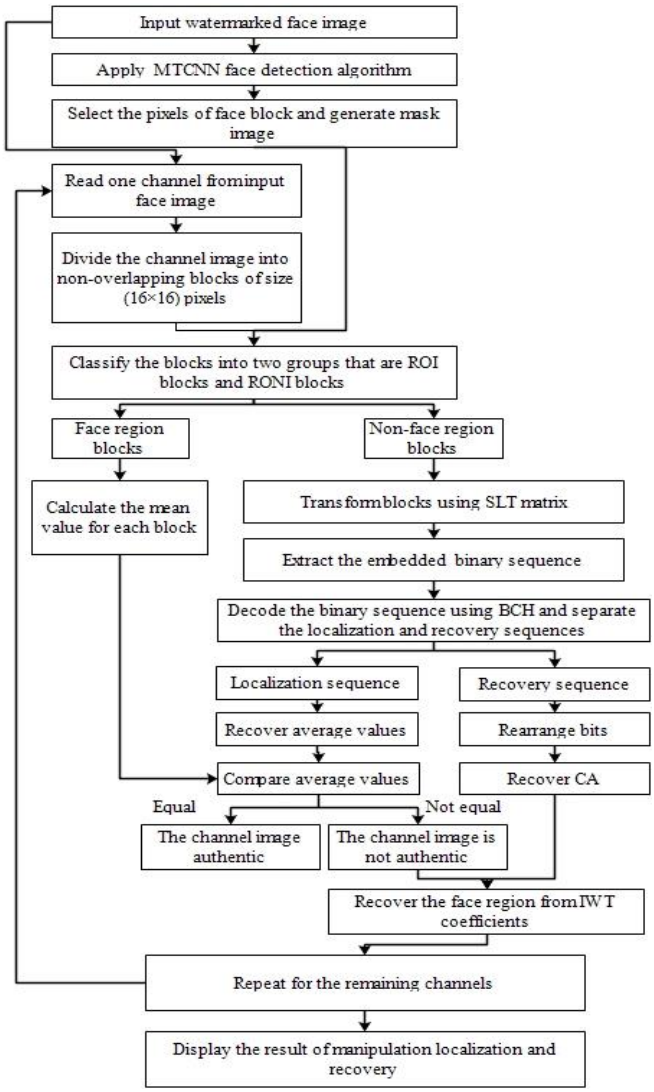

Figure 6. Block diagram for the proposed extraction algorithm

3.3.1 Face area detection and selection

The process that was applied at subsection (3.2.1) is replicated here.

3.3.2 Mask image generation

The process that was applied at subsection (3.2.2) is replicated here.

3.3.3 Dividing images and blocks classification

The process that was applied at subsection (3.2.4) is replicated here.

3.3.4 Binary sequence extraction and separate sequences

At this point, the binary sequence encoded in the blocks outside of the face region is extracted using the data extraction technique from [32]. Two subsequences are separated from the extracted Bin_{Seq} : one for recovery and the other for localization (i.e., Bin_{CA} and Bin_{AVG}).

3.3.5 Manipulation localization

Calculate the average values of the face region blocks from step 3.3.3. Recover the average values from binary sequence Bin_{AVG} that have been extracted in step 3.3.4. Compare the extracted and calculated average values to detect alterations. If the average values are identical, the block is deemed authentic. If the average values differ, the block is considered unauthentic, and it is localized by drawing a border around its pixels. The procedure then continues to recover the face region.

3.3.6 Recovery of face region

Recover CA coefficients from the binary sequence (Bin_{CA}) then set the other subbands (i.e., CH, CV, and CD) to zeros. Apply inverse IWT (cdf 3.5) to recover the face region.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Using a range of test images, experiments have been conducted to assess the effectiveness of the suggested technique. Examples of these test images can be seen in Figure 7, featuring images of varying dimensions and different sizes of facial regions [65-67]. The subsequent sections will present the experiments along with their discussions, culminating in a comprehensive comparison with existing FIMD schemes.

Figure 7. Sample test image

The major findings of these tests showed that the size of the chosen face region affects the overall number of bits in the binary sequence, while the dimensions of the original image and the face region dictate the embedding capacity. The

suggested approach is able to recover the face region with excellent visual quality and embeds the binary sequence into the image correctly and without producing any visual distortions. Various manipulations have been imposed to the test images and the outcomes are promising as illustrated in the following subsections. The efficiency of the suggested algorithms in identifying various manipulations and retrieving the original face region when such manipulations are detected was demonstrated by a comparison with earlier FIMD systems.

4.1 Capacity and payload test

The payload is the length of the binary sequence created from the localization and recovery data, whereas capacity is the total amount of bits that can be placed in the non-face region. The experiment's findings for the sample photos in Figure 7 are presented in Table 2. The results verified that the payload grows along with the face region size and vice versa. The size of the face region and the original image both affect capacity; a larger ratio between the size of the face region and the original image results in a lower capacity.

Table 2. Capacity and payload test results

Image Name	Image Size M×N×3	Size of Face Region H×W×3	Payload (bits)	Capacity (bits)
Img_1	570×1014	160×128	170880	404352
Img_2	600×1200	112×80	74880	523584
Img_3	1152×2048	432×384	1379520	1639872
Img_4	1152×2048	288×224	537216	1714752
Img_5	536×1024	144×144	172992	386304
Img_6	600×1200	224×192	358464	495360
Img_7	640×800	80×64	43008	378240
Img_8	1669×2500	256×224	477120	3069312
Img_9	455×728	128×128	136896	226368
Img_10	608×1080	224×176	328512	454272

Notes: M = height and N = width of the original face image.
H = height and W = width of the selected face region.

4.2 Visual quality and time complexity test

The PSNR between the original facial image and its watermarked version was computed in order to assess the images' visual quality after embedding data. To calculate the time complexity of the extraction and embedding techniques, MATLAB's "tic toc" commands were utilized. Eight gigabytes of RAM and a 2.60GHz Intel® Core TM i7 CPU powered the experiment's PC. The outcomes, displayed in Table 3, demonstrate the efficacy of the suggested algorithms in generating images of superior quality. In addition, the analysis in this experiment showed that the embedding process is faster than the extraction procedure.

Table 3. Visual quality and time complexity test results

Image Name	PSNR (dB)	Embedding Time (sec.)	Extraction Time (sec.)
Img_1	45.0877	0.872964	11.360993
Img_2	54.0449	0.960417	7.56287
Img_3	44.3761	3.070259	23.72201
Img_4	48.2253	2.479163	37.787807
Img_5	42.1705	0.881936	17.123695
Img_6	46.1548	1.094932	24.019591
Img_7	53.0178	0.76407	10.367643
Img_8	56.5061	3.642619	35.756902
Img_9	38.336	0.867687	13.792615
Img_10	46.1699	1.021984	24.837479

Figure 8. Manipulations reveal and face region recovery for 'Img_1'

Figure 9. Manipulations reveal and face region recovery for 'Img_8'

Figure 10. Manipulations reveal and face region recovery for 'Img_7'

4.3 Face manipulation localization and recovery

The watermarked facial images were subjected to a variety of attacks in order to evaluate the scheme's capability to identify manipulations within the facial region and restore the original facial area when such manipulations occur. Sample findings, shown in Figures 8 to 10, show that, regardless of the size of the modified area, the suggested scheme can successfully recover the original facial area and identify the altered region with accuracy.

4.4 Comparison with the state-of-the-art schemes

By efficiently detecting different types of manipulations, precisely identifying the modified parts within the facial area, and reconstructing the facial region when manipulations are identified, the suggested methodology outperforms multiple state-of-the-art methods. A thorough comparison of this method and other FIMD systems is provided in Table 4.

Compared to deep learning-based algorithms that necessitate extensive training times, watermarking-based algorithms demonstrate greater efficiency in terms of both

speed and accuracy. Deep learning-based algorithms tend to excel when applied to specific test images closely resembling the training dataset, while watermarking-based algorithms exhibit versatility, capable of being employed on a wider range of input images. Notably, the proposed algorithm outperforms the scheme described in studies [32, 33] because it possesses the capability to recover the original facial region, an aspect in which the latter schemes fall short.

Table 4. Comparison with existing FIMD schemes

Characteristics	Schemes [25-30]	Scheme [32, 33]	Proposed Scheme
Methodology	Deep-learning	Watermarking	Watermarking & IWT
Detection of various manipulations	×	✓	✓
Manipulation localization	×	✓	✓
Required training	✓	×	×
Accuracy	Less than 100%	100%	100%
Specific test images	✓	×	×
Recover the face region after detection	×	×	✓

5. CONCLUSIONS

This paper introduces a new algorithm designed to detect alterations in digital facial images and, in the event of detecting manipulations, restore the original facial region. Three distinct algorithms have been proposed to generate recovery information, each based on different approaches: an average (2×2) method, an average (4×4) method, and an IWT method. When employing the IWT-based algorithm, various wavelet families were tested to identify the one yielding the highest visual quality in the recovered facial region. The outcomes of the best IWT variant (specifically, cdf 3.5) were compared with the results obtained from the other two algorithms. The IWT (cdf 3.5)-based approach struck a favorable balance between visual quality and payload length, leading to its inclusion in the proposed facial image authentication scheme.

Extensive experiments were conducted to assess the effectiveness of the proposed scheme, demonstrating its capability to detect various facial image manipulations and subsequently restore the facial region following manipulation detection. A comprehensive comparison with existing state-of-the-art methods affirmed the superiority of the proposed scheme.

The proposed algorithms carry significant implications across various domains. In digital forensics, it can be used to ensure the reliability of evidence by preserving the integrity of images, strengthening their admissibility in legal proceedings, and aiding in the identification of cybercriminals, fraudsters, and identity thieves. It can help in uncover critical information hidden within manipulated images, facilitating thorough analysis and potentially solving complex cases. On the privacy front, face recovery safeguards individuals by countering invasive image manipulation and harassment, maintaining their privacy and dignity. Moreover, it plays a pivotal role in preventing false accusations, exonerating innocent individuals wrongly implicated through manipulated images. This contributes to a fairer legal system and protects innocent individuals from unwarranted harm. In matters of public safety, face recovery ensures accurate identification, enhancing security measures and aiding law enforcement in maintaining order and protecting citizens.

As illustrated in this work, the proposed algorithm can successfully be applied in various fields, however, the restricted embedding capacity can be considered as a limitation. For instance, the algorithm cannot be applied when the face region is very large compared to the size of the original image. The future researches can be conducted in different directions such as enhancing the embedding capacity, implementing a real-time detection system for live video streams, and investigating the main requirements for efficient algorithm's execution on hardware devices.

ACKNOWLEDGMENT

The authors express their gratitude to Al-Iraqia University and Dijlah University College for supporting and encouraging their researches.

REFERENCES

[1] Al-Najjar, H., Alharthi, S., Atrey, P.K. (2016). Secure image sharing method over unsecured channels. Multimedia Tools and Applications, 75: 2249-2274. https://doi.org/10.1007/s11042-014-2404-5

[2] Hodeish, M.E., Bukauskas, L., Humbe, V.T. (2022). A new efficient TKHC-based image sharing scheme over unsecured channel. Journal of King Saud University-Computer and Information Sciences, 34(4): 1246-1262. https://doi.org/10.1016/j.jksuci.2019.08.004

[3] Faircloth, J. (2014). Chapter 5-Information Security. In Enterprise Applications Administration, Boston: Morgan Kaufmann, pp. 175-220. https://doi.org/10.1016/B978-0-12-407773-7.00005-3

[4] Manivannan, D., Brindha, M. (2022). Secure image cloud storage using homomorphic password authentication with ECC based cryptosystem. Advances in Systems Science and Applications, 22(1): 92-116. https://doi.org/10.25728/assa.2022.22.1.1175

[5] Schultz, V.L., Kul'ba, V.V., Zaikin, O.A., Shelkov, A.B., Chernov, I.V. (2017). Regional security: Analysis of the emergency management effectiveness based on the scenario approach. Advances in Systems Science and Applications, 17(1): 9-24. https://doi.org/10.25728/assa.2017.17.1.252

[6] Abuali, M.S., Rashidi, C.B.M., Raof, R.A.A., Ku Azir, K.N.F., Hussein, S.S., Abd-Alhasan, A.Q. (2024). Enhancing security with multi-level steganography: A dynamic least significant bit and wavelet-based approach. Mathematical Modelling of Engineering Problems, 11(6): 1403-1416. https://doi.org/10.18280/mmep.110602

[7] Abood, M.H., Abdulmajeed S.W. (2022). High security image cryptographic algorithm using chaotic encryption algorithm with Hash-LSB steganography. Al-Iraqia Journal for Scientific Engineering Research, 1(2): 65-74. https://doi.org/10.58564/IJSER.1.2.2022.53

[8] Abdulridha Muttashar, R., Sami Fyath, R. (2023). Triple color image encryption using hybrid digital/optical scheme supported by high-order chaos. Al-Iraqia Journal for Scientific Engineering Research, 2(1): 68-79.

[9] Kloppenburg, S., Van der Ploeg, I. (2020). Securing identities: Biometric technologies and the enactment of

human bodily differences. Science as Culture, 29(1): 57-76. https://doi.org/10.1080/09505431.2018.1519534
[10] Galbally, J., Marcel, S., Fierrez, J. (2014). Biometric antispoofing methods: A survey in face recognition. IEEE Access, 2: 1530-1552. https://doi.org/10.1109/ACCESS.2014.2381273
[11] Zhao, K., Wang, D., Wang, Y. (2019). A face recognition algorithm based on optimal feature selection. Revue d'Intelligence Artificielle, 33(2): 105-109. https://doi.org/10.18280/ria.330204
[12] Marcel, S., Nixon, M.S., Fierrez, J., Evans, N. (2019). Handbook of biometric anti-spoofing: Presentation attack detection. Cham, Switzerland: Springer, vol. 2. https://doi.org/10.1007/978-3-319-92627-8
[13] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64: 131-148. https://doi.org/10.1016/j.inffus.2020.06.014
[14] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5): 910-932. https://doi.org/10.1109/JSTSP.2020.3002101
[15] Czajka, A., Kasprzak, W., Wilkowski, A. (2016). Verification of iris image authenticity using fragile watermarking. Bulletin of the Polish Academy of Sciences. Technical Sciences, 64(4): 807-819. http://dx.doi.org/10.1515%2Fbpasts-2016-0090
[16] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems.
[17] Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2): 135-146. https://doi.org/10.1016/j.bushor.2019.11.006
[18] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. (2022). An introduction to digital face manipulation. In Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition. Springer, Cham. https://doi.org/10.1007/978-3-030-87664-7_1
[19] Vamsi, V.V.V.N.S., Shet, S.S., Reddy, S.S.M., Rose, S.S., Shetty, S.R., Sathvika, S., Supriya, M.S., Shankar, S.P. (2022). Deepfake detection in digital media forensics. Global Transitions Proceedings, 3(1): 74-79. https://doi.org/10.1016/j.gltp.2022.04.017
[20] Talib, D.A., Abed, A.A. (2023). Real-time deepfake image generation based on stylegan2-ADA. Revue d'Intelligence Artificielle, 37(2): 397-405. https://doi.org/10.18280/ria.370216
[21] Sharma, J., Sharma, S., Kumar, V., Hussein, H.S., Alshazly, H. (2022). Deepfakes classification of faces using convolutional neural networks. Traitement du Signal, 39(3): 1027-1037. https://doi.org/10.18280/ts.390330
[22] Korshunov, P., Marcel, S. (2018). Deepfakes: A new threat to face recognition? Assessment and detection. arXiv Preprint arXiv: 1812.08685. https://doi.org/10.48550/arXiv.1812.08685
[23] Saxena, A., Yadav, D., Gupta, M., Phulre, S., Arjariya, T., Jaiswal, V., Bhujade, R.K. (2023). Detecting deepfakes: A novel framework employing XceptionNet-based convolutional neural networks. Traitement du Signal, 40(3): 835-846. https://doi.org/10.18280/ts.400301
[24] Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, pp. 2382-2390. https://doi.org/10.1145/3394171.3413769
[25] Matern, F., Riess, C., Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, USA, pp. 83-92. https://doi.org/10.1109/WACVW.2019.00020
[26] Hu, S., Li, Y., Lyu, S. (2021). Exposing GAN-generated faces using inconsistent corneal specular highlights. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 2500-2504. https://doi.org/10.1109/ICASSP39728.2021.9414582
[27] Han, X., Ji, Z., Wang, W. (2020). Low resolution facial manipulation detection. In 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, pp. 431-434. https://doi.org/10.1109/VCIP49819.2020.9301796
[28] Yang, X., Li, Y., Qi, H., Lyu, S. (2019). Exposing GAN-synthesized faces using landmark locations. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, pp. 113-118. https://doi.org/10.1145/3335203.3335724
[29] McCloskey, S., Albright, M. (2019). Detecting GAN-generated imagery using saturation cues. In 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 4584-4588. https://doi.org/10.1109/ICIP.2019.8803661
[30] Li, H., Li, B., Tan, S., Huang, J. (2018). Detection of deep network generated images using disparities in color components. arXiv 2018. arXiv Preprint arXiv: 1808.07276.
[31] Salih, Z.A., Thabit, R., Zidan, K.A., Khoo, B.E. (2022). Challenges of face image authentication and suggested solutions. In 2022 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, pp. 189-193. https://doi.org/10.1109/ICITSI56531.2022.9970797
[32] Salih, Z.A., Thabit, R., Zidan, K.A., Khoo, B.E. (2022). A new face image manipulation reveal scheme based on face detection and image watermarking. In 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, pp. 1-6. https://doi.org/10.1109/IICAIET55139.2022.9936838
[33] Salih, Z.A., Thabit, R.A.S.H.A., Zidan, K.A. (2023). A new manipulation detection and localization scheme for digital face images. Journal of Engineering Science and Technology, 18(2): 1164-1183.
[34] Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503. https://doi.org/10.1109/LSP.2016.2603342
[35] Thabit, R., Khoo, B.E. (2014). A new robust reversible watermarking method in the transform domain. In the 8th International Conference on Robotic, Vision, Signal

Processing & Power Applications: Innovation Excellence Towards Humanistic Technology, Springer Singapore, pp. 161-168. https://doi.org/10.1007/978-981-4585-42-2_19

[36] Li, L.D., Qian, J.S., Pan, J.S. (2011). Characteristic region based watermark embedding with RST invariance and high capacity. AEU - International Journal of Electronics and Communications, 65(5): 435-442. https://doi.org/10.1016/j.aeue.2010.06.001

[37] Allaf, A.H., Kbir, M.A. (2019). A review of digital watermarking applications for medical image exchange security. In: Ben Ahmed, M., Boudhir, A., Younes, A. (eds) Innovations in Smart Cities Applications. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham. https://doi.org/10.1007/978-3-030-11196-0_40

[38] Swaraja, K. (2018). Medical image region based watermarking for secured telemedicine. Multimedia Tools and Applications, 77(21): 28249-28280. https://doi.org/10.1007/s11042-018-6020-7

[39] Yu, X., Wang, C., Zhou, X. (2017). Review on semi-fragile watermarking algorithms for content authentication of digital images. Future Internet, 9(4): 56. https://doi.org/10.3390/fi9040056

[40] BW, T.A., Permana, F.P. (2012). Medical image watermarking with tamper detection and recovery using reversible watermarking with LSB modification and run length encoding (RLE) compression. In 2012 IEEE International Conference on Communication, Networks and Satellite (ComNetSat), Bali, Indonesia, pp. 167-171. https://doi.org/10.1109/ComNetSat.2012.6380799

[41] Zain, J.M., Fauzi, A.R. (2007). Evaluation of medical image watermarking with tamper detection and recovery (AW-TDR). In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, pp. 5661-5664. https://doi.org/10.1109/IEMBS.2007.4353631

[42] Zain, J.M., Fauzi, A.R. (2006). Medical image watermarking with tamper detection and recovery. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, USA, pp. 3270-3273. https://doi.org/10.1109/IEMBS.2006.260767

[43] Chiang, K.H., Chang-Chien, K.C., Chang, R.F., Yen, H.Y. (2008). Tamper detection and restoring system for medical images using wavelet-based reversible data embedding. Journal of Digital Imaging, 21: 77-90. https://doi.org/10.1007/s10278-007-9012-0

[44] Kulkarni, M.B., Patil, R.T. (2012). Tamper detection & recovery in medical Image with secure data hiding using Reversible watermarking. International Journal of Emerging Technology and Advanced Engineering, 2(3): 370-373.

[45] Tareef, A., Al-Ani, A., Nguyen, H., Chung, Y.Y. (2014). A novel tamper detection-recovery and watermarking system for medical image authentication and EPR hiding. In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, USA, pp. 5554-5557. https://doi.org/10.1109/EMBC.2014.6944885

[46] Bappy, J.H., Roy-Chowdhury, A.K., Bunk, J., Nataraj, L., Manjunath, B.S. (2017). Exploiting spatial structure for localizing manipulated image regions. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 4970-4979. https://doi.org/10.1109/ICCV.2017.532

[47] Cristin, R., Ananth, J.P., Cyril Raj, V. (2018). Illumination-based texture descriptor and fruit fly support vector neural network for image forgery detection in face images. IET Image Processing, 12(8): 1439-1449. https://doi.org/10.1049/iet-ipr.2017.1120

[48] Dang, L.M., Hassan, S.I., Im, S., Moon, H. (2019). Face image manipulation detection based on a convolutional neural network. Expert Systems with Applications, 129: 156-168. https://doi.org/10.1016/j.eswa.2019.04.005

[49] Nguyen, H.H., Yamagishi, J., Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv Preprint arXiv: 1910.12467. https://doi.org/10.48550/arXiv.1910.12467

[50] Nguyen, H.H., Yamagishi, J., Echizen, I. (2022). Capsule-forensics networks for deepfake detection. In Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks. Cham: Springer, pp. 275-301. https://doi.org/10.1007/978-3-030-87664-7

[51] Khalil, S.S., Youssef, S.M., Saleh, S.N. (2021). A multi-layer capsule-based forensics model for fake detection of digital visual media. In 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, United Arab Emirates, pp. 1-6. https://doi.org/10.1109/ICCSPA49915.2021.9385719

[52] Cao, L., Sheng, W., Zhang, F., Du, K., Fu, C., Song, P. (2021). Face manipulation detection based on supervised multi-feature fusion attention network. Sensors, 21(24): 8181. https://doi.org/10.3390/s21248181

[53] Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, USA, pp. 384-389. https://doi.org/10.1109/MIPR.2018.00084

[54] Nataraj, L., Mohammed, T.M., Chandrasekaran, S., Flenner, A., Bappy, J.H., Roy-Chowdhury, A.K., Manjunath, B.S. (2019). Detecting GAN generated fake images using co-occurrence matrices. arXiv Preprint arXiv: 1903.06836. https://doi.org/10.48550/arXiv.1903.06836

[55] Barni, M., Kallas, K., Nowroozi, E., Tondi, B. (2020). CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, USA, pp. 1-6. https://doi.org/10.1109/WIFS49906.2020.9360905

[56] Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, Seattle, USA, pp. 5781-5790. https://doi.org/10.1109/CVPR42600.2020.00582.

[57] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In International Conference on Machine Learning. PMLR, pp. 3247-3258.

[58] Mi, Z., Jiang, X., Sun, T., Xu, K. (2020). GAN-generated image detection with self-attention mechanism against GAN generator defect. IEEE Journal of Selected Topics in Signal Processing, 14(5): 969-981. https://doi.org/10.1109/JSTSP.2020.2994523

[59] Zhang, X., Karaman, S., Chang, S.F. (2019). Detecting and simulating artifacts in gan fake images. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS), Delft, Netherlands, pp. 1-6. https://doi.org/10.1109/WIFS47025.2019.9035107

[60] Dzanic, T., Shah, K., Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. Advances in Neural Information Processing Systems. In NIPS'20, vol. 2020- Decem. Red Hook, NY, USA: Curran Associates Inc., 33: 3022-3032.

[61] Durall, R., Keuper, M., Keuper, J. (2020). Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, pp. 7890-7899. https://doi.org/10.1109/CVPR42600.2020.00791.

[62] Bonettini, N., Bestagini, P., Milani, S., Tubaro, S. (2021). On the use of Benford's law to detect GAN-generated images. In 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, pp. 5495-5502. https://doi.org/10.1109/ICPR48806.2021.9412944

[63] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251-1258.

[64] Sitawarin, C., Wagner, D. (2020). Minimum-norm adversarial examples on KNN and KNN based models. In 2020 IEEE Security and Privacy Workshops (SPW), San Francisco, USA, pp. 34-40. https://doi.org/10.1109/SPW50608.2020.00023

[65] Bainbridge, W.A., Isola, P., Oliva, A. (2013). The intrinsic memorability of face photographs. Journal of Experimental Psychology: General, 142(4): 1323-1334. https://psycnet.apa.org/doi/10.1037/a0033872

[66] 5 million faces – Top 14 free image datasets for facial recognition, iMerit, 2021. https://imerit.net/blog/5-million-faces-top-17-free-image-datasets-for-facial-recognition-all-pbm/.

[67] Human faces, Kaggle, 2020. https://www.kaggle.com/datasets/ashwingupta3012/human-faces.