ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 10, Issue 6, June 2023

Video Tamper Detection and Forgery Localization using Hybrid Deep Learning Framework

[1] Malle Raveendra*, [2] K Nagireddy

[1] Electronics and Communication Engineering, Jawaharlal Nehru Technological University, Anantapuramu, India
[2] Electronics and Communication Engineering, NBKR Institute of Science and Technology, Andhra Pradesh, India
Corresponding Author Email: [1] ravindra439@gmail.com, [2] knreddy1974@gmail.com

Abstract*— Nowadays, anomalous detection is of utmost importance in videos because of unusual activity (or) unauthorized changes in the video frame. Several techniques have been involved in authenticating and localizing video tampering but most of them are inefficient. To enhance the detection of video tampering, the new Hybrid adversarial deep network (Hybrid DeepNet) has been proposed. The videos are collected from three datasets such as VTD, MFC 18 and VIRAT. This model mainly consists of five ways to identify the tampered region of videos. They are double compression detection, noise filtering process, segmentation, feature extraction and tamper detection. Initially, the double compression process is evolved using an improved group of pictures (GOP) method and then, the rapid bilateral filter is used to remove the noise involved in the video. By using wavelet-based feature extraction, the features are extracted. The new hybrid adversarial deep network with white shark optimization (WSO) has been performed to detect and localize the forged region of video. Because of the WSO algorithm with a deep model, the proposed method is helpful for removing network loss. The performance of the proposed model is evaluated by comparing with existing model. The proposed model obtained 96.21%, 95.15%, 96.13% and 97.15% for accuracy, recall, F-score and precision, respectively.

***Keywords* - Hybrid DeepNet, segmentation, feature extraction, rapid bilateral filter, tamper detection, double compression.**

I. INTRODUCTION

Because of the rapid advancement of digital video methods, tampering with for editing the sequence of videos is more accessible than before [1]. Anyone can change (or) remove the content of a video sequence using video editing tools like Apple Final Cut Pro, Adobe after effects and Adobe premiere [2]. With the alteration of objects in the video, the same can sometimes spread over the internet to cause social severe security events. Many researchers have focused on detecting video tampering in recent years [3]. Based on a standard video tampering technique, the hackers try to remove existing things (or) add new objects to a video sequence [4]. Detecting video object tampering is a challenging exercise due to the image copy-move scientific approaches. The usage of new image forensics algorithms is not helpful because of computational cost [5]. Therefore, the image forensics based techniques are not applied in the video forensics process. To minimize the difficulty of the video forensics process, the temporal relationship between the video frames is analysed [6].

The devices such as gadgets, mobile phones and video sharing websites play a vital role in daily life for sharing visual information from one place to another place [7]. Based on the consideration of a court of law, the visual information is served as sufficient proof for verifying a person when the person is an accused (or) a victim [8]. Nowadays, User-friendly video editing software is easily accessed by anyone to change the object of the video source. So, the uniqueness (or) genuineness of videos cannot be granted [9]. By deleting (or) inserting objects, videos are edited based on bad (or) good purposes [10]. The tampering method mainly contains three steps [11]. They are re-compressing the video, decoding the video streams into image data and tampering with videos for specific purposes. A video is termed a frames that can be denoted as a sequence of images. The performance of video tampering attacks can happen in three ways [12]. To minimize the noise in the video tampering attacks, image tampering detection techniques are involved in videos [13].

Due to complex scenarios such as noise incurred (or) moving objects, the detection methods may not obtain favourable results. If the video edited, then the authentication process will provide the footprints of certain changes in the videos [14]. Video object segmentation, online and offline segmentation are the three variations of segmentation [15]. For the video object segmentation process, the object extraction has happened from the background of videos.

To find altered areas in videos that have been produced fraudulently, numerous research projects have been developed. Deep learning techniques based on CNN involve the processes of segmentation, detection and classification. However these models don't have the complexity or accuracy needed. The new model includes a hybrid Deepnet technique to analyse the video data and determine whether the input video is authentic (or fabricated) in order to address the aforementioned problem. This strategy is perfect for identifying video using a variety of unique functions. Both frame-level analysis and pixel-level analysis are aspects of this technique. The primary objective of this proposed model is presented as follows:

- To enhance the detection of video tampering, the new Hybrid adversarial deep network (Hybrid DeepNet) has been proposed
- The videos are collected from three datasets such as VTD, MFC 18 and VIRAT. This model mainly consists of five ways to identify the tampered region of videos.
- The double compression process is evolved using an improved group of pictures (GOP) method and then, the rapid bilateral filter is used to remove the noise involved in the video. By using wavelet-based feature extraction, the features are extracted.
- The segmentation process is happened using the AHSW algorithm. The proposed method is helpful for removing network loss.

The rest of the new model is mentioned as follows: Section 2 relates to various related papers in recent years. Section 3 denotes the proposed methodology. Section 4 denotes the results and discussion based on the MATLAB tool with three video datasets such as MFC 18, VIRAT and VTD datasets. Section 5 shows the conclusion and future work.

II. RELATED WORK

Raveendra et al. (2022) [16] implemented an adaptive segmentation based deep network for video tamper detection and location process. Initially, the compression of input video was processed using double compression with discrete cosine transform (DCT) method. To improve the compressed frame quality, double compression model was applied. For the segmentation process, the frames have segmented into different regions. The extraction process was attained using gabor wavelet transform (GWT) with hybrid deep neural (HDN) network based galactic swarm optimization. Three datasets were used such as VTD, VIRAT and MFC-18 using MATLAB tool.

Raveendra et al. (2020) [17] attained fine-tuned AlexNet with DWT-DCT markov features model to identify inter frame tampering detection. For the detection of particular object, Markov based technique was used. The ImageNet dataset was applied to retraining the data and the detection of inter frame video tampering achieved using convolution neural (CN) network model.

Natarajan et al. (2021) [18] introduced a rapid bilateral filter process based smoothen edge preservation. The captured image contains noise that can affect the analysis of better results. This model involved enhancing technique to improve the contrast and brightness of the digital images. For the usage of edge filtering scheme, the quality of both image and video was improved because of extraction and preservation for minimizing the noise. The multi-resolution processing was reduced significantly. To filter the contrast image edges, rapid bilateral filtering technique was involved. Using this method, the edge preservation of the dynamic contrast images were filtered.

Yang et al (2020) [19] implemented a video tampering detection based depth face forgery using convolutional neural network. With the reduction of forgery in the face, deep convolution neural (DCN) network used to analyse the challenges involved in the information security. To obtain better evaluation, this paper involved four different methods for the evaluation of a video based face forgery. They were face location, detection, scaling and interception. The performance of face forgery was detected using two different levels such as video level and frame level. The dataset used for the experimentation purpose using Celeb-DF.

Problem Statement:

Numerous problems with fabricating evidence using altered videos or images are present in modern society. An advanced algorithm is required to spot the altered area in fake images (or) videos in order to reveal the truth. Although numerous strategies have been used, they cannot be used effectively in video tamper detection due to low feature matching rates, lengthy computation times, and poor accuracy rates. The goal of many academics is to extract the best features from the image data. Thus, those techniques are ineffective for using an image comparison to identify a forgeries position. As a result, this research proposed a new hybrid deepnet method for the detection of video data tampering.

III. PROPOSED METHODOLOGY

Recently, video frame tampering has become very popular and received much attention with the development of computing devices. Digital videos are easy to record due to the accessibility of modern digital low-cost video cameras in smart phones and useful in sharing and disseminating visual information. With the emergence of advanced video editing tools, the authenticity of videos may not be considered for granted. Several DL (Deep Learning) models have been introduced to localize and authenticate the video tampering. However, these techniques reported severe challenges such as high computational complexity and low accuracy while managing the huge volume of video data.

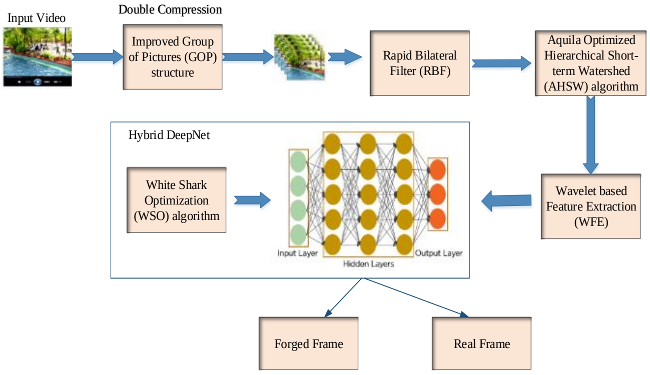

Input Video
Double Compression
Improved Group of Pictures (GOP) structure
Rapid Bilateral Filter (RBF)
Aquila Optimized Hierarchical Short-term Watershed (AHSW) algorithm
Hybrid DeepNet
White Shark Optimization (WSO) algorithm
Input Layer
Hidden Layers
Output Layer
Wavelet based Feature Extraction (WFE)
Forged Frame
Real Frame

Figure 1. Block Diagram of Proposed Hyrid DeepNet Method

Figure 1 represents the block diagram of new hybrid deepnet technique. Each stages of data analysis from the input video to the output detection of forged (or) real frame can be obtained based on the above mentioned block diagram. This work aims to present a novel video tamper detection and forgery localization using Hybrid Adversarial Deep Network (Hybrid DeepNet) to ensure the video authenticity and to identify the tampered region. The steps followed in video tamper detection are Video Data collection, Double Compression Detection (DCD), Noise filtering and Contrast Enhancement, Swarm Optimized Watershed Segmentation, Feature Extraction, Video Tamper detection and forgery Localization. Initially, the forgery video data is collected from different datasets such as VTD, MFC 18 and VIRAT dataset. Next, the detection of double compression is processed using Improved GOP (Group of Pictures) structure. The noise in the video data is removed and contrast is enhanced using Rapid Bilateral Filter (RBF). Next, segmentation is performed using Aquila Optimized Hierarchical Short-term Watershed (AHSW) algorithm. The features are extracted using Wavelet based Feature Extraction (WFE). Finally, the detection and localization of forged region is performed using Hybrid Adversarial Deep Network with White Shark Optimization (Hybrid DeepNet). The loss in the deep network is optimized using WSO (White Shark Optimization) algorithm. The performance is evaluated based on the metrics like precision, recall, accuracy, F-score etc.

A. Improved Group of Pictures (IGOP) Structure

In this research, an improved group of pictures (GOP) method is used to develop the double compression process. The Group of Pictures (GoPs) consist of three frames, such as I-frames, B-frames and P-frames based on a video coding approach. The video frames begin with the I-frame and are followed by the P-frame and the B-frame. With the video compression methods, the GoPs size is fixed as constant. But, this is not suitable for fast moving videos because of two different adjacent frames. It increases the gap between predictive frames (P-frames) and adjacent frames (B-frames). Due to this gap, the intra-coding is not good to provide directly. For dealing these kind of situation, this model involves an improved GoPs method. In this method, high definition H.264 is utilized for the improved GoPs technique. It involves extra one (or) more I-frames in terms of B- and P-frames. The improved GoPs contain large threshold value of current frame than the previous adjacent frame. This technique is suitable to fast changing contents.

Consider the detection process is under fixed GoPs. The size of GoPs in the first compression is k_1 and the second compression is k_2. Here the assumption of fixed GoPs is constant such as k_1 and k_2, respectively. The video tampering can lead to non-aligned double compression $k_2 \neq k_1$. In this paper. It considers the non-aligned double compression. The single compressed video can be denoted with $n-$ frames sequence $S_1, S_2, S_3, \ldots\ldots S_n$ which has decoded to pixel domain based on the GoPs size k_1. The doubly compressed video that the frames are decoded as $N_1, N_2, N_3, \ldots\ldots N_n$ and the GoPs size k_2. Based on the convenient method, the GoPs structure $I\ P\ P\ P\ P\ P\ P\ldots\ldots$ is used in first and second compression in the H.264 baseline with no B-frames. The first and second compression of decoded frames can be denoted as follows:

$$N_i = S_i + t_i, \quad i \in [1, n] \tag{1}$$

Here, t_i denotes the quantization loss. The data that is saved in the stream is denoted as quantized residuals. The p-frames based residuals can be modelled before quantization as $B_j = S_j - u(N_{j-1}, S_j)$, where $u(N_{j-1}, S_j)$ represents motion compensation. The P-frames based index denotes j and the reference value to predict B_j is shown as N_{j-1}. Due to the similarity between adjacent frames, B_j is slightly smaller. Hence, $j = k_1 + 1$

$$B_{k_1+1} = S_{k_1+1} - u(N_{k_1}, S_{k_1+1}) \tag{2}$$

$$B_{k_1+1} = S_{k_1+1} - u(S_{k_1} + t_{k_1}, S_{k_1+1})$$
$$\approx S_{k_1+1} - u(S_{k_1}, S_{k_1+1}) \tag{3}$$

The representation of first compression of I-frame and P-frame is denotes as S_{k_1+1} and S_{k_1}, respectively. To make growing prediction error at zero, I-frames involves a unique inta-coding method. It develops a large split between S_{k_1+1} and S_{k_1}. When S_{k_1+1} uses S_{k_1} as the prediction reference, B_{k_1+1} becomes greater than other factors. Based on the same process, the residual value will be abnormally increased. When the factor $j = a \times k_1 + 1, \quad a = 2,3,4,..$ then the abnormal condition is denoted as relocated I-frame. Using this method, the detection of double compression is achieved.

B. Rapid Bilateral Filter (RBF)

The RBF [20] is used in pre-processing stage which applied to the video for reducing noise and improve contrast. Based on the multi-resolution images with strong dynamic contrast, the RBF filters the edges in a shorter amount of time. In the first step, the input data is filtered on the sharp edge position of the data. Using the rapid bilateral filtering technique, the exact position of pixel is extracted. Therefore, the RBF technique calculates the intensity values for weighted standard distance. To observe the image noise level, both base and upcoming layer's intensity values are calculated. For performing the sharp localization based edge filtering, the detection of multi-resolution high contrast data are obtained.

The high contrast frame is measured using the following formula:

$$Contrast\ Frame = pixel\,(P_{(a,b)}) \tag{4}$$

From the equation 4, P denotes the contrast frame of the pixel intensity value. The pixel point is represented using longitude and latitude axis of pixel point x and nearer pixel point y .

C. Aquila Optimized Hierarchical Short-term Watershed (AHSW) algorithm for feature selection

AHSW [21] algorithm is used to segment the data which pre-processed by rapid bilateral filtering algorithm. Based on equation 4, position of search agents are gets updated.

$$A_3(s+1) = (A_{best}(s) - A_D(s)) \times \beta - Rand + ((U_b - L_b) \times Rand + L_b) \times \alpha \tag{5}$$

Where, the output of the next iteration of s is termed as $A_3(s+1)$, the exact position of prey is $A_{best}(s)$, the current solution based mean value is $A_D(s)$. In the proposed work, the Aquila optimisation is hybridised with Hierarchical short-term watershed algorithm for selecting the best features.

1) Hierarchical short-term watershed algorithm

The segmentation process of the video is processed sequentially one by one of the frame. To identify the boundaries of separate locations, the AHSW algorithm is used. Initially, an image $t(a,b)$ with the regional maxima (R_m) is the pixel values. The co-ordinate value with regional maxima is denoted as $N(R_m)$ and the co-ordinate set is $G[c]$.

$$N_c(R_m) = N(R_m) \cap G[c] \tag{6}$$

Consider, $N_c(R_m) = 1$, the value denotes the location of (a,b) .

If $N_c(R_m) = 0$, it shows a flooded catchment basin and is represented as follows:

$$N[c] = \cup_1^R N_c(R_m) \tag{7}$$

If, the connected components are more than one then it shows the separation of ridges in two (or) more catchment basin. Therefore, until the value of $c = \min - 1$, the steps are repeated to obtain the points of ridges. This operation is to achieve the temporal connection of frames. The AHSW algorithm is termed as shift invariant in two (or) more dimensions and directionally selective. Using the un-decimated wavelet transforms, it achieves the redundancy based on the factor of $2d$ for d-dimensions. The loading process of consecutive frames with three levels of dual tree complex wavelet (DTCW) transform is obtained using high and low intensity pixels between the frames. Finally, the segmentation of the video is efficiently obtained.

D. Wavelet Based Feature Extraction (WFE)

After segmentation, Wavelet Based Feature Extraction is used to extract the data. For the identification of detailed information, wavelet coefficients are featured for the extraction process. Therefore, normalized Shannon entropy, standard deviation, energy and mean are the local features is proposed in the dual tree complex wavelet (DTCW) transform coefficients in order to differentiate out and in focus locations of the video (or) image. To calculate the local features, a local window 3×3 is involved at six directions. The calculation is expressed in the following equation:

$$M_{mean}(p,q) = \frac{1}{9} \sum_{x=-1}^{1} \sum_{y=-1}^{1} \left| a(p+x, q+y) \right| \tag{8}$$

$$E_{energy}(p,q) = \frac{1}{9} \sum_{x=-1}^{1} \sum_{y=-1}^{1} \left| a(p+x, q+y) \right|^2 \tag{9}$$

$$SD_{s\tan d.devi}(p,q) = \left(\frac{1}{9} \sum_{x=-1}^{1} \sum_{y=-1}^{1} \left(\left| a(p+x, q+y) \right| - M_{mean}(p,q) \right)^2 \right)^{\frac{1}{2}} \tag{10}$$

$$SH_{shannon-entr}(p,q) = \sum_{i=0}^{255} j^i(p,q) \times \log_2 \left(\frac{1}{j^i(p,q)} \right) \tag{11}$$

Here, the DTCW transform coefficients with b orientation is denoted as $a(p,q)$ that is calculated as

$$a_b(p,q) = \left(a_{\mathrm{Re},b}(p,q)^2 + a_{im,b}(p,q)^2 \right)^{\frac{1}{2}} \tag{12}$$

Here, the absolute values of DTCW transform coefficients in the i^{th} value of normalized histogram with local window (3×3) are shown as $j^i(p,q)$. In the extraction process, the four feature vectors are derived at h scale. Two source images C, D with b directional sub-band based different focus points is denoted as $LF_{X,b}^h$ and $LF_{Y,b}^h$.

High frequency coefficients based feature vectors are evaluated using above equation. Using the high frequency coefficients for down sampled decision matrices, the low frequency fusion coefficients are obtained.

E. Hybrid Adversarial Deep Network with White Shark Optimization (Hybrid DeepNet)

The hybrid model contains two principal streams. The first model focuses on an auto-encoder to emphasize normal features at patch level. In the second stream, it performs a classification based on the extracted characteristics in the former stream. Considering, a 3-dimensional cuboid of frame concatenation with small spatial resolution is resized to 160×120 with gray scale. The input model is a cuboid size $10\times10\times3$. Hyper parameters of the classifier is given by; Fmax (% Maximum frequency of the wavy motion) is 0.75, Fmin (% Minimum frequency of the wavy motion) is 0.07, tau of 4.11 with mu of 2/abs(2-tau-sqrt(tau^2-4*tau)), pmin of 0.5, pmax of 1.5, a0 of 6.250, a1 and a2 of 100 and 0.0005, respectively.

1) Convolutional auto-encoder based common features

The auto-encoder based hybrid network is to study local features. It operates as a reconstruction method related to a bottleneck framework. By concatenating consecutive frames, the embedded process is learned by temporal factor. Few analysis employed optical way of representation in terms of data, which may cause noisy outputs and may require a training stage like Flow-net. Based on the process of encoder, the feature maps are reduced with a stack of layer blocks that contains two convolutional (Convl) layers such as activation and normalization. In the first convl layer, the spatial dimensions of the input is converted. In the second convl layer, the specific number of channels are obtained due to the transformation of features. The first convl layer is an alternative for the pooling process.

2) Sub-network Classification

By using the auto encoder process, common local features can be learned. In addition to the local features, a constraint is added to the encoder to extract the data to reveal spatial positions. Therefore, the encoder output blocks are vectorized and the elements of 7104 feature vector are concatenated. In order to reduce the channel number except latent variables, the convolution of 1×1 filters is applied before the process of vectorization. The partition of each frame size 160×120 is non-overlapped with 10×10 patches. Hence, a total of an input cuboid has attained 16×12=192 possible locations. Due to 192 classification problems, a small cuboid is complicated. For the simplification of this problem, every patch location is denoted in terms of its spatial dimensions. Hence, the performance of classification model is of two branches of image patches such as vertical and horizontal indices. By using this way, the total class number is reduced to 16+12=28. This method helps to learn common features of cuboids for both in the same row and column.

3) White Shark (WS) Optimizer

This White Shark (WS) Optimizer is used to remove network loss and enhance the performance of classification stage. The WS optimizer is a population based method for randomly generating an initial solutions. Based on the initialization process, this technique is helpful for solving an optimization problem. Population size is denoted as n white sharks and each white shark's position is described as $2d$ matrix.

$$s=\begin{bmatrix} s_1^1 & s_2^1 & \dots & \dots & s_d^1 \\ s_1^2 & s_2^2 & \dots & \dots & s_d^2 \\ & & \vdots & & \\ & & \vdots & & \\ s_1^n & s_2^n & \dots & \dots & s_d^n \end{bmatrix} \tag{13}$$

Where, the location of search space for all white shark stands as s, the decision variable number of a problem denotes as d and the d^{th} dimension of the location for e^{th} white shark denotes as s_d^e.

Initial population based on a uniform random initialization in the search domain is given as follows:

$$s_p^e = L_p + t\times(U_p - L_p) \tag{14}$$

Here, the e^{th} white shark in the p^{th} dimension for the initial vector represents s_p^e. The upper bound and lower bound are represented as U_p and L_p, respectively. t shows a random number in the range of $[0,1]$. The white sharks can search the exact prey in order to identify the location. Commonly, the white sharks locations are changing based on the time limit. Using the movement of waves, it can identify the exact position of prey. Using equation 14, the optimal parameters are fine tuned. The equation is given by:

$$s_{v+1}^e = \begin{cases} s_v^e.\neg\oplus s_o + U_p..x + L_p.y; & Rand < m_f \\ s_v^e + k_v^e/l; & Rand \geq m_f \end{cases} \tag{15}$$

Where, the new position vector for e^{th} white shark is shown as s_{v+1}^e, a negation operator is $\neg$, one-dimensional binary vectors are x and y, respectively. The search space of both lower and upper limits is denoted with the factors of L_p and U_p, respectively. A logical vector is shown as s_0. l denotes the frequency of the wavy motion. $Rand$ shows a random number. The movement force increases related to the number of iterations is denoted as m_f .

4) Adversarial training

In order to reduce losses, an additional discriminative technique is added to improve the quality of redeveloped cuboids using a generative adversarial (GA) network. It consists of a generator G_r and a discriminator D_r. The GA network consists of binary classifier in order to differentiate the data based on training set with real patterns. In this model, two GA network components are involved to obtain better results. G_r value tries to dominate D_r values to generate outputs that are related to training samples. Then, the D_r classifies this outputs as fake information. This model is useful to achieve better performance for image translation and video frame prediction. The equation is denoted as follows:

$$K_{g_r}(d,N) = -\alpha_{g_r} \log C(N(d)) + \alpha_S K_S(d,N) + \alpha_d K_d(d) \tag{16}$$

$$K_C(d,N) = -\frac{1}{2}\log C(d) - \frac{1}{2}\log[1 - C(N(d))] \tag{17}$$

Here, input cuboids with training data is denoted as d , the auto-encoder is shown as N , the classification, adversarial and reconstruction are shown as α_{g_r} , α_S and α_d , respectively. Using the optimization process, the two objective functions are denoted as K_{g_r} and K_C . During the optimization operation, the discriminator value D_r is employed but it has no role in the detection process.*5) Frame level detection*

After the completion of optimization process, the hybrid network is to provide softmax results $S_d(c)$ for the classification of spatial position as well as to reconstruct an input cuboid c through the auto-encoder. The location of ground truth c is shown as (c_a, c_b) and the three normality scores of a cuboid are denoted as $F_R(c)$, $F_a(c)$ and $F_b(c)$.

$$F_R(c) = \max\left(\left\|c - \lambda(c)\right|^{\beta}\right); \quad F_{d\in\{a,b\}}(c) = mean\left(\left|M(c_d) - S_d(c)\right|^{\alpha}\right) \tag{18}$$

Where, the conversion of an input label to a one-hot vector is termed as $M(.)$, the output functions of average and maximum values are represented as $\max(.)$ and $mean\ (.)$. There are three types of cuboid level scores that are obtained using 3 score map size 16×12 for each frame. One of the frame is a concatenation and the other two frames are the next consecutive frames. In order to combine the three maps to obtain an improved output, the weighted sum of final normality score is mentioned in the below equation:

$$F_{R,a,b}(c) = \sum_{k\in\{R,a,b\}} \left[1 - \frac{1}{\|Q\|} \sum_{a\in Q} F_k(a)\right] F_k(c) \tag{19}$$

Where, $\left[1 - \frac{1}{\|Q\|} \sum_{a\in Q} F_k(a)\right]$ denotes the weighted sum, $F_k(c)$ shows cuboid c estimated scores, Q represents cuboid collection of normal events in the training set at the spatial location (c_a, c_b). According to the training data of its efficiency, each score type of weighted cuboid is assigned at a specific location. Consider, F_k should be small value. Equation (19) is helpful to implement at frame level.

For the measurement of frame level normality, the standard deviation is termed as $f(s)$ with cuboid level scores $F_{R,a,b}(c)$ attained from same cuboid frame. $F_{a,b}$ is the combination of F_a and F_b. Therefore, the weight of $F_R(c)$ is considered to 0 for calculating the value $F_{a,b}$.

The frame level score $f(s)$ is normalized in terms of each evaluated video with the frames n as

$$\hat{f}(s_i) = \frac{f(s_i)}{\max[f(s_1),\ldots, f(s_n)]}, \quad 1 \le i \le n \tag{20}$$

The scores of the frames are high when anomalous events are high otherwise it is low.

IV. RESULTS AND DISCUSSION

In this section, the results and discussion part are mainly divided into four parts such as simulation environment, dataset description, performance metrics and comparison analysis. In terms of accuracy, precision, recall, F1-score, EER, AUC are used to compare with existing models like SIF, MPS, MST, SRCS, DNG-PSO etc. The evaluation are shown in following sections. The hyper parameters is given by; numEpochs is 500, miniBatchSize is 128, learnRate is 0.0002 with the dropout of 0.5. The simulation is achieved using the MATLAB tool.

A. Simulation environment

The proposed hybrid deepnet technique has been simulated using MATLAB R2015a tool, 6GB RAM and 2.00 GHZ intel core 2 dual processor. The hybrid deepnet model consists of three dataset such as VTD, VIRAT and MFC datasets. For using this MATLAB tool, the input video is simulated in terms of different datasets to identify the data whether the video is forged (or) real.

B. Dataset Description

The analysis of each recognition process mainly depends on training, evaluation and testing methods. In order to obtain these operations, the dataset is essential to analyse the video information. In this model, three different types of datasets are used for analyse the video data such as Media forensics challenge (MFC) 18, Video image retrieval analysis (VIRAT) and Video tampering dataset (VTD) dataset. The videos are gathered from youtube to evaluate the forgery detection. In the MFC 18 dataset process, this involves 1267 videos in terms of validation (15%), testing (15%) and training (70%). Based on the VIRAT dataset process, the ground portion of data is gathered about 25 hours using stationary cameras. The data consists of 16 scenes. The VTD dataset of 10 videos are considered for forgery detection. Using 10 videos, 4490 frames are attained along with training (70%), validation (15%) and testing (15%).

C. Performance Metrics

The performance of the new method is evaluated using the following values.

1) Receiver Operating Characteristics (ROC)

The ROC value is obtained using two different rates like false positive and true positive based on different threshold limits. The false positive value can be termed as probability (or) specificity. The true positive value is denotes as sensitivity (or) recall.

2) Equal Error Rate (EER)

In the EER process, the proposed hybrid deepnet is considered for both frame level and pixel level values in terms of threshold level of both false acceptance and false rejection rate.

3) Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (21)$$

Here, the accuracy value is denoted as ACC, the true positive value is termed as TP, the true negative value is TN, the false positive and false negative are represented as FP and FN, respectively.

4) Precision

$$PR = \frac{TP}{FP + TP} \qquad (22)$$

5) Recall

$$R_{ec} = \frac{TP}{FN + TP} \qquad (23)$$

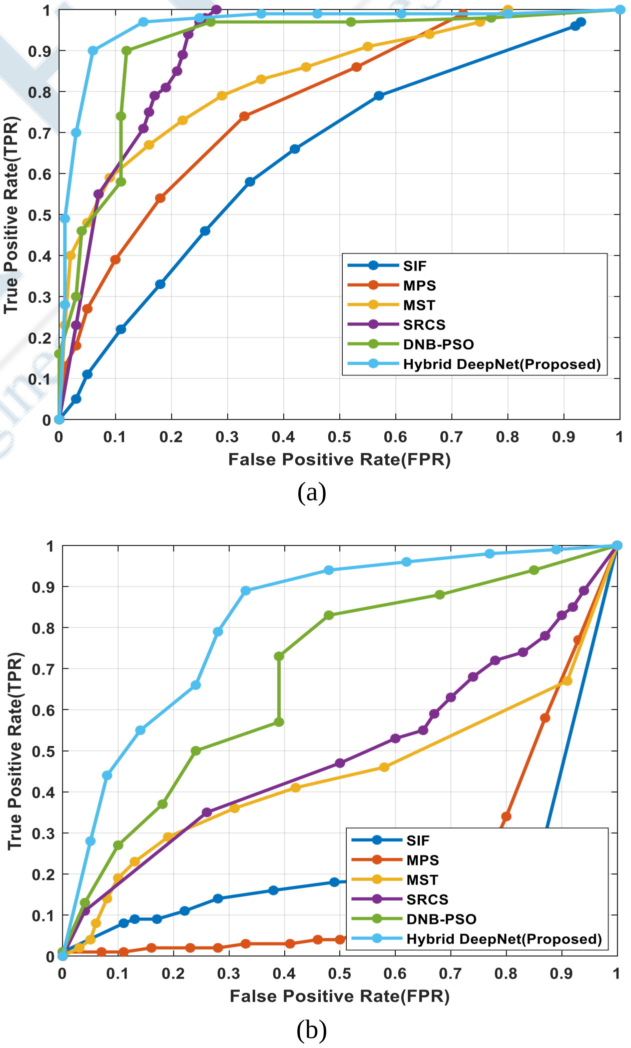
6) F1-Score

$$F1 = 2 \times \frac{PR \times R_{ec}}{PR + R_{ec}} \qquad (24)$$

D. Comparison Analysis

The new proposed Hybrid DeepNet is related to other existing techniques such as SIF, MPS, MST, SRCS and DNS-PSO algorithms. In the analysis process, the results of Frame level ROC curve, Pixel level ROC curve, accuracy, precision, recall, F-score, EER, AUC (Area under the curve) and EDC (Equal detected rate).

In the Figure 3 (a), the results of frame level ROC curve of proposed hybrid DeepNet compared with other existing models [22]. The propose model has attained better performance due to the minimization of loss in the false positive rate and true positive rate. This is happened because of WSO algorithm based deep learning process to detect the forged region effectively. All the other models are not achieved better results than the new hybrid deepnet model.



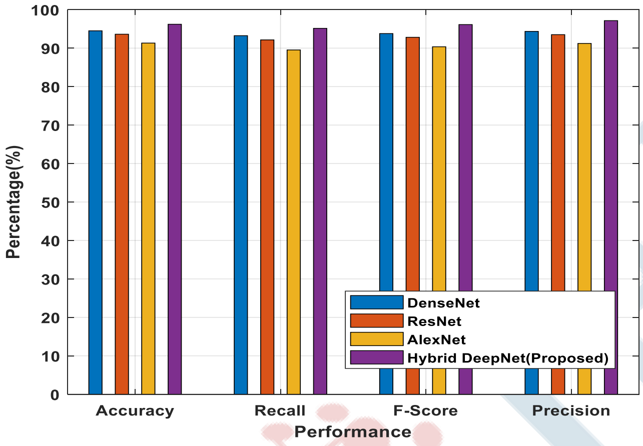
(a)

(b)

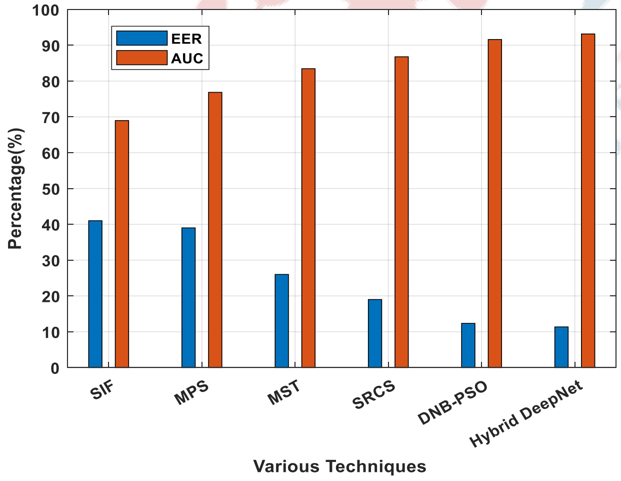
Figure 2. (a) Frame level ROC curve, (b) Pixel level ROC curve

From the figure 2(b), the results of pixel level ROC curve are shown for the proposed compared to other existing models. In the pixel level ROC values, the hybrid deepnet has attained better results than the other models [22]. Because of the improve GOP structure, the double compression helps to enhance the output of hybrid deepnet in a proper way.

Figure 3(a) shows the comparison based on the several technique for the proposed model related to other Densenet, Resnet, Alexnet. The accuracy of the hybrid deepnet shows the values of 96.21% rate. This new technique has enhanced the performance for the values of 95.15%, 96.13% and 97.15% for Recall, F-score and Precision, respectively. The existing methods has shown low performance than the new method. The accuracy rate of existing model has reached the values of 91.32%, 93.62% and 94.51% for Alexnet, Resnet and Densenet, respectively.



(a)



(b)

Figure 3. (a) Performance analysis of various models, (b) Performance of EER and AUC values

Figure 3(b) shows the performance of frame level based EER and AUC values. The above graph clearly explains that the EER rate and the AUC value are obtained better results than the existing models. The EER rate of the hybrid deepnet has attained about 11.35% and the AUC value of new model has reached as 93.15%. By using the wavelet based feature extraction, the performance is improved when compared to other models.

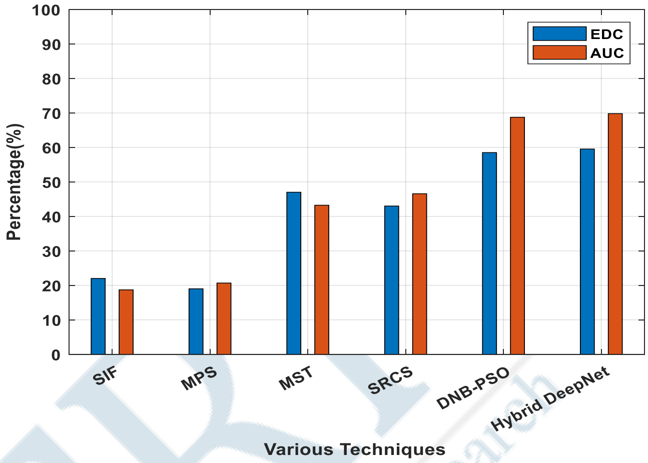


Figure 4: Performance of EDC and AUC values

Figure 4 represents the performance of EDC and AUC values based on pixel level method. Due to the deep network analysis, the forged region of the frame (or) pixel is exactly identified and the noise of the video information is minimized using rapid bilateral filter method. The new proposed Hybrid DeepNet has achieved the values of 59.53% and 69.79% for EDC and AUC, respectively. The existing method like DNB-PSO technique attained the values of 58.49% and 68.74% for EDC and AUC, respectively. Therefore, the new model has reached better results than the other existing methods in terms of pixel level performance.

V. CONCLUSIONS

In this paper, the proposed Hybrid DeepNet is involved for the identification of tamper detection in video. Several existing model involves convolutional neural CN (network) to detect the forged video. But, the effectiveness of values like accuracy, precision, recall and ROC are not achieved. This model is well suited than the other techniques like DenseNet, ResNet and AlexNet. Therefore, the hybrid deepnet is applied for the detection and localization of tampered location of video and the loss is minimized using WSO algorithm. The results are evaluated using MATLAB tool. By using this new hybrid deepnet model, the outputs have obtained in terms of 96.21%, 95.15%, 96.13% and 97.15% for accuracy, recall, F-score and precision, respectively. From the experimental outputs, the new method is more powerful than the existing methods. Using hybrid deep learning method, the detection and localization of the tamper video are achieved. In order to spot tampering in the future, this research wants to contribute 3D videos as well as criminal videos. The efficiency of the proposed model would be significantly higher, if the implementation of feature selection in this research.

Acknowledgments:

None.

REFERENCES

[1] W. Wei, X. Fan, H. Song & H. Wang, “Video tamper detection based on multi-scale mutual information”. Multimedia Tools and Applications, vol. 78, no. 19, pp. 27109-27126, 2019.

[2] M.E. Purwitasari, "The Process of Making Motion Video Tutorials Graphic by Using Adobe Premiere Cc 2015." IJOTECH: International Journal of Science and technology vol. 1, no. 2, pp. 102-106, 2023.

[3] D. Myvizhi and J.M.J. Pamila, "Extensive analysis of deep learning-based deepfake video detection." Journal of Ubiquitous Computing and Communication Technologies vol. 4, no. 1, pp. 1-8, 2022.

[4] H. Wu, Y. Zhou and Z. Wen, "Video tamper detection based on convolutional neural network and perceptual hashing learning." In Computer Graphics International Conference, Springer, Cham, pp. 107-118, 2019.

[5] Y. Al Balushi, H. Shaker and B. Kumar, "The Use of Machine Learning in Digital Forensics." In 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022), Atlantis Press, pp. 96-113, 2023.

[6] J. Hu, X. Liao, W. Wang and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network." IEEE Transactions on Circuits and Systems for Video Technology vol. 32, no. 3, pp. 1089-1102, 2021.

[7] M. Raveendra and K. Nagireddy, "DNN based moth search optimization for video forgery detection." International Journal of Engineering and Advanced Technology vol. 9, no. 1, pp. 1190-1199, 2019.

[8] P. Johnston and E. Elyan, "A review of digital video tampering: From simple editing to full synthesis." Digital Investigation vol. 29, pp. 67-81, 2019.

[9] M. Palioura and C. Dimoulas, "Digital Storytelling in Education: A Transmedia Integration Approach for the Non-Developers." Education Sciences vol. 12, no. 8, pp. 559, 2022.

[10] M. Zampoglou, F. Markatopoulou, G. Mercier, D. Touska, E. Apostolidis, S. Papadopoulos, R. Cozien, I. Patras, V. Mezaris and I. Kompatsiaris, "Detecting tampered videos with multimedia forensics and deep learning." In International Conference on Multimedia Modeling, Springer, Cham, pp. 374-386, 2019.

[11] H. Wu, P. Wang, X. Wang, J. Xiang and R. Gong, "GGViT: Multistream Vision Transformer Network in Face2Face Facial Reenactment Detection." In 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, pp. 2335-2341, 2022.

[12] S. Tariq, S. Lee and S. Woo, "One detector to rule them all: Towards a general deepfake attack detection framework." In Proceedings of the web conference, pp. 3625-3637. 2021.

[13] X. Liu, W. Lu, W. Liu, S. Luo, Y. Liang and M. Li. "Image deblocking detection based on a convolutional neural network." IEEE Access vol. 7, pp. 26432-26439, 2019.

[14] P. Yang, D. Baracchi, M. Iuliani, D. Shullani, R. Ni, Y. Zhao and A. Piva, "Efficient video integrity analysis through container characterization." IEEE Journal of Selected Topics in Signal Processing vol. 14, no. 5, pp. 947-954, 2020.

[15] S.W. Oh, J.-Y. Lee, N. Xu and S.J. Kim, "Video object segmentation using space-time memory networks." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9226-9235, 2019.

[16] M. Raveendra and K. Nagireddy, "Tamper video detection and localization using an adaptive segmentation and deep network technique." Journal of Visual Communication and Image Representation vol. 82, pp. 103401, 2022.

[17] M. Raveendra and K. Nagireddy, “Inter frame detection based on DWT-DCT Markov features and Fine-tuned AlexNet Model.” IJCSNS International journal of computer science and network security, vol. 20, no. 12, Dec 2020.

[18] B. Natarajan, "Smoothen Edge Preservation using Rapid Bilateral Filter Process." Annals of the Romanian Society for Cell Biology pp. 11585-11590, 2021.

[19] T. Yang, J. Wu, L. Liu, X. Chang, and G. Feng, "VTD-Net: depth face forgery oriented video tampering detection based on convolutional neural network." In 2020 39th chinese control conference (CCC), IEEE, pp. 7247-7251, 2020.

[20] B. Natarajan and P. Krishnan, "Contrast Enhancement Based Image Detection Using Edge Preserved Key Pixel Point Filtering." Comput. Syst. Sci. Eng. vol. 42, no. 2, pp. 423-438, 2022.

[21] L. Abualigah, D. Yousri, M.A. Elaziz, A.A. Ewees, M.A.A. Al-Qaness and A.H. Gandomi, "Aquila optimizer: a novel meta-heuristic optimization algorithm." Computers & Industrial Engineering vol. 157, pp. 107250, 2021.

[22] D.R. Kishore, D. Suneetha, G.S.P. Ghantasala and B.R. Sankar. "Anomaly Detection in Real-Time Videos Using Match Subspace System and Deep Belief Networks." Multimedia Computing Systems and Virtual Reality, vol. 151.