ISSN: 0067-2904

Enhancement Digital Forensic Approach for Inter-Frame Video Forgery Detection Using a Deep Learning Technique

Mohammed R. Oraibi *, Abdulkareem M. Radhi
Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

Received: 26/5/2021 Accepted: 22/8/2021 Published: 30/6/2022

Abstract

The digital world has been witnessing a fast progress in technology, which led to an enormous increase in using digital devices, such as cell phones, laptops, and digital cameras. Thus, photographs and videos function as the primary sources of legal proof in courtrooms concerning any incident or crime. It has become important to prove the trustworthiness of digital multimedia. Inter-frame video forgery one of common types of video manipulation performed in temporal domain. It deals with inter-frame video forgery detection that involves frame deletion, insertion, duplication, and shuffling. Deep Learning (DL) techniques have been proven effective in analysis and processing of visual media. Dealing with video data needs to handle the third dimension (the time dimension), which means extracting temporal features as well as spatial features. The proposed model is built based on the Three Dimension Convolution Neural Network (3D-CNN). Through pre-processing operation that introduced difference frames that pick up the difference in successive adjacent frames, which provide a large quantity of temporal information and lead to enhance the effectiveness of the proposed model. The model achieves high accuracy of 99%.

Keywords: Digital video, Deep learning, Video forensics, Inter-frame video forgery, Classification.

تعزيز نهج الطب الشرعي الرقمي لاكتشاف تزوير الفيديو بين الإطارات باستخدام تقنية التعلم العميق

محمد راضي عريبي*, عبدالكريم مرهج راضي
قسم الحاسبات، كلية العلوم، جامعة النهرين، بغداد، العراق

الخلاصة

يشهد العالم الرقمي تقدمًا سريعًا للتكنولوجيا ، مما أدى إلى زيادة هائلة في استخدام الأجهزة الرقمية ، مثل الهواتف المحمولة وأجهزة الكمبيوتر المحمولة والكاميرات الرقمية .حيث تعتبر الصور الفوتوغرافية ومقاطع الفيديو مصدر أساسي للإثبات القانوني في قاعات المحاكم فيما يتعلق بأي حادث أو جريمة. لذا أصبح من المهم إثبات مصداقية الوسائط المتعددة الرقمية. تزوير الفيديو بين الإطارات أحد أنواع الهجمات الشائعة للتلاعب بالفيديو الذي يتم تنفيذه في المجال الزمني. تتضمن عملية اكتشاف تزييف الفيديو بين الإطارات التعامل مع الأنواع التالية حذف الإطار ، والإدخال ، والنسخ ، والخلط. أثبتت تقنيات التعلم العميق فعاليتها في

*Email: contcreatahyzt@gmail.com

تحليل ومعالجة الوسائط المرئية. التعامل مع بيانات الفيديو يحتاج إلى معالجة البعد الثالث (البعد الزمني) ، وهو ما يعني استخراج السمات الزمنية وكذلك السمات المكانية. يعتمد النموذج المقترح على الشبكة العصبية الملتفة ثلاثية الأبعاد (D–CNN3). من خلال عملية المعالجة المسبقة ، أدخلت إطارات الفرق التي تلتقط الفرق في الإطارات المجاورة المتتالية ، والتي توفر كمية كبيرة من المعلومات الزمنية وتؤدي إلى تعزيز فعالية النموذج المقترح. النموذج المقترح حقق دقة عالية بلغت 99٪.

1. Introduction

Digital multimedia management has been easier recently because of the availability of powerful computers, advanced editing software products, and new multimedia capture equipment. This led to making every normal person's daily life includes the exchange and distribution of large quantities of digital media, in particular digital images and videos. In several ways, photographs and video clips act as the main sources of legal proof in courtrooms concerning any incident or crime. Manipulation of these types of multimedia has, however, become an exceedingly simple job for even a layman with little cost. due to the Simplicity of utilizing image and video processing program and desktop equipment. Therefore, before being viewed as evidence in a courtroom, it has become increasingly important to authenticate and prove the trustworthiness of digital multimedia by employing video forensic techniques [1].

Video forensics investigates, compares, or analyzes videos in scientific aspects. Such operations must also be technically accurate, because the data will be presented to a prosecutor in most cases. Video sequences are often assumed to be better forensic evidence than still images. Surveillance video thus is regarded as valuable evidence [2].

Performs the attacks of video manipulation,In temporal and Spatial-temporal domains. Splicing and modification of region and copy-paste usually occur within the spatial and spatial-temporal domain. Frame deletion, insertion, duplication, and shuffling occur in the temporal domain. However, these attacks cannot produce satisfactory results, because videos can involve complex scenarios such as continuously moving objects or compression-related noise. Video can be like an image series called frames [3].

Video Forgery Detection (VFD) has two primary approaches: Active approach and passive approach, illustrated in Figure 1. Active forgery detection involves techniques such as digital watermarking and digital signatures that confirm authentic content ownership and copyright infringements. The techniques for passive forgery detection are viewed as an advanced digital safety path [4].

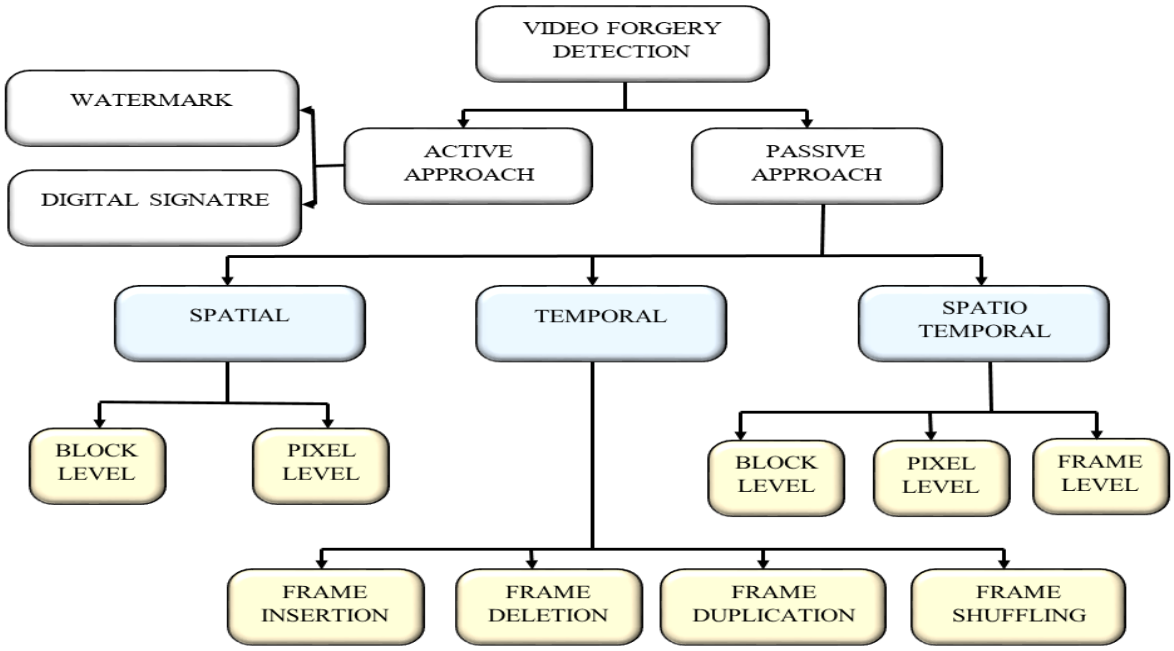

Figure 1- VFD Techniques [5].

The passive approach works in contrast to the active approach where it deals with specialized hardware without restriction and does not require any details on the content of the video. It is often referred to as a passive blind method. The core principle of this approach is that the videos have some intrinsic characteristics or attributes that are compatible with the original videos that will suffer distortion when the video is compromised. These features are extracted from a video by passive approaches and evaluated for various forgery operations detection purposes [6].

DL techniques prove effective in analysis and processing of visual media. They can extract and learn complex features compared to shallow methods [7]. DL methods do not require the extraction and representation of handcrafted features [8]. This allows for benefitting from increasing computation power and data without the intervention of domain experts [9]. During the training in DL, the extraction and classification of features are combined in an end-to-end [10].

2. Literature Review

In the field of digital multimedia forensics, much substantial research progress has been made. This includes a series of research into video forgery. Many works have focused mainly on VFD frame insertion, removing, shuffling, and duplication. Presented in this section is a concise review of relevant research to identify the mentioned video forgeries.

[11] Proposed forensic technique allowed the detection of inter-frame forgery, in H.264 and Moving Picture Experts Group (MPEG-2) encoded videos. This implemented an objectivity approach for the automated detection of position and the manipulation by using optical flow and the residual gradient. Inter-frame forgery can be diagnosed with 90% by the proposed technique. Even when alternative bit rates are employed to record and reconstruct a video, the detection results of this technique cannot be affected. This technique tends to suffer from the loss of performance in videos with extremely slow motion.

[12] developed a forgery detection method based on chromatic moments of the opposing Zernike. The falsification characteristic analysis is based on the matching coarse to fine models. Careful identification is done first for the extraction of abnormal points by transforming each frame into Two Dimensions (2Ds) chromaticity space from Three

Dimensions (3Ds) Red, Green, Blue (RGB) color, along with the correlation with the momentary Zernike. Experimental results show a greater precision in this approach of 97.5% in detection types (copy-move, insert, delete, replace).

In [13] A Multi-Level subtraction (MLS) method for video frame insertion forgery detection with a 93.92% recall rate on a forensically realistic video database has been proposed by the authors.

[14] Authors extracted video stream residue data from each frame. Spatial and temporal energy was then used to demonstrate the data flow, and the manipulated frames were detected by abnormal points. For the distinction of insertion from duplication attacks, noise ratios of forged and original frames were calculated. Anomalies are not captured by the approach suggested, and identification fails, when some frames are removed from the static scene.

[15] The authors proposed an algorithm that comprises feature extraction and localize an abnormal point. It extracts the 2D phase congruence of each frame during extraction because it is a good characteristic of each frame. The correlation between consecutive frames is computed. The abnormal points were identified with k-means clustering algorithms in the second level. The result two categories divided up normal and abnormal points.

In [16] An Optical Flow (OF) and stable parameter coarse-to-fine detection strategy proposed. To find suspected forgery points, coarse diagnosis specifically analyses OF sum consistency. Fine detection is then performed for the exact location of the forgery, including duplicate frame pairs that fit according to OF correlation, and validation test to further minimize the false detection. Experimental results show that the suggested technique for detection provides excellent precision in several popular attacks with minimal computational complexity and strong applicability.

In [17] A Deep Convolution Neural Network (DCNN) approach for the detection of forged content in videos was presented. A classifier categorizing the frames as authentic and forged based on spatial or temporal interrelationships was supplied with the prepared data collection. When compressed videos were processed using YouTube, the proposed algorithm averaged 98% accuracy.

[18] Recommended to re-trained by using spatial-temporal relationships in a video to detect inter-frame forgery based on CNN models. They use the confidence score instead of the raw network performance score to prevent errors because of the network. In this study, the suggested technique has shown that it is considerably more precise than the recent methods on the same data set and has reached 99.17% of accuracy.

3. Methodology

This part is dedicated to the basic background of VFD, its types, and the detection techniques that have been used previously. A Digital Video (DV) comprises a digital, rather than an analog, signal-based electronic recording. It is used to construct an image sequence people can easily understand and analyzed with computer algorithms. The key areas of DV appliances include movie making, news reporting, tracking systems, and acceptable court evidence [19].

Video tampering is a video forensics subcategory that investigates the Video to locate spatial or temporal locations of forgery and for content alterations detection. VFD aims to determine video authenticity and to discover the potentially manipulated and fabricated video. The digital forensic solutions for forgery authentication and validation divided into two methods, active and passive. The Active method suggests a forgery detection method in which it keeps some data side by side at the source (camera). Digital watermarking or digital signature can be this information. Passive forgery detection methods consider as progressing in digital security. It operates with no limitations on special hardware or requires any details from the first hand on the content of the video [20]. If a video is forged, its fundamental properties change and these changes are observed by those techniques used to detect video forgery [21].

3.1 Video Tampering Attack

Researchers have proposed various methods and algorithms to identify digital video tampering based on different features and the boundary of their occurrence. This paper will explore the passive VFD methods, based on the type of forgery they address (Temporal tampering method).

Spatial Tampering Attack: Changes are made to the frame material (x-y axis) that shows visual information on the video. Cropping and replacement, painting, alteration, object addition, and deletion are spatial manipulation operations. Spatial manipulation at the block or pixel level can be performed [22].

Temporal Tampering Attack: It is executed with the frame sequence. Concentrated on temporal dependency. These attacks primarily affect the time sequence of the visual information collected by the system. Common attacks are frame insertion, frame deletion, frame duplication, frame shuffling [23].

Spatio-Temporal Tampering Attack: It is a mixture of temporal and spatial manipulation. The frames and the visual contents in the same video will be changed. Here you can find intra-frame manipulation and inter-frame manipulation combinations [23].

Video tampering attacks represented in Figure 2, F_i and P_{ij} stand for the i^{th} frame and pixel intensity respectively. Height and width are established by x and y [24].

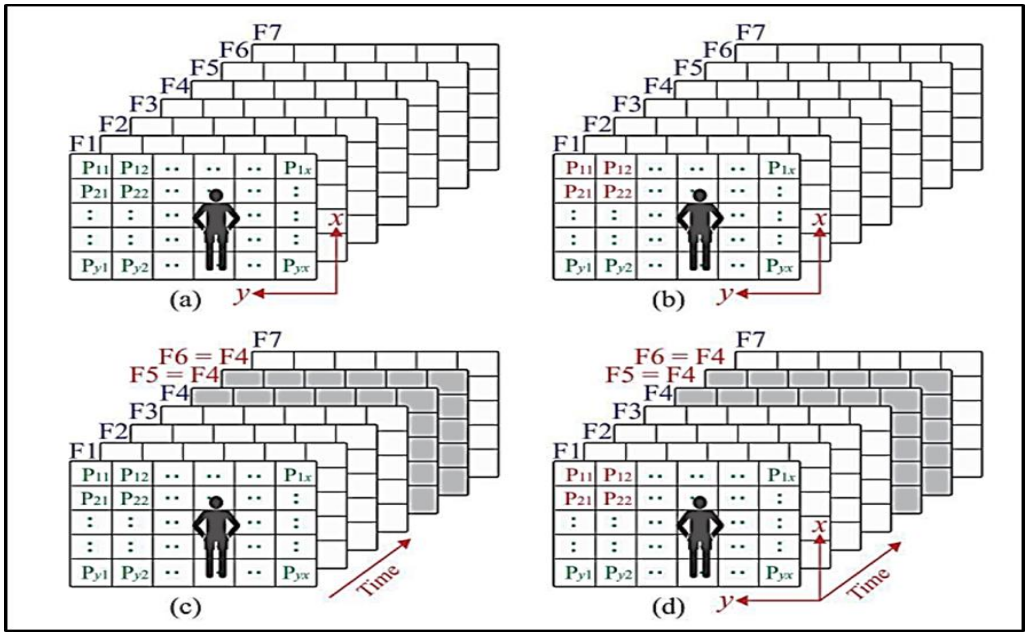

Figure 2- Video tampering attacks, where (a) Represent original video, (b) Spatially tampered video, (c) Temporally tampered video, (d) Spatio-temporal tampered video [24].

3.2 Inter-Frame Video Forgery Types

The Inter-frame video forgery types are as follow [25]:

Frame Insertion**:** frames from another video or the same video sequence are inserted into the original video sequence.

Frame Duplication**:** Any frames of the original video sequence are copied from a temporary location and pasted to a temporal location in the same video.

Frame Deletion**:** Several frames from the original video sequence are removed. The cumulative number of video frames decreases relative to the number of original videos.

Frame Shuffling: Is another type of frame duplication in which the copied frames are temporarily reordered before insertion.

3.3 DL for Computer Vision

DL or hierarchical learning or deep structured learning is a sub-area of Machine Learning (ML) that deals with algorithms that have inspiration by structure and function of the brain called Artificial Neural Network (ANN). NN is a technique of ML, similar and inspired by the human brain and nerve system. It consists of three major layers: first one is input, the second is hidden and at last output-layers arranged processing units. Every layer has units or nodes linked to neighbouring layers [26].

Machine vision or computer vision deals with developing a system in which the input is an image, and the output is some information. DCNNs also benefitted from the stylization of visual content. echo the transition from hand-designed to end-to-end training solutions of visual identity pipelines, DL has changed our capacity by example to learn creative styles. and transfer the design to new imagery — addressing a big sub problem in the field of computer graphics Non-Photo Realistic Rendering (NPRR) [27].

A video is just a series of frames. In the direction of time, the video adds a new dimension to the picture. To achieve a better result, the spatial features of images and temporal features of the video can be compiled. The extra dimension also gives much space and thus increases training and inference complexity. The specifications for computing a video are extremely high. Video classification is the process of labelling a video with a class. On the frame level or for the entire video one category may be. Video also affects the design of DL models since temporal features must be considred.. A video classification can therefore identify the objects in the video or label the behaviour in the video [27].

In projects like Google's Deep Dream, stylized production has already been explored, where backpropagation has been used to optimize an image input for an image that maximizes one hot output on the discriminatory network (such as GoogLeNet). The initialization of such a network with white noise converges the entry for the desired object category to an optimal trigger image. More interestingly, initializing a photo optimization (plus Gaussian additive, i.e. white noise) is to hallucinate a locally optimal image, with structures that mimic the one-hot object being transformed to represent the object more closely [28].

4. The Build of Proposed VFDS

This paper aims primarily to use an algorithm that can boost classification efficiency using a DL method in the development of a robust and adaptable VFDS. Where the input is a suspicious video, and the output is two class values (pristine or forgery). The general block diagram of the proposed inter-frame video forgery detection system is stated in Figure 3.

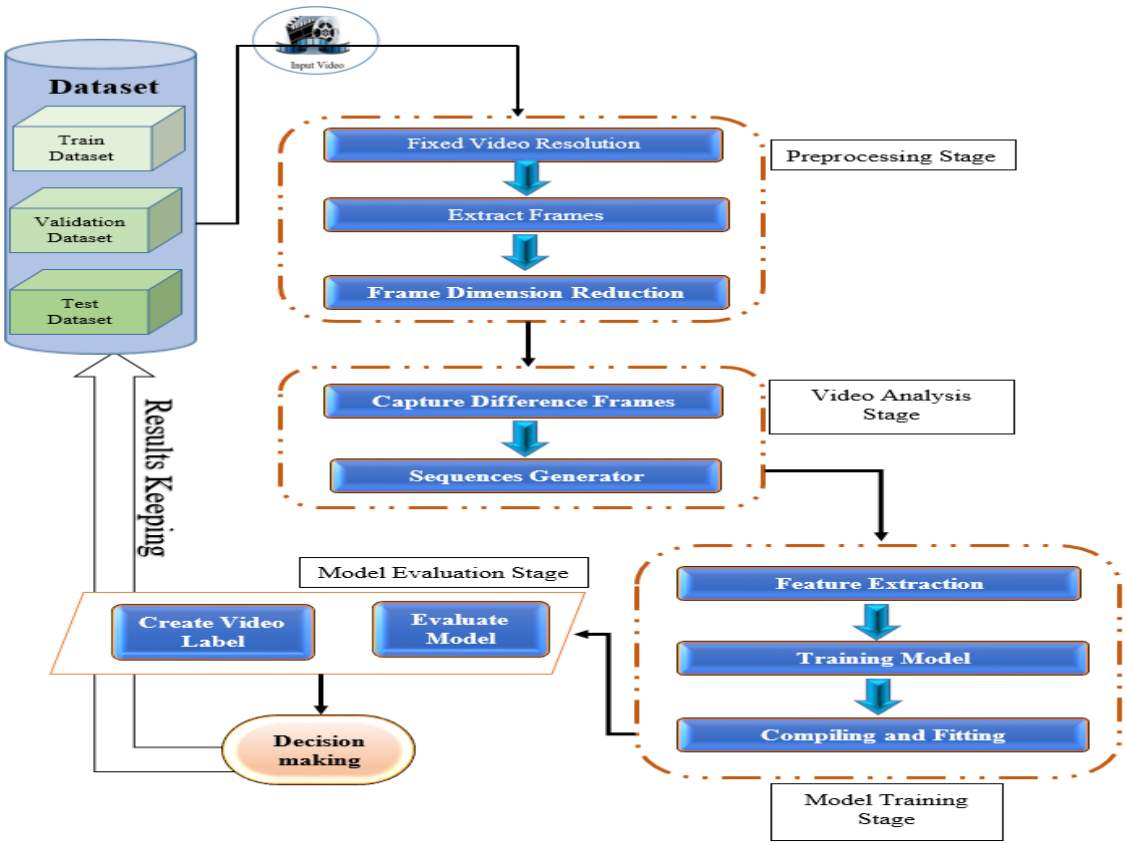

Figure 3- The proposed inter-frame video forgery detection system.

This section presents the proposed forensic system developed for detection of inter-frame forgery. The work in this paper proposes to build a model based on deep learning techniques, the 3D-CNN model. The proposed models can automatically distinguish pristine and forged video with no interaction with a computer user. VFDS comprises four stages. It starts with feeding video files into the system until it had detection results. The major strategy for detection of inter-frame forgery method is represents in Algorithm 1.

Algorithm 1- The proposed inter-frame video forgery detection system

Input	Video \\ Video files
Output	Detection Result #Pristine or Forgery
Procedure Begin	***Step 1***: • ***Set hyper_parameters***: set of parameters tuning to enhance training operation • ***Set total_frame***: the total number of frames in video clip • ***Set nb_frame***: the length of desired sequence • ***Set deff_frame***: the difference frames getting from two adjacent frames • ***Set frames_batches***: a new frames batch • ***Set frames_batches***: the total of batches that generated from each video. • ***Set input_shape***: input_shape = *NON x* 36 *x* 64 *x* 64 *x* 3, ***Where NON*** represent video file number ***Step 2***: Read video from dataset, the video collection selected from dataset is explained by Equation:

End	$V_i = \{V_1, V_2, V_3, \dots\dots, V_k\}$ $Where\ 1 \geq V \leq$ k ***Step_2***: Extracted frames from each video selected in step_1, frames are representing by using Equation: $F_i = \{F_1, F_2, F_3, \dots\dots, F_k\}$ $Where\ 1 \geq F \leq$ k ***Step_3***: Resize the extracted frames high=64, width=64 ***Step_4***: Image_to_array convert ***Step_5***: produce the difference frames by apply Equation: deff_frames= $\lvert I_{t-1}(x,y) - I_t(x,y)\rvert$ ***Step_6***: Create frames batches, where it number get from: frames_batches = total_frame / nb_frame, then ***Step_7***: Reading video frames sequences ***Step_8***: Apply 3D-CNN model training on all data. ***Step_9***: Pass result of 3D-Convolution layer to next ConvLSTM2D layer. ***Step_10***: Enhancement of the selected features by analysing them depending on hyper-layer ConvLSTM2D to maintain temporal features. ***Step_11***: FC layer gets the result of the ConvLSTM2D layer, converts them into a singular vector, implement feature analysis, gives the final probabilities. ***Step_12***: showing detection results.

4.1 Pre-processing Stage

Preprocessing is the first stage where the initial processes are executed for the initialization of the video file to manage it during the training. This stage comprises three levels that run respectively. These levels involve fixing the video resolution, frame extraction, and frame dimension reduction. It is an introductory process employed for improving video clips to use as data by VFDS. These processes are described in the following sections:

i. *Fix the Video Resolution*: The high-frequency detail videos proved important to get high accuracy in video forgery detection systems, but more resolution means more information and an increase in the computational cost. Therefore, attempting to downscale video resolution to reduce the computational cost and running time without decreasing performance.

ii. *Frame extraction*: Most video (films and television) programs are shot with 24-30 frame per second, and each image is called a frame where one can see the term fps. To detect forgery in a video, we need to extract individual frames from the video first to process the video. The extracted frames will be kept in A frame-buffer, which is a part of random-access memory (RAM). The principal functions of the frame buffer are the storage, conditioning, and production of the video signals that stimulate the display device.

iii. *Frame Dimension Reduction*: When implementing dimension reduction in the initial process, it is evident that some loss of information occurs during the operation. Thus, the goal is to save as much information as possible from the original color image. The next step is to resize each frame to 64x64 to reduce the size and make it fixed for all test videos. The purpose of this operation are to pick up more useful features and decrease the complexity in the computation operations.

4.2 Video Analysis Stage

Video content analysis, also known as video analysis or video analytics is the capability of automatically analyzing video to detect and determine temporal and spatial events. Many functionalities can be implemented in video content analysis. The frame difference is one of the simpler forms where an alter in correlation factor value is detected. Frames sequences are generated to achieve the best performance as explained in the next sections.

i. *The Difference Frame capturing*: It is an approach where the computer finds out the difference between successive video frames, as shown in Figure 4.

Figure 4- The original frames vs difference frames.

The set of frames below represent raw frames while the above frames represent frames product by applying difference frames approach. When the pixel value changes, it refers to the variation in frames content through time. The frame difference technique is the simplest way for distinguishing temporal alterations in the intensity and correlation of video frames. In RGB color image can compute the absolute difference for each pixel with coordinates (x, y) in a frame I_{t-1} with its corresponding coordinates in the next frame I_t, by using the absolute difference method as follows:

$$d(x,y) = |I_{t-1}(x,y) - I_t(x,y)| \quad (1)$$

ii. *Generate Frames Batches*: The desire is to send a sequence that involves several frames. The coveted shape is (n, f, h, w, ch) where n is the number of videos, f is the number of frames for a sequence, h and w are the height and width of the frame respectively, and ch is the number of color channels. The filter size is denoted by d × k × k, where d is the temporal depth of the kernel, k is its spatial size for 3D convolution and pooling layers. For example, if train a sequence of 5 images that are RBG and with 64 x 64 size, the shape should be (n, 5, 64, 64, 3). Videos are split into 37-frame batches, which become 36-frame after getting the frame difference to produce the input dimension are 36x64x64x3 as input to the model.

4.3 Training Stage

In general, executed DL mechanization using a sequence of convolutional, pooling, ConvLSTM, and a classification layer. The proposed 3D-CNN model architecture comprises 4 convolution layers,3 pooling layers, 2 Fully Connected Layers (FCLs), 5 dropout layers, and sigmoid layer to conclude the category labels. This CNN network begins with DV as entries. Then divides videos into frame sequences, which inputs to the first Conv3D layer of the network. The input shape is 36x320x240x3, where this matrix comprises the used features as mentioned in Figure 3 in the training model stage. Additionally, do jittering of the input DV, by using arbitrary crops with a size of 36x64x64x3, through training.

The collection of kernels for 4 convolution layers from 1 to 4, are 8,8,16, and 16, respectively. For all convolution layers, the size of the kernel is 3×3×3, and used a stride of 1; therefore, the input and output sizes of convolution layers are the same. use the 3D max-pooling kernel with a size of 2×2×2 just in the last two layers. During the first pooling layer, adopted a temporal kernel including depth d = 1, and a spatial kernel by size k = 2, so as not to incorporate the temporal signal too early. The network begins the learning procedure with a learning rate of 0.001, and a momentum of 0.9. To obviate over-fitting, use five dropout layers with a probability of 0.2 for all.

4.4 Predict Values and Evaluate Model Stage

The last phase of building the model is to make some prediction and evaluate the performance of the model, which are explained as shown Figure 5. Will be correct some parameters like learning rate, momentum, epochs number, and batch size.

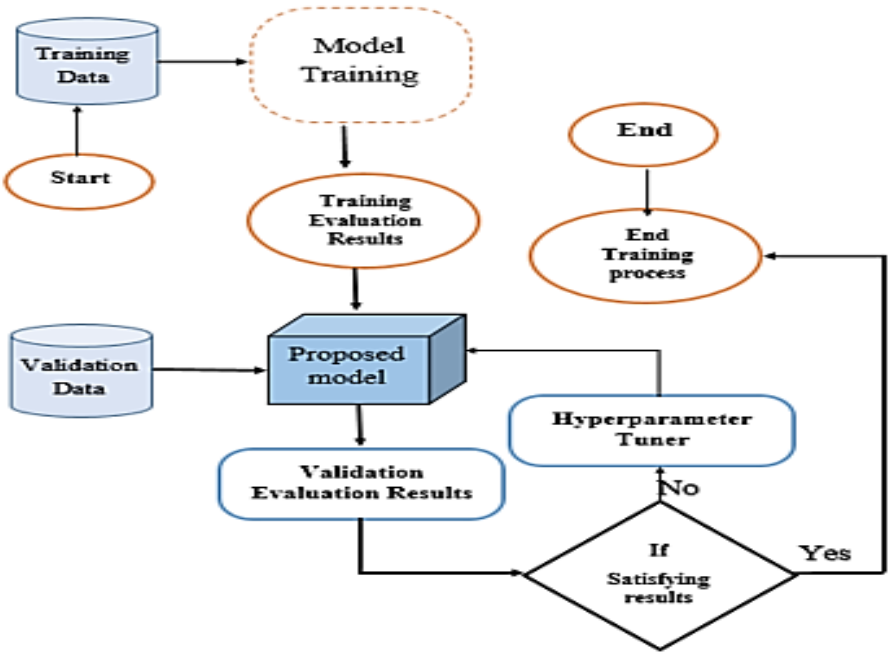

Figure 5- The proposed 3D-CNN model evaluation.

5. Datasets

DL offers an excellent toolkit for exploring features, but it is still hungry for data. An extensive dataset that includes many examples from several recent techniques is still not available in areas such as video manipulation. Two datasets were used in this work as follows:

5.1 The University of Central Florida UCF101 dataset

Considered the biggest data collection presently and includes 101 human action category. Table 1 described intrinsic properties for the UCF101 dataset.

Table 1- UCF101 characteristics Summary

Video Clips	13320
Mean Clip Length	7.21 second
Total Duration	1600 minutes
Min Clip Length	1.06 sec
Max Clip Length	71.04 second
Resolution	320 x 240
Frame Rate	25 frames/second

The dataset contains user-uploaded videos containing camera motion and a cluttered environment. It comprises of 13320 videos of different time lengths. Videos of the dataset are divided into three groups; the first is for training with 8000 videos, the second 3500 video are the validation dataset, and a third group for testing that consist of 2000 video.

5.2 Selected dataset

Adopted a video recorded with a surveillance camera at one intersection of the main road, and it comprised 12 hours and divided into 8,500 video clips. Also divided the original videos into training dataset 4000 videos, validation dataset 2500 videos, and a test dataset 2000 videos. By using the Fast Forward Movie Picture Experiment Group (FFmpeg) tool, we create forged video datasets from UCF101 and selected datasets for each type of inter-frame forgery, such as frame insertion, remove, shuffling, and duplication.

6. Results and Discussion

The difference frames offer statistical information that supports the proposed models to determine video behavior. Depending on this information, to build a decision through identifying the patterns of normal and tampered video. Table 2 illustrated the statistical information gained from the difference frame production operation.

Table 2- One value represent value mean of each frame.

Frame No.	Pristine	Insert	Delete	Duplicate	Shuffling
1	8.24E-05	1.01E-04	8.87E-05	1.09E-04	7.45E-05
2	9.48E-05	4.82E-05	4.39E-05	5.17E-05	4.71E-05
3	4.79E-05	2.64E-05	2.25E-05	3.65E-05	2.64E-05
4	1.01E-04	1.53E-04	1.39E-04	1.41E-04	1.40E-04
5	1.02E-02	1.02E-02	1.02E-02	1.02E-02	1.02E-02
6	4.16E-03	3.13E-03	3.15E-03	3.05E-03	3.01E-03
7	1.00E-04	2.08E-04	2.35E-04	1.73E-04	1.51E-04
8	4.86E-05	9.06E-05	8.25E-05	6.80E-05	9.19E-05
9	8.11E-05	1.67E-04	1.51E-04	1.17E-04	1.66E-04
10	2.88E-04	5.11E-04	4.75E-04	5.37E-04	5.37E-04
11	1.22E-02	1.19E-02	1.19E-02	1.20E-02	1.18E-02
12	3.30E-04	2.73E-04	3.09E-04	3.79E-04	3.65E-04
13	8.20E-04	6.56E-04	7.36E-04	1.93E-02	1.92E-02
14	2.17E-04	2.44E-04	3.93E-04	3.23E-03	2.87E-03
15	1.13E-02	1.12E-02	1.12E-02	2.32E-04	3.39E-04
16	2.21E-04	2.84E-04	3.35E-04	8.10E-05	1.40E-04
17	9.49E-04	1.14E-03	1.73E-02	2.01E-04	3.16E-04
18	1.73E-04	2.18E-02	9.41E-03	9.76E-03	9.77E-03
19	1.06E-02	8.96E-05	5.51E-04	3.60E-04	3.28E-04
20	1.32E-03	1.61E-04	1.78E-04	1.80E-04	1.69E-04
21	1.98E-04	5.36E-04	3.10E-04	4.20E-04	4.85E-04
22	1.69E-04	9.20E-03	1.02E-02	1.01E-02	1.02E-02
23	1.06E-02	3.35E-04	3.88E-04	1.37E-02	2.15E-02
24	5.98E-03	2.58E-04	1.27E-04	4.64E-04	1.68E-04
25	6.69E-05	1.15E-03	2.77E-04	2.75E-04	1.08E-02
26	4.54E-05	9.25E-03	9.98E-03	1.46E-04	2.37E-04
27	8.98E-05	5.61E-04	5.55E-04	7.69E-05	3.77E-04

28	9.46E-03	4.20E-04	1.51E-04	9.46E-03	2.30E-04
29	4.59E-04	4.24E-04	5.03E-04	3.44E-04	1.05E-02
30	1.08E-04	1.86E-02	4.03E-04	1.96E-04	5.72E-04
31	4.26E-04	3.12E-04	9.01E-03	1.92E-04	1.51E-04
32	1.00E-02	9.92E-03	3.34E-04	1.00E-02	2.15E-04
33	4.19E-04	2.55E-04	1.48E-04	3.31E-04	1.73E-02
34	9.25E-05	8.97E-05	1.16E-03	1.60E-04	2.31E-04
35	4.07E-04	1.98E-04	9.28E-03	1.62E-04	2.99E-04
36	9.89E-03	9.89E-03	6.04E-04	9.84E-03	9.94E-03

The arithmetic mean is the total of data divided by the whole of data-points, a measure of the middle location of data in a collection of values that differ in range. By employing the individual value of frames we can get more understanding of patterns that formulate video behavior, as shown in Figure 6, and how to distinguish between pristine and forged video using this technique.

The inter-frame forgery process will decrease the correlation between neighboring frames at tampering places. Then, the consistency of the connected correlation coefficients is disordered. Forgery detection could be obtained by identifying these discontinuous points, called abnormal points. When computing the correlation of contiguous frames, we use the pixel-wise difference as a metric to the value of the correlation factor. Then, the classifier layers classify the regular and abnormal points into two classes.

As shown in Table 2, the contents of contiguous frames in the video are mostly close with special pattern, while the contents of remote frames may vary. Consequently, used the correlation factor as a metric of the connection of the inter-frame content. As illustrated in Figure 6-b a series of frames with inter-frame correlation factors were damaged because the insertion of 12 frames. This result, as can see in the two peaks at frames 18 and 30, occurred because the correlation factor at these frames differs from that in the original video. The value of the correlation factors is close to each other in the original video. Thus, after it was subjected to frame insertion forgery, the value of the correlation factor reduced at the forgery position in frame 18 and frame 30.

Figure 6-d gives an instance of frames duplication attack, where the original video was subjected to duplication of 10 frames from 14 to 24. The value of the correlation factor reduced at the tamper position at frame 14 and frame 24. The result of frame deletion was presented in Figure 6-c where 12 frames were deleted. The correlation factor decreased at the location where frames are removed from the video. In this state, the difference in value occurs just at the position of frame 18.

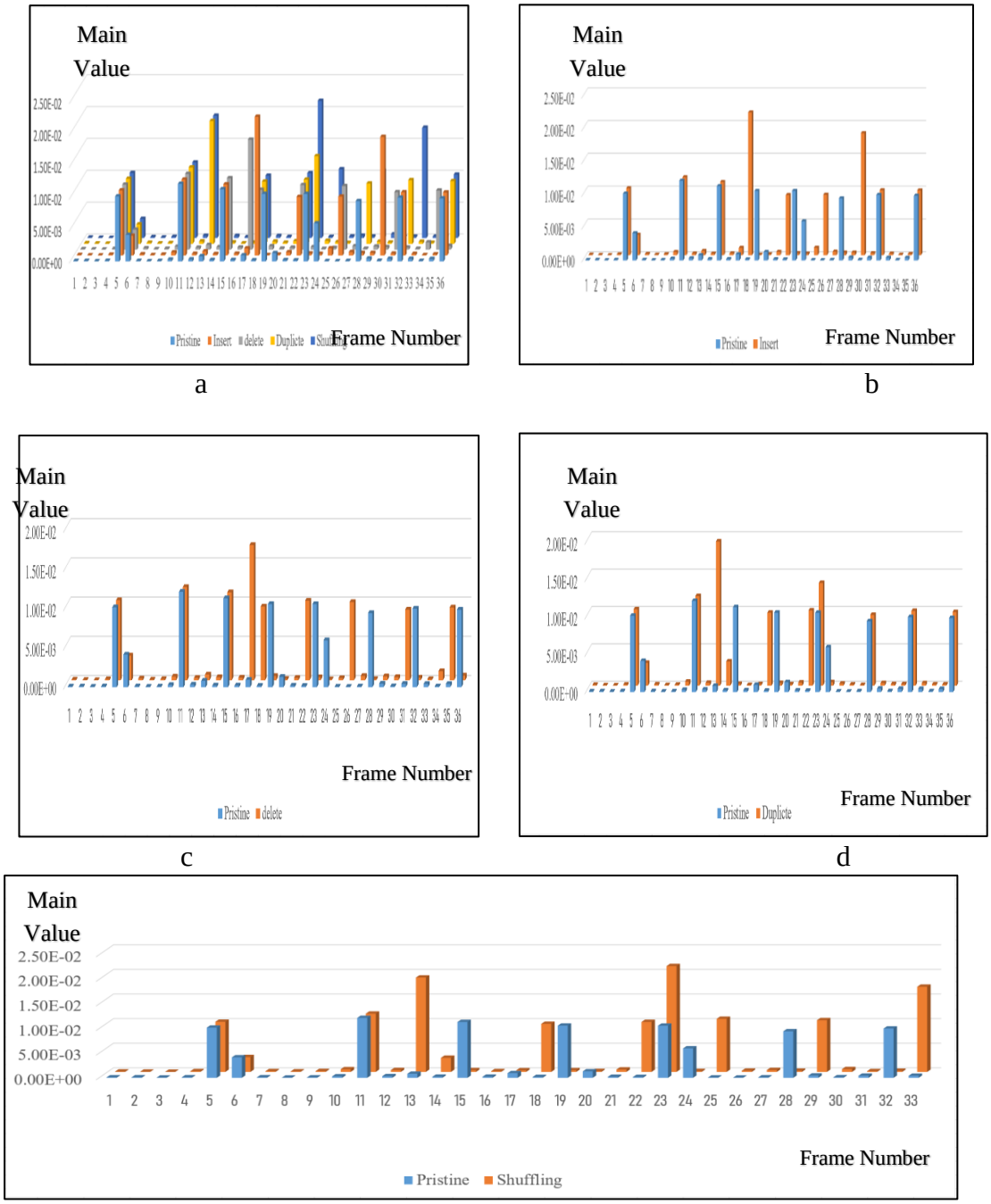

Figure 6- The forgery operation effected on video frames correlation, where (a)Impact forgery process on video behaviour, (b) The insertion forgery pattern, (c) The deletion forgery pattern, (d). The duplicate forgery pattern, (e) The shuffling forgery Pattern

6.1 Result Performance Evaluation

The production of the CNN is a binary probability matrix, adopted for mathematical function, to distinguish the actual and counterfeited videos. Through this experiment, the performance of the suggested models in terms of accuracy were calculated using the equation:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (2)$$

Where TP denotes the number of True Positives or the number of forged videos rightly detected to be forged. FP denotes the number of False Positives or the number of pristine videos detected as forged. TN describes the number of True Negatives or the number of pristine videos accurately detected as pristine FN denotes the number of False Negatives, and

the number of forged videos falsely detected as pristine. Table 3 and Table 4 display the results of the proposed training model by the UCF101 and selected datasets, respectively.

Table 3- The gained results (%) of training proposed model by the UCF101 dataset.

Types of Attacks	Frames	3D-CNN
Frame Deletion	10	96.2
Frame Deletion	30	98.08
Frame Insertion	10	99.1
Frame Insertion	30	99.17
Frame Duplication	10	99.3
Frame Duplication	30	99.41
Frame Shuffling	10	98.19
Frame Shuffling	30	98.04

Table 4- The gained results (%) of training proposed model by the selected dataset.

Types of Attacks	Frames	3D-CNN
Frame Deletion	10	98.2
Frame Insertion	10	99.51
Frame Duplication	10	99.34
Frame Shuffling	10	99.54

Table 5 summarizes the forgery types that were detect in related literature according to the publication date. Table 6 lists the average accuracy value of the proposed models and the set of methods that handled the inter-frame video forgery.

Table 5- Comparison of the proposed models with related literature regarded detected types

References No.	Year	Frame Insertion	Frame Deletion	Frame Duplication	Frame Shuffling
[11]	2017	√	√	√	X
[12]	2017	√	√	X	X
[13]	2017	√	X	X	X
[14]	2018	√	√	√	X
[15]	2018	√	√	X	X
[18]	2020	√	√	X	X
The Proposed Model	*2021*	√	√	√	√

Table 6- Accuracy of proposed models compared with other literature

References No.	Year	Accuracy (%)
[11]	2017	*90*
[12]	2017	*97.5*
[13]	2017	*93.92*
[14]	2018	*97*
[15]	2018	*94.47*
[18]	2020	*99.17*
The Proposed Model	*2021*	*99.14*

7. Conclusions

In this paper, we presented a forensic system comprising a simple yet powerful detection of falsifying method. The proposed model is based on the deep learning algorithm: 3D-CNN for digital video counterfeit detection. It provided the pixel difference that transfer temporal information to the subsequent CNN layer, which corresponds to a video batch. In consideration of video forgery, inter-frame is a temporal forgery operation. The subsequent convolution layers of the CNN may find better features by leveraging this temporal information to identify forgery inter-frame video.

For video with static and dynamic background, the proposed model can detect and locate frames inserted, deleted, mixed and duplicated, and establish their superiority regarding forgery detection accuracy.

References

[1] J. Bakas and R. Naskar, "A Digital Forensic Technique for Inter–Frame Video Forgery Detection Based on 3D CNN," in *International Conference on Information Systems Security*, 2018, pp. 304-317: Springer.

[2] A. Katsaounidou, C. Dimoulas, and A. Veglis, *Cross-Media Authentication and Verification: Emerging Research and Opportunities: Emerging Research and Opportunities*. IGI Global, 2018.

[3] K. Sitara and B. M. J. D. I. Mehtre, "Digital video tampering detection: An overview of passive techniques," vol. 18, pp. 8-22, 2016.

[4] G. Ulutas and G. J. M. P. i. E. Muzaffer, "A new copy move forgery detection method resistant to object removal with uniform background forgery," vol. 2016, 2016.

[5] K. Sowmya, H. J. I. J. o. C. E. Chennamma, and Applications, "A survey on video forgery detection," vol. 9, no. 2, pp. 17-27, 2015.

[6] V. Kumar, A. Singh, and M. Gaur, "A Comprehensive Analysis on Video Forgery Detection Techniques," 2020.

[7] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. J. N. Lew, "Deep learning for visual understanding: A review," vol. 187, pp. 27-48, 2016.

[8] M. J. a. p. a. Moustafa, "Applying deep learning to classify pornographic images and videos," 2015.

[9] Y. LeCun, Y. Bengio, and G. J. n. Hinton, "Deep learning," vol. 521, no. 7553, pp. 436-444, 2015.

[10] J. Xing, K. Li, W. Hu, C. Yuan, and H. J. P. R. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," vol. 66, pp. 106-116, 2017.

[11] S. Kingra, N. Aggarwal, R. D. J. I. J. o. E. Singh, and C. Engineering, "Video Inter-frame Forgery Detection Approach for Surveillance and Mobile Recorded Videos," vol. 7, no. 2, 2017.

[12] Y. Liu and T. J. M. S. Huang, "Exposing video inter-frame forgery by Zernike opponent chromaticity moments and coarseness analysis," vol. 23, no. 2, pp. 223-238, 2017.

[13] C. C. Huang, Y. Zhang, and V. L. Thing, "Inter-frame video forgery detection based on multi-level subtraction approach for realistic video forensic applications," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 2017, pp. 20-24: IEEE.

[14] S. M. Fadl, Q. Han, and Q. J. I. I. P. Li, "Inter-frame forgery detection based on differential energy of residue," vol. 13, no. 3, pp. 522-528, 2018.

[15] Q. Li, R. Wang, and D. J. I. Xu, "An Inter-Frame Forgery Detection Algorithm for Surveillance Video," vol. 9, no. 12, p. 301, 2018.

[16] S. Jia, Z. Xu, H. Wang, C. Feng, and T. J. I. A. Wang, "Coarse-to-fine copy-move forgery detection for video forensics," vol. 6, pp. 25323-25335, 2018.

[17] H. Kaur and N. J. W. P. C. Jindal, "Deep Convolutional Neural Network for Graphics Forgery Detection in Video," pp. 1-19, 2020.

[18] X. H. Nguyena, Y. HUb, K. G. Hayatc, V. T. Led, T. D. J. A. J. o. C. S. Truonge, and Applications, "Detecting Video Inter-Frame Forgeries Based on Convolutional Neural Network Models," vol. 3, 2020.

[19] B. Aminu Mustapha, "Passive video forgery detection using frame correlation statistical features/Aminu Mustapha Bagiwa," University of Malaya, 2017.

[20] H. Kaur and N. J. W. P. C. Jindal, "Image and Video Forensics: A Critical Survey," pp. 1-22, 2020.
[21] K. Sitara and B. Mehtre, "A comprehensive approach for exposing inter-frame video forgeries," in *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, 2017, pp. 73-78: IEEE.
[22] M. U. Mulla and P. R. J. I. J. S. R. C. S. E. I. T. Bevinamarad, "Review of techniques for the detection of passive video forgeries," vol. 2, pp. 199-203, 2017.
[23] R. Sawant and M. J. J. o. C. E. Sabnis, "A Review of Video Forgery and Its Detection," vol. 20, pp. 1-4, 2018.
[24] O. I. Al-Sanjary, A. A. Ahmed, and G. J. F. s. i. Sulong, "Development of a video tampering dataset for forensic investigation," vol. 266, pp. 565-572, 2016.
[25] K. Sitara and B. J. F. s. i. Mehtre, "Detection of inter-frame forgeries in digital videos," vol. 289, pp. 186-206, 2018.
[26] A. Shrestha and A. J. I. A. Mahmood, "Review of deep learning algorithms and architectures," vol. 7, pp. 53040-53065, 2019.
[27] J. E. Kyprianidis, J. Collomosse, T. Wang, T. J. I. t. o. v. Isenberg, and c. graphics, "State of the" Art": A Taxonomy of Artistic Stylization Techniques for Images and Video," vol. 19, no. 5, pp. 866-885, 2012.
[28] A. Mordvintsev, C. Olah, and M. J. G. R. Tyka, "Deepdream-a code example for visualizing neural networks," vol. 2, no. 5, 2015.