Deep Convolutional Neural Network for Object Forgery Detection in Video

A Thesis Submitted in Fulfillment of the Requirement for the Award of the Degree of

Master of Engineering

In

Electronics and Communication Engineering

Submitted By

Harpreet Kaur

Roll No: 801661009

Under Supervision of

Dr. Neeru Jindal

Assistant Professor,ECED

ELECTRONICS AND COMMUNICATION ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY), PATIALA, PUNJAB

JULY, 2018

DECLARATION

I, Harpreet Kaur hereby declare that the work presented in this thesis entitled **"Deep Convolutional Neural Network for Object Forgery Detection in Video"** in fulfilment of the requirement for the award of degree of Master of Engineering (ECE) submitted at Electronics and Communication department, Thapar Institute of Engineering and Technology, Patiala is an authentic record of work carried out under supervision of **Dr. Neeru Jindal** (Assistant Professor), Electronics and Communication Department, Thapar Institute of Engineering and Technology, Patiala from 2017 to 2018.

The matter presented in this thesis has not been submitted either in part or full to any other university or institute for the award of any other degree.

Harpreet Kaur.

Date: 11/7/18

Harpreet Kaur
801661009

It is certified that the above statement made by the candidate is correct to the best of my knowledge and belief.

Neeru Jindal

Date: 11/7/18

Neeru Jindal
Assistant Professor
Electronics and Communication Engineering Department
Thapar Institute of Engineering & Technology
(Deemed to be University), Patiala, Punjab

ACKNOWLEDGEMENT

From the depth of my heart, I express my deep earnest gratitude to the Almighty for the blessings, wisdom, good health and pertinence. He had bestowed upon me to successfully accomplish this work.

In the endeavour of journey towards M.E., this thesis has been finalized with the support and inspiration of numerous individuals including my mentors, friends and well-wishers. I would like to express my profound gratitude and heartfelt thanks to my supervisor **Dr. Neeru Jindal**, Assistant Professor, Electronics and Communication Engineering Department, Thapar Institute of Engineering and Technology, Patiala who has always uplifted me throughout the work. I am truly privileged with this opportunity to work under her administration. Her exemplary guidance, valuable advice, constructive suggestions and extensive discussions during my work secured the final completion of this thesis.

I am extremely gratified and wish to owe my honest gratitude to **Dr. Alpana Agarwal**, Head of the Department, and **Dr. Amit Mishra**, Program Coordinator, Electronics and Communication Engineering Department, Thapar Institute of Engineering and Technology, Patiala who ensured the availability of all learning facilities and infrastructure in ECED.

I am thankful to my friends for providing a stimulating and fun filled environment and indebted to my family for the sincere encouragement in needy times and being my inexhaustible pillar of supports.

Last but not least, I would like to pay best regards to all individuals who directly or indirectly assisted me in the development and completion towards this work.

Harpreet Kaur

Harpreet Kaur

ABSTRACT

Talking of today's digital revolution, where visual data is playing an imperative role, accessing, processing, and sharing of most of the information is typically attained with the help of video. These video sequences have shown their significance in various fields like news broadcasting, legal trials in court rooms, and many more but the doctoring of authentic visual content has made it uncertain to use as an evidence. Doctored video generation with a fast-growing rate done by easily accessible editing software like Adobe Photoshop, filmora, etc. have proved to be a major problem in maintaining its authenticity. The extent of forging is so vast that video spoofs reach our electronic-mail in-boxes, WhatsApp, Facebook or any other social media every minute and this fakery is totally indistinguishable that hence raise a demand for a new versatile field to perceive any alteration. Video forgery detection aims at restoring the trust and validating the authenticity by uncovering the counterfeits. But the traditional approaches used so far to detect forgeries have faced difficulties like less accurate detection rate and more false negatives. Nowadays, deep neural networks have been recognized as an effective technique in eradicating such troubles by learning significant features. The increasing attempt of video modification has drawn greater attention towards Deep Convolutional Neural Networks (DCNN) for achieving better counterfeits recognition.

The proposed work is about "**Deep Convolutional Neural Network for Object Forgery Detection in Video**" that aims to detect forgery without requiring additional pre-embedded information of the frame. The proposed DCNN consists of various neurons where weights and biases are defined for individual neuron which helps the network to learn the data properly. Unlike other pre-existing learning-techniques, the proposed algorithm classifies the forged frames on the basis of correlation among them and the observed abnormalities using DCNN. The decoders used for batch normalization of input improves the training swiftness. It leads to an inordinate evidence in recognizing and discovering the fake regions. Simulation results are obtained on MATLAB 2018a with NVIDIA Cuda Graphics with REWIND and GRIP dataset which is rich in video inter-frame forgery effects. The outcomes so obtained with an average accuracy of 99% shows the superiority of the proposed algorithm as compared to existing one. The robustness of proposed algorithm is also tested on You Tube compressed video sequences. Recurrent Neural Networks can be combined with DCNN to achieve comparatively remarkable results in future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CFA	Color Filter Array
CMFD	Copy Move Forgery Detection
CNN	Convolutional Neural Networks
FC	Fully Connected
FNR	False Negative Rate
FPR	False Positive Rate
GRIP	Image Processing Research Group
MIFT	Mirror-reflection Invariant Feature Transformation
MPEG	Moving Picture Experts Group
ReLU	Rectified Linear Unit
REWIND	REVerse Engineering of audio-Visual content Data
RNN	Recurrent Neural Networks
SIFT	Scale Invariant Feature Transformation
SPN	Sensor Pattern Noise
SULFA	Surrey University Library for Forensic Analysis
SURF	Speeded UP Robust Features
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate

CHAPTER 1
INRTODUCTION

1.1 PREAMBLE

Digital media corresponds to the primary means of communication. Video demonstration have proved to be the productive way of sharing sentiments and thoughts. Video is defined as a continuous or analog signal denoted by $f(x, y, t)$, and x and y are the space-coordinates and t is the time variable [1]. The extensive creation of reasonable-priced and portable video capturing devices, like digital cameras and cell phones, has activated a rapid improvement in the generation of visual data. The multimedia content on the internet have been altered with a growing rate and now a data the easy accessibility of powerful editing software like Adobe Photoshop, GIMP, filmora, Pinnacle, Nero, etc. has drawn our concern towards the authenticity of such media.

Many researchers [11,12,20,32,54] from the last decade have played an important role in detecting these falsifications in video by giving several techniques. Some pre-existing techniques are SIFT (Scale Invariant Feature Transformation), MIFT (Mirror-reflection Invariant Feature Transformation), SURF (Speeded UP Robust Features) [53,54], moment and noise correlation [29], optical flow, etc. But the approaches based on these concepts suffer various drawbacks like:

- Low detection accuracy,
- More false negatives,
- Less detection rate, etc.

These drawbacks demanded a new research field that effortlessly perceives the modifications in the video if any. Deep Learning has gain recent significance in building an effective and highly accurate framework by working on large set databases to detect such alterations. Deep Convolutional Neural Networks (DCNN) falls in deep research field of Artificial Intelligence which are largely utilized to categorize images and grouping them by means of similarity in features so extracted and finally performing further investigations within the frames like object or pattern recognition, forgery detection and localization. From the last two decades, more and more information has been captured because of the rise of mobile phones and less expensive digital cameras. The computing power is also on the rise as high-speed CPUs are coming into

picture and GPUs are becoming a general computing tool. Hence, the rise of both tools and data have made the accessing of neural networks an interesting task.

Although different types of neural networks do exist but while dealing with the large information at once like it is in video-frames, Deep Convolutional Neural Networks proved to be the best. As DCNN makes the best use of encoded information within the video-frame. Other types of forgeries like audio, text also exists but the proposed work focuses on video forgery and its detection.

1.2 VIDEO FORGERY

Video forgery is intended to hide or erase some important particulars [2] from a recorded sequence and hence aims to create anomalous scenarios that appear ordinary or make illegitimate duplicates of the original.

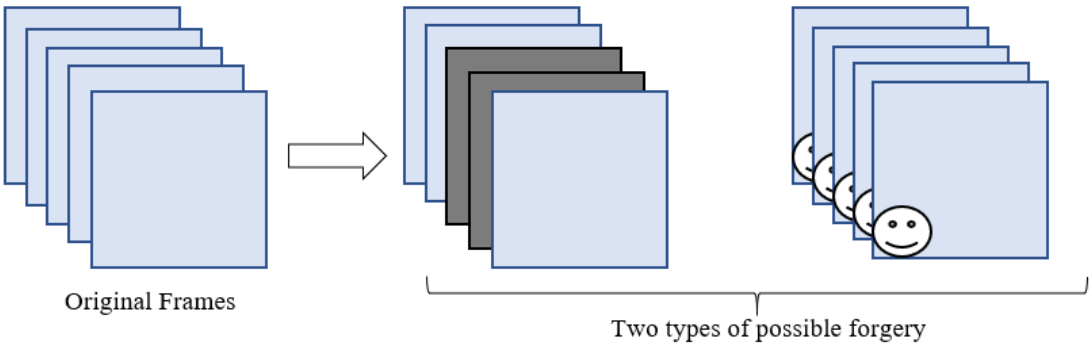


Figure 1.1 Illustration of basic video frame forgery where original frames suffer from frame duplication and region duplication

In dictionary, Video forgery or tampering means falsely interfering with authentic contents of the video in order to cause the mutilation alterations. High-end profile visual data capturing devices have played a significant role in making such aggressive alterations purely in-distinguishable to human eyes. Technological advancement of numerous video processing tools has made tampering of digital video data easy and faster.

Video tampering can be done by replication of frames in the video sequence called Frame Duplication. This type of alteration is tough to perceive. An example of doctored video is given in the Figure 1.1 which shows how some frames are duplicated in original set of authentic frames forming frame duplication forgery and whenever objects are introduced or pasted in sequence of frames to generate extra objects or cover a region inside the scene, it is called as

Region Duplication. To fool people for personal interest, video forgery is created. Figure 1.2 is a basic illustration of some original and forged video frames. It shows how the scene of the Television inside the frames is changed and it is hardly perceived by human eye. Hence, the need for the detection of such falsification is required.

Figure 1.2 Top row shows the original frames and bottom row is of forged frames from REWIND [3] video dataset

1.3 NEED OF VIDEO FORGERY DETECTION

It is difficult to perceive any fake content by the naked eye. Digital cinema uses various computer technologies to produce unreal movies like creating an animal out of human. This exactly shows how vast digital tampering has created it roots. But this sharing is not done as such. Following are some reasons why the voice for the need of forgery detection has been raised.

- The abundant availability of visual-data editing software like Adobe Suite, OpenShot, filmora, etc., allow anyone to make in-distinguishable alterations. The utilization of high quality software tools which can easily alter the content within the video has interrogated the genuineness of the video sequences so far.

- Various key fields like journalism, justice courtrooms, worldwide conferences use video media as a means of communication. But the guarantee about the validation of such a content can never be provided.

Hence the necessity of such a field arose which effortlessly detects the alterations in the video if any. Video forensics is a discipline of research that targets to validate the authenticity of such tampered video by recovering the information that has been forged. The dictionary meaning of 'authenticity' is 'the property of being genuine or valid, not being a fake or forgery.' Hence this field helps a lot in maintaining a multimedia security. There are various techniques [1-3] that helps in achieving this task.

1.4 CLASSIFICATION OF VIDEO FORGERY DETECTION TECHNIQUES

Various approaches exist to verify the tampered content of video. Researchers in the history of forensics classified forgery detection approaches into two areas as: Active approach like digital watermarking or digital signatures to defend the truthfulness of visual information and Passive approach [4] or Blind Detection.

1.4.1 Active Methods

Researchers in premature time came up with **Active Forensics** to defend the truthfulness of visual information. Active approach [5] was proposed in past where some information has to be kept at source (camera) side. This information can be Digital watermark or Digital signature. Later during the acquisition process, any alterations in the image can be observed by comparing values of such watermarks or signatures. Though these methods require some early source data related to capture forgery, so the detection becomes a challenging task. These approaches need an equipment of such facility with digital cameras to store Digital signature or Digital watermark. Some companies place a large watermark that creates a distraction in visualizing the content of the video. Many de-watermarking software exist in market that does not make the active approaches fool-proof. Common standard protocols are needed to be set for digital cameras for the applicability of this technique. Satisfying such requirement for cameras is too difficult. This puts a constraint in applying solutions to limited scenarios.

1.4.2 Passive Methods

The problems for active methods like requirement of pre-embedded information related to source camera, large watermarking, etc. has been conquered by a new method for authentication of video-frames. **Passive Forensics** methods depicted our concern towards it in the past few decades. Passive clears its meaning by blindly examining the binary information with no external data required [6]. It aims at localizing the tampering on raw video. It is based

upon the assumption that tampering the contents of video may likely interrupt the semantics and statistical properties or inconsistencies in the contents that introduce some new artifacts in the image. Study of these artifacts can help in finding the forged region in video. These methods proved efficient for the images but the challenge for detecting doctored encoded video stood still in terms of more accuracy and authenticity. Moreover, the forged exposure for the video with high resolution and long length with the traditional methods showed low-performance measures and less efficient.

Several types of tampering attacks have been observed in the video under passive methods. The broad classification of such ways of tampering has been listed in the next section.

1.5 TAMPERING OF VIDEO SEQUENCE IN PASSIVE METHODS

There are numerous different classes of video tampering or forgery, but all of them typically fall under two categories only: Inter-frame Video Forgery and Intra-Frame Video Forgery [7]. The basic structure is shown in Figure 1.3.

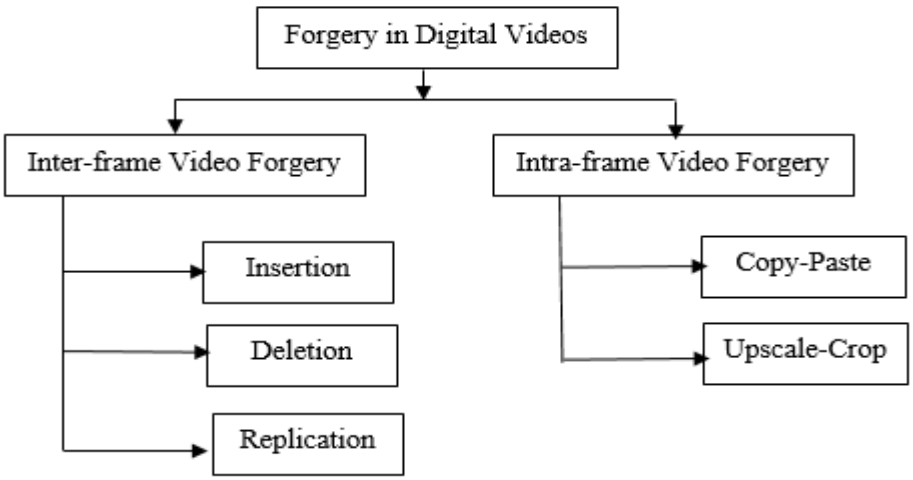


Figure 1.3 Various Methods for the accomplishment of Digital Video Forgery

1.5.1 Inter-frame Video Forgery

The inter-frame forgery means, tampering the contents within the same frame of video. This forgery is able to affect the sequence of frames in a video. Usually, the set of frames from the video is copied and inserted into the same video [7] at another temporal location. One more kind that is considered as inter-frame forgery is temporal splicing, where frames of various video are inserted in order produce a new video. Figure 1.4 is an illustration of different ways of inter-frame video forgery. As the relation of adjacent frames is quite the same, it becomes

computationally an expensive task to search for all those possible frames with their locations and sizes. Following is its descriptive analysis:

i. *Frame-Insertion Video Forgery:* As clear from the name, some set of frames from some other video sequence are inserted into the original video sequence. To keep the video of equal characteristics, the inserted frames are sometimes resized. The frame rate of forged video is also kept same as that of original video but the total number of frames in forged video increases. The true depiction of this type of forgery is shown in Figure 1.4 (b).

ii. *Frame-Deletion Video Forgery:* As clear from the name, some set of frames from original video sequence are deleted. This is usually done to delete out the important clues from the video. Like during criminal trials in courtrooms, if the convict shows its proof by submitting a spoofed video, then a convict can be proved innocent. The total number of frames from the forged video decreases than that of original video. Figure 1.4 (c) shows that the frame 7,8 and 9 has been deleted from the original video.

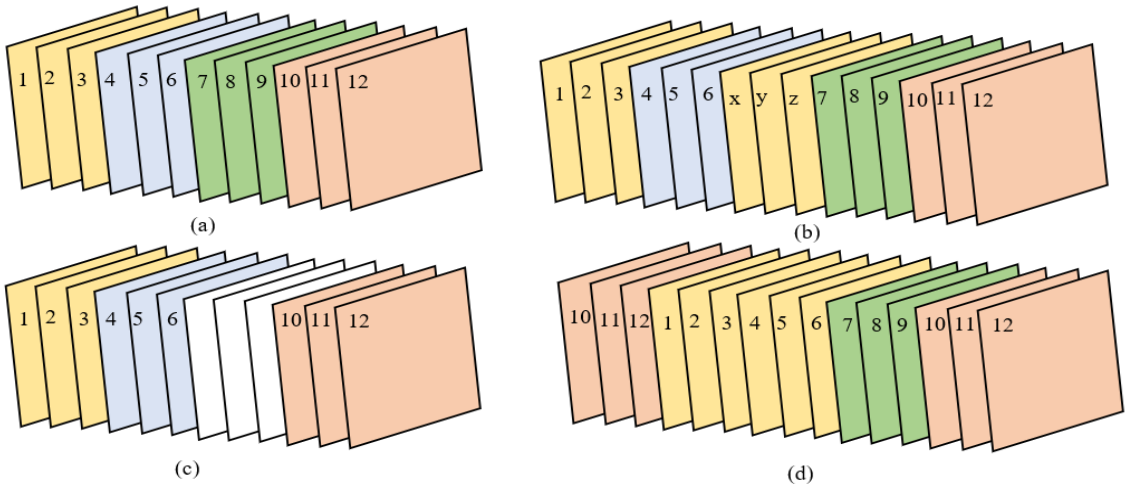


Figure 1.4 Various kinds of inter-frame video forgery. (a) illustrates Original Frame Sequence (b) illustrates Frame-Insertion Forgery (c) picturizes Frame-Deletion Video Forgery (d) is a graphical representation of Frame-Duplication

iii. *Frame-Duplication Video Forgery:* When some set of frames from original video sequence are copied from one temporal location and pasted to the same video to some other temporal location, it is known as Frame-Duplication. Following Figure 1.5 is an example that shows this type of video forgery. Figure 1.5 comprises of

frames from the inter-frame video forgery dataset copyright REWIND (REVerse Engineering of audio-Visual content Data) [3].

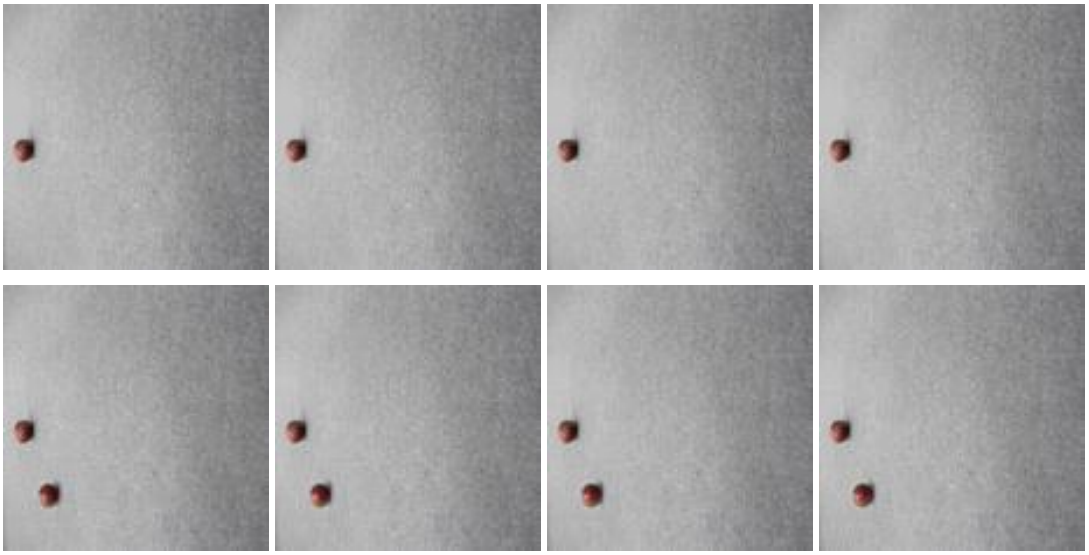


Figure 1.5 Frames in top row are original and bottom row shows duplicated frames from REWIND video dataset [3]

1.5.2 Intra-frame Video Forgery

In an intra-frame forgery, the actual contents of individual frames are modified. Analysis is performed by considering one only frame at a time. When some copy move segment is added to the video frame whether inter or intra, it adds a new relationship between the segment of original frame and that of the pasted one (original pixel-alignment gets changed). Cloned parts (whether patches or blocks) can attain any shape and can be present at any spatial or temporal location. Sometimes the objects are inserted in full temporal length of the video to make un-recognizable. Following are the two different types of intra-frame video forgery:

i. *Copy-Paste or Partial Video Forgery:* In this type of video forgery, an external object is added or sometimes removed to or from the authentic video in order to form a new video sequence. The word 'partial' indicated here is because only a minor section of the frame suffers alteration whereas the remaining frame remains unharmed. The following Figure 1.6 shows the example of the same from the GRIP (Image Processing Research Group) copyright dataset where the object tank has been copied form some other video and pasted into the original video frames to make it an in-distinguishable part of it.

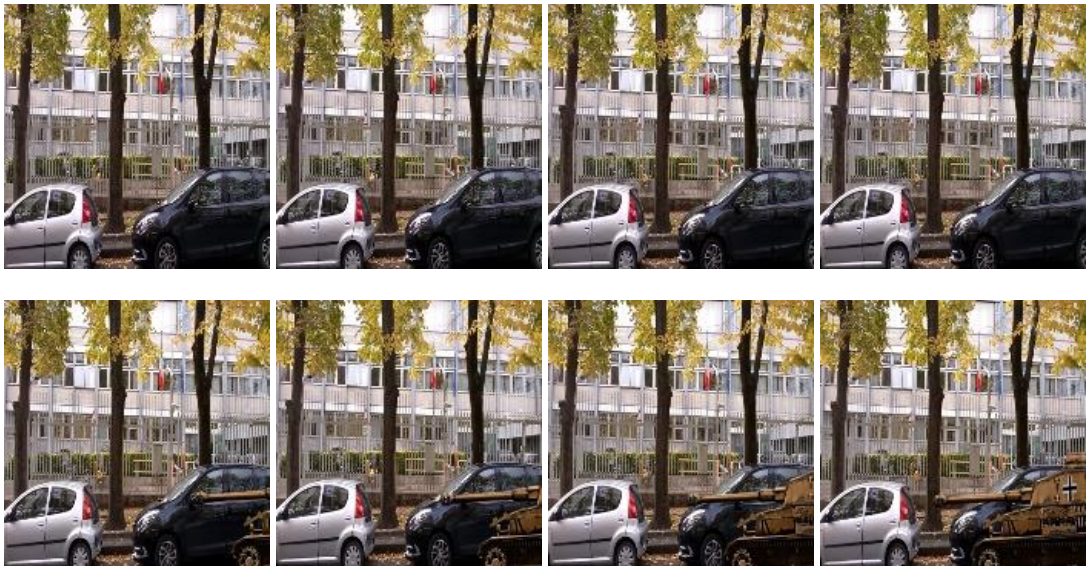


Figure 1.6 Top row shows the original frames and bottom four are forged frames where an external object tank is pasted inside it. [8]

Green-screening or blue-screening is also an example of intra-frame video forgery. Many movie producers use green-screen as the background during shoot and convert them to some attractive background to give a real view. Figure 1.7 is a perfect example of this type of alteration form movie Avengers where green color background has been changed to some other background to give the scene an attractive view.

Figure 1.7 Some scenes from the movie Avengers showing how green screen has been changed [9]

ii. *Upscale Crop Video Forgery:* In these type of tampering, the some of the uttermost part of frame in video is cropped to eradicate evidence of incidence and to match the inner dimensions of the frames in the entire video, these frames are again enlarged.

As the above discussion shows how difficult it is to perceive the difference between the doctored and original content of the video, there exist various methods to detect these forgeries.

1.6 VIDEO FORGERY DETECTION APPROACHES

1.6.1 Conventional Approaches

The generalized algorithmic steps for video forgery detection using traditional approaches are shown in Figure 1.8. The first step for the detection of forgery always start with acquiring a video or shooting a video. During the acquisition process, due to the inconsistencies in the camera's optical lenses, the imaging sensor can introduce artifacts in the frames captures at the initial stage. The visual data capturing sensor is considered as the heart of every digital camera which consists of an array of photo detectors. After this, filtering of light is performed by the [10] Color Filter Array (CFA) whose pattern further depends on the manufacturer. The missing pixel values are estimated by the process of Demosaicing. Before the final storage in the memory device additional processing is performed that includes enhancement, white balance and gamma correction. Hence camera itself introduces some artifacts. So, it is necessary to record the video from a good quality camera.

A video can be taken as collection of images varying with time. The video-frames are first extracted from it and then frames can be saved in any image format based on choice of the researchers. This format can be jpg, bmp, png or any other. Now in further steps involved converting a colored image into Gray scale image as it simplifies the algorithm and reduces the computational requirements. Indeed, color may introduce some unnecessary information [11] could increase the amount of training data and performance time to achieve good performance. Color information [12] is sometimes of limited benefit in many applications like it doesn't help us identify important edges that can be done by noticing the high frequency changes that are easily perceived in Gray-scale images.

Then the Dimension Reduction algorithm is applied to reduce the computational complexity further. By doing so, some of the features present in the image may get reduced but the feature extraction algorithm should be so efficient that the acceptable accuracy is achieved. Then the frame is divided into overlapping or non-overlapping blocks depending upon the demand of the accuracy of algorithm. Now various features are extracted from these blocks using various scheme [13] for feature extraction like Scale Invariant Feature Transformation (SIFT), Speeded UP Robust Features (SURF), Mirror-reflection Invariant Feature Transformation (MIFT), etc.

Following step is to collect these features are then stored them in a feature matrix and are sorted. This brings all the similar blocks close to each other. Matching feature vector pairs are searched among the nearest neighbours using either threshold or various minimum distance calculations.

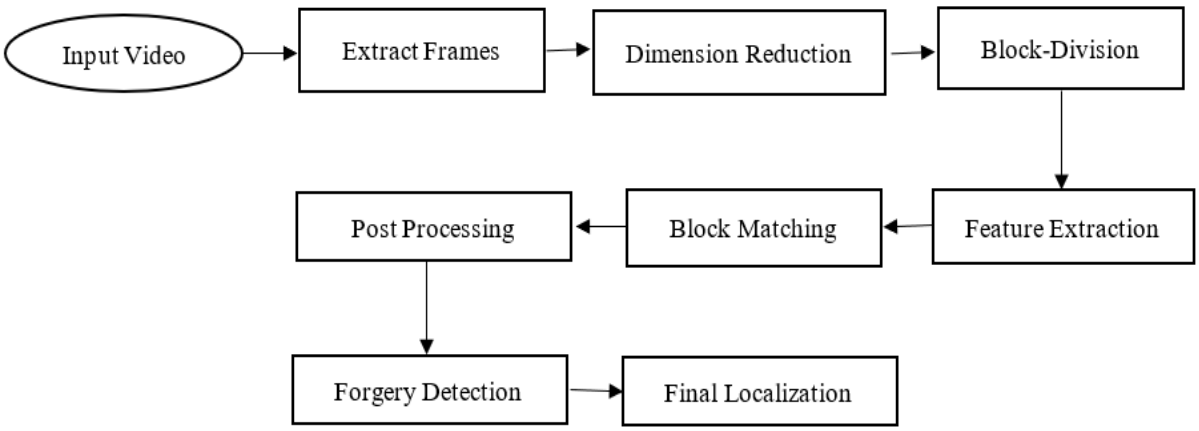


Figure 1.8 Flow chart of Video Tampering Detection Algorithm using Conventional Approaches

Then post-processing operations like mathematical morphological operations are performed for improvement of localization of the tampered region. The robustness to accuracy of the algorithm is being checked by various attacks. But these traditional approaches still need more accurate forgery detection results. So, deep learning approaches came into play.

1.6.2 Deep Learning Approaches

The comprehensive phases for detecting video forgery by the means of Deep Learning Approaches [14] has been described in detail. Figure 1.9 is an illustration for the same. First step two steps in this approach till dimension reduction are similar expect for the fact that it requires millions of frames for training and hence we need to load large video dataset unlike one video in traditional methods. Once the process of frame extraction has accomplished, we define Convolutional Network with number of hidden layers where each layer of the model extracts some significant features that allows the network to learn about the objects and scenes inside the video.

Next step is to train the input video for learning development. Once the data is trained, the extracted frames of test video sequence can be fed to the network that passes through the classifier. The classified output will be either authentic frames or forged frames on the basis of matched features of the frames from the test data with the trained data. Fully Connected (FC)

layer acts as a classifier for all those extracted features that allots each class a probability score on the basis of which the number of forged frames is calculated. Once the classification is done, the forgery can be located from classified forged frames using some applicable algorithms. There are various Deep Learning approaches followed by researchers to detect the counterfeits in doctored video sequences. These methods can be:

i. Convolutional Neural Networks (CNN) [15]
ii. Recurrent Neural Networks (RNN) [16]
iii. CNN with Auto encoders
iv. RNN with Auto encoders
v. Adaptive Neural Networks (ANN) [17]

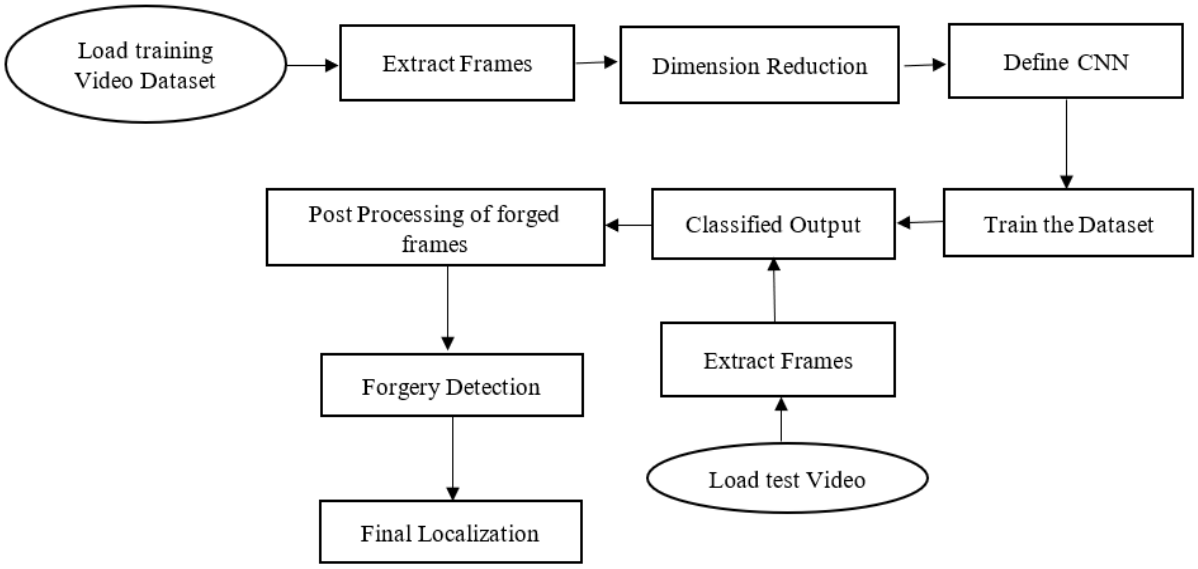


Figure 1.9 Flow chart showing the Tampering Detection Algorithm in video using Deep Learning (Convolutional Neural Network) Approaches

There has been noteworthy expansion in the extent of convolutional neural networks recently. The advantages of using DCNN for forgery detection over the conventional methods is given as follows:

i. Image classification and detection is quite rugged to falsifications such as extra artifacts due to camera lens, brightness problems during video acquisition, occlusions, etc. they are also shift invariant due to the presence fully connected layers in the network.

ii. It benefits of easier and better training opens the road for moving towards DCNN. The structure of standard neural network corresponding to that of DCNN will contain large number of parameters but DCNN has less parameters and hence less training time will be faced.

The performance of algorithm is checked with various parameters like Precision, Recall, F1 measures, True positive, True negative, etc. To check the robustness of the algorithm, different attacks are necessary to be implemented.

1.7 ATTACKS ON VIDEO

The optimization and efficiency of the algorithm for detecting video forgeries is checked by putting or adding some of the attacks to the video. These attacks [18] effect the internal parameters of the objects and hence create the counterfeit part more difficult to detect. But when the proposed algorithm calculates the acceptable accuracy even with the attacks, the algorithm is then considered to be reliable. There are many attacks used by researchers but the normally occurred attacks on visual content on Web with their parameters are described in the Table 1.1.

Table 1.1. Various Attacks used on Video Contents

Sr. No.	Attacks	Further Division	
1.	Compression	i.	You Tube (YT) Compression
		ii.	Double MPEG Compression
2.	Noise	i.	Pepper and Salt Noise
		ii.	Additive Gaussian Noise
3.	Changing Contrast	i.	Increasing contrast
		ii.	Decreasing contrast
4.	Frame Rate	i.	Slow motion video
		ii.	Fast Motion Video

1. *Compression Attacks:* Compression is an exercise of decreasing the number of bits i.e. zeros or ones. A video file typically can contain more than one type of compression. As it is clear that video is transmitted through a wireless channel as an electrical signal. These signals generally flow via cables, radio waves or micro waves, etc. But

the bandwidth spectrum is quite limited and hence compression of video is needed by reducing the file sizes, efficient use of bandwidth can be made.

i. *YT Compression:* You Tube (YT) platform are the fast-growing platform when it comes to interaction with world. But sometimes the video seuwnces are altered or its properties like brightness, picture quality or background are modified before uploading them to the platform. Hence the video gets compressed after modification. This compression is further increased when these are finally uploaded and re-saved. Hence the video content suffers much more compression then it actually had.

ii. *Double MPEG Compression:* MPEG codecs [19] have extensively been employed in numerous applications and video capturing devices. But the attackers can maliciously create some counterfeits in the contents of video sequence and most of these counterfeits experience recompression and hence double compression happens.

2. *Noise: T*he noise is usually studied in signal processing. But video forensics aims to remove noise from video. On the contrary, newly some investigators have remarkably made an effort to successfully utilize noise [20] for forgery detection rather than removing it.

   i. *Salt and Pepper Noise:* This type of noise usually is observed on images or frames of the video. This noise is produced by abrupt and sharp turbulences in the image signal. It is presented by lightly displaying white and black pixels. This type of noise can be reduced by median or a morphological filtering.

   ii. *Additive Gaussian Noise:* Major foundations of Additive Gaussian noise rise during capturing or acquisition process. Its example can be considered as sensor noise which is further produced by deprived illumination or sometimes due to high temperature. This type of noise can be decreased with a spatial filter.

3. *Changing Contrast:* To make the appearances in the movies less visual by changing its contrast and brightness levels. This may get worse when the video sequence is comprised of the objects with corners and edges and due to change in contrast

properties, they might not be visible properly. This variation can be done in two ways either decreasing the contrast value or increasing it and hence increasing the darkness in the scene. To remove this, some pre-processing operations are performed by researchers before actually feeding the video to the algorithm.

4. *Frame Rate:* Frame rate is the rate at which the [21] video frames are moving within a second. If a video is 30fps then it means thirty frames of the video are moving within in a second. Changing frame rate is also one type of attacks used by attackers.

    i. *Fast Motion Video:* It may appear abnormal to visual eye. Such objects in the movie will run faster than the average.

    ii. *Slow Motion Video:* This style is attained when the frame is acquired at a rate much sooner than played back but when repeated at regular speed, time seems to move more slowly.

The discussion of above all methods and numerous attacks helped to focus our main concentration in implementing a new and better technique for detecting inter-frame forged objects in video.

1.8 THESIS OUTLINE

This thesis focuses on detecting and localizing video inter-frame forgery by the use of Deep Convolutional Neural Networks (DCNN). It can be structured as follows:

Chapter 1 comprises of the detailed introduction of video forgery and its types. It also expresses about the necessity of video forgery detection and the numerous approaches to attain the same. It also gives a quick view about the common attacks used by researchers on video to examine the reliability of the algorithm.

Chapter 2 includes the comprehensive review on literature for video forgery detection techniques. It discusses the development made in the area of video forensics from the pre-mature time to the advancement in convolutional neural networks. It also describes how numerous video attacks effect the algorithmic efficiency.

Chapter 3 discusses about the generalized structure of CNN. It covers the entire mathematical study for deep convolutional neural networks that are used in proposed algorithm.

Chapter 4 covers the explanation about the suggested methodology. It briefs the approach used to expose and localize the counterfeit part in inter-frame doctored video and hence also displaying out the simulation outcomes with several performance parameters.

Chapter 5 sums up the recommended effort by comparing it with the conventional approaches and ends up with the advancements that can be made in future for more optimized technology.

CHAPTER 2
LITERATURE SURVEY

2.1 INTRODUCTION

Over the years, image forensics has come a long way in exposing tampering in visual content using simple approaches. Although many researchers have performed various algorithms for the image forgery detection but working on them becomes a challenging task when the disadvantages overcome its advantages. This chapter includes the precise techniques used till now to make better progress day by day.

2.2 REVIEW OF VIDEO FORGERY DETECTION METHODS

This section presents a wide-ranging and deeply analysed catalogue addressing the available literature in the field of blind video validation, with chief concentration on forgery detection. Many researchers have explored this field these approaches are attracting significance day by day. There are two main classes of tampering detection in video sequences, namely inter-frame (copied section from some other frame) video forgery detection and intra-frame (copied section from same frame) video forgery detection. This type of tampering attack can be intended at manipulating background that are static, interpolating or deleting anything in foreground.

Cheng *et. al.* [22,23] presented region tampering detection in video by using pixel intensity and its coherence investigation across spatio-temporal portions of the doubtful video. Some other reviewers [24] have cleared their idea on pixels to perceive spatio-temporal counterfeits in de-interlacing algorithms. Some authors have cracked tampering problems with newly introduced CNN framework. B Ravi Kiran *et. al.* proposed an unsupervised and semi-supervised technique [25] for irregularity detection by exploring those various specimens that utilize linear estimates by PCA and nonlinear estimates by auto encoders and deep neural networks. Their results were way better than the conventional approaches both in terms of accuracy and recall.

Following is the analysis on the literature framework designed for recognition of fake visual data. The literature framework has been detailed in three sections: literature survey on Conventional approaches, review on deep learning-based approaches and last section shows the brief study on various post-processing attacks used for checking robustness of the algorithm.

2.2.1 Standard Existing Techniques

With alteration in video, new relationship gets added in between the section of original frame and that of the newly inserted section which helps in forgery detection (CMFD). Replicated fragments can accomplish any shape and can exist at any position, so it turns out to be computationally an expensive job to locate all possible forged frames with the specific shapes and dimensions. Guo-Shiang Lin, *et. al.* learned this idea by introducing a scheme [26] for detecting frame duplication by finding the difference in histogram of two end-to-end frames. The similarity of the patches is evaluated by using block-based procedure that measure spatial correlation of respective frames between the fake clip and the original one. Wang *et. al.* [27] worked on this similarity impression. They defined a matrix on temporal correlation that symbolizes the similarities between all sets of frames in a sequence which are utilized to distinguish replicated frames from original ones in input video. This method was only able to detect static forgeries and for some video, accuracy wasn't appreciable. M. Mathai *et. al.* [28] used on arithmetical moment descriptors using prediction error concept computed for individual block from each frame of the video. They exposed out the forgery setting threshold for standardized cross correlation but still the true matching rate was 52% on an average.

Dai-Kyung Hyun *et. al.* used the analogous concept of correlation and applied it on the basis of noise. They presented a forensic method [29] to perceive modifications in surveillance video using sensor pattern noise (SPN). The algorithm recognizes partially altered regions by omitting the high frequency components in the given video using scaling factor on the basis of noise correlation. But this method also focused on detecting partial static forgeries. Jie Xu *et. al.* [30] explained correlation index on the basis of histogram intersection and achieved a recall of 90.4%. Their examination results showed that any tampering in the video will create a disturbance in the regularly occurring values of correlation coefficients. Syaiful Andy *et. al.* [31] changed the similarity index to hash values changes between two consecutive frames. They defined a hash function that takes an adjustable extent block of information as an input and static hash value so produced can be utilized to localize the replication.

Another technique [32-34] were projected to detect forged regions in a static-scene video using noise characteristics and noise feature residues. Pixels found to be fake distinguished on the basis of maximum a posteriori estimation (MAP) when the Noise Level Functions of the areas doesn't diverge with the residual video. Mandeep Kaur *et.al.* [35] showed noise is not consistent throughout the forged image. And this is what they used to expose the forgery using

HOG features and noise estimation. This estimation is done with the help of the features that were extracted from individual image block. In [36], projected an effort using HOG descriptors and compression characteristics. Video interfering discovery is completed by captivating all the benefits of HOG features. Julian Goodwin *et.al.* introduced a new algorithm [37] dependent on beneficial information fusion and conversion of descriptors in cross modal subspace for numerous residual features so extracted from inter-frame and intra-frame so formed blocks in video for the detection of fakery in digital video.

Some researchers estimated motion in the video using PSNR [38] and it helps the others to use this computation further to estimate doctored content. Hany Farid *et.al.* in [39] gave an idea 3-D ballistic motion inside the video usually observed in flights. This scheme assumes that the route of the object so chosen is influenced by gravity only and this needs that estimations of motion of the camera from its elements in the background and authors in [40] used video inpainting mechanism to target the moving objects in a video sequence and some [41] uses motion prediction error to find any insertion or deletion of frames in the video.

Some authors gave the idea for forgery detection using the MPEG compression methods. Some of them are given here. Hany Farid *et.al.* came up with a model [42] in which manipulation of the static and timely artifacts come to involve whenever a video goes through double MPEG compression. Also, temporally speaking, frames that can be copied-moved from one Group of Pictures to another, because of omission or addition of frames in the video, these leads to large motion approximation errors. A new methodology has been introduced [43] by Yuting Su *et.al.* to find interfering in video reliant on the supremacy of high frequency zone features that exists in the tampered sequences. These features present in the components of high-frequency of DCT coefficient chunks which are then drew so as to expose the visuals those were left by MPEG compression. And at final step, a recognition function is designed to explore for the coding-type modification in the uncertain MPEG-2 video given. Jieyuan Chen et.al. proposed [44] arithmetical descriptor related to macroblock mode (MBM) to recognize the double MPEG compression. This comprises of motion vector in P-frames and a macroblock type. MBM numerical features are captured out throughout multiple decoding measures for repeatedly compressed video that have same QS for numerous periods. The so introduced features are then united with the support vector machine (SVM) for classification of the single and double MPEG compression.

Wang and Farid gave a technique in 2009 [45] for knowing double quantization effects in some video. The consequences are collected from the double MPEG compression or when the two video sequences of different qualities are united and hence solving the problem of observing unseen arithmetic artifacts with double quantization can introduce and showed that these can be enumerated, restrained, and in addition to it, these are helpful in detecting tampering. Girija Chetty presented in their scheme that [46] the noise residue along with the quantization features can play a significant role in perceiving tampering. Conversion in cross-modal subspace in copy-move altered scenario demonstrated a suggestively enhanced scheme in terms detection accuracy.

Other different techniques have been proposed by numerous investigators. Some of them are based on Coherence properties, SIFT, various moments, CDN and FFTs. In [47], Coherence Based Forgery Detection (CBFD) based algorithm was recommended. This method first splits the given image into either non-overlapping or over-lapping blocks and then extracts feature vectors that has been lexicographically sorted. A proposed method in 2016 [48], a circular block that were overlapped is used to split up the tampered image into further overlapped circular blocks. The image features are then extracted by the Discrete Radial Harmonic Fourier Moments (DRHFMs) by using the overlapping block (circular) from the doubtful image. Euclidean distance and correlation coefficient are used to filter some of these features so as to erase the wrong matches. The Fast Fourier Transform (FFT) is considered for fast learning and detecting the tampered content from the video.

A method was proposed [49] in 2012 by taking fruitful analysis of Content Delivery Network (CDN) so as to improve scalability. With this, large amount of Internet video can be accessed in real-time application areas. A CDN-based Resource-Aware Scheduling (CRAS) algorithm plans the work efficiently in the *DFSP* depending on some parameters and these can be delay and computation load. A new approach [50] for detecting the counterfeits in video was suggested by making the use of the ghost shadow artifacts that can be detected accurately by irregularities in the moving foreground made to be segmented from the original frames in the video and such a moving person can be obtained from the static backgrounds because of frame differences and hence the video forgery is detected. Rui Ma *et.al* gave [51] a method based on MI-SIFT descriptors and this keeps the advantages of the standard SIFT and with an addition to this that these are invariant to mirror images and inverted images. MI-SIFT is accomplished by combining SIFT histogram bins at the slight expense of distinctiveness. Many investigators

[52-59] have utilized SIFT and MIFT features in order to improve accuracy by showing how MI-SIFT can be applied to mirror-like images that were abundant in real world. Chi-Man Pun *et.al.* in [60] proposed method in 2015 for detecting forgery that are copied and moved to some other part by using adaptive over-segmentation. This algorithm of Adaptive Over-Segmentation splits the input image into non-overlapping blocks and this process is done adaptively. Then, the extraction of feature points process is done. The small super pixels feature points act as blocks that merges the neighbouring blocks with some similarity of local color features into the feature blocks in order to generate the doctored regions. In 2010, a method [61] was proposed in which block matching procedures are used. Firstly, the division of the image is done, then singular value decomposition (SVD) for dimension reduction process for forming the feature matrix and then these are lexicographically sorted. Declaration of matched block is done only if correlation coefficient crosses the defined threshold.

As we know there is some relation among the frames itself and hence that can be considered in two ways i.e. either spatially or temporally. Researchers engaged their continuous time in this spatio-temporal context learning features and utilized them in predicting any falseness in the video. Confidence map [62] is utilized as a problem solver that fetches the past data of the target. To point out replicated regions [63, 64] in the temporal domain, the histogram variation between the adjacent frames is adopted. And further the similarity content is captured out by the use of block-based procedure that measures spatial relationship of corresponding frame. Another procedure uses 4-staged algorithm that [65] uses histogram difference of corresponding frames. Various authors [66-68] emphasized on observed challenges and fetched out prospects in the field blind video forensics by revising the prevailing literature in this field.

Apart from the traditional long and exhaustive approaches, some authors played their part in reducing the computational time and hence they took some help of machine learning area where machine can be automatically able to extract the features out of the frames and therefore reducing some continuous efforts of the user. In method [69], an algorithm in two stages has been recommended to locate the counterfeit in video sections precisely in the fake video. In order to form the suggested frame distorted indicator, the target video motion residuals are produced. The object-based alteration in frames use the descriptors extracted from frames which are firstly built. An innovative PES feature [70] was introduced from Discrete Wavelet

Transform (DWT) domain and are trained in Support Vector Machine (SVM) and further it is used for unknown video to expose forgery.

Table 2.1 Various Conventional Methods for video forgery detection with specified Gaps

Name (Year)	Technique	Gaps
Dai-Kyung Hyun [29] (2013)	Sensor Pattern Noise and Noise Correlation patterns	Detects forgeries in partial static surveillance video only
Wang [27] (2007)	Temporal frame-correlation	Valid only for static video scenes
M. Mathai [28] (2016)	Arithmetical moments using the concept of prediction error	Matching rate reached up to 52% only
Hany Farid [42] (2006)	Temporal correlation patterns on double MPEG compressed video	Results large motion approximation errors
Rui Ma [51] (2010)	Mirror-SIFT or MIFT by combining SIFT histogram bins distinctively	Doesn't work for shaky video

Deep learning has been utilized expansively in countless fields. Further in deep learning, Convolutional Neural Networks are developed to give the greatest precise outcomes in solving physical world difficulties.

2.2.2 Deep Learning Schemes

CNN is better than other deep learning methods in applications pertaining to computer vision and natural language processing because it mitigates most of the traditional problems. Various researchers have put their efforts to bring out the best work [71] done till today. Dan C. *et. al.* showed [72] how the CNN are efficient and highly flexible techniques for image classification. Another investigator in analyses the state-of-the-art that are based on deep learning approaches for video irregularity finding and classifies them on the basis of prototype and principles of recognition. Researchers in [73] exploit the property of adjacent correlation with the use

of auto regressive (AR) coefficients as the feature vector for identifying the location of digital forgery in a sample image. They trained the artificial neural network with 300 feature vectors and finally tested with additional 300 feature vectors. Others in [74] employs a convolutional neural network to acquire graded illustrations from the RGB color input to expose out both copy-move and spliced part. The first layer weights of network are initialized using high-pass filter (HPF) and made the residual maps assist to efficiently conquer the effect and collect the artifacts included by alteration. is Some researchers [75] used various codecs with strong first and next compression to detect all alterations like insertion and deletion of frames to or from the authentic video sequence.

Peisong He *et. al.* presented an CNN based agenda [76] that prepares pre-processing layer to detect repositioned frames. Furthermore, over-fitting is mitigated by adding convolutional in the proposed neural network. But they didn't clear out the video wise detection of forgeries. Ying Zhang *et. al.* develops [77] a Stacked Auto encoder prototype for training the network and hence making it learn all the composite features for respective patch. Also, they integrated the contextual data of each and every patch to make the accurate detection. Jiansheng Chen *et. al.* proposed an equivalent scheme [78] to work against the challenge of perceiving small-size images with median filtering by using the characteristics of median filtering with the help of an algorithm using convolutional neural networks (CNNs). Other investigators in [79] used ANN for images to detect forgery by training the network with ICA coefficients that were extracted in AR domain of the picture data. Similar method used for images with three different classifiers and are made to train image interpolation algorithm so proposed that also identifies the identify source cameras and expose digital counterfeits. Chengjiang Long *et. al.* proposed a parallel scheme [80] for frame drop detection by utilizing the spatial and temporal relationships inside a video segment. the errors in the network are supressed by periodically examining a confidence score. But the run-time was quite a big challenge.

Jonathan Long *et. al.* shaped a fully convolutional neural network [81] to which input of any size can be fed and it will yield output of corresponding size with effectual interpretation and learning. To make it fully convolutional, they described an architecture that combines semantic data from a deep layer with large amount of material from a shallow and a fine layer to yield precise and comprehensive segmentations. Liang *et. al.* proposed a methodology for exact object segmentation. They used invariance characteristics that makes Deep CNNs a better methodology for high level errands using Conditional Random Field (CRF).

Roman Sizyakin *et. al.* considered a technique [82] for fault recognition, which contains frame compensation steps followed by pixel values pre-processing ad further followed by categorization of all pixels having inconsistent standards using convolutional neural networks. Another researcher used projected a high resolution for multi-frame that gives every pixel information from consecutive frames in the given sequence. Simon *et. al.* [83] recommended a deep and fully convolutional neural network that estimates pairs of 1D kernels from the given input simultaneously for all pixels.

Not only forged regions but anomalies like blur recognition was carried out with neural network-based algorithms. Li Xu *et. al.* introduced a model [84] for image deconvolution by presenting a deep CNN to capture the features of blur degradation. The relation between traditional arrangements and a proposed architecture is calculated to describe deconvolution in contrast to various blur artifacts. Ye Yao *et. al.* proposed an algorithm [85] for object-based tampering detection for advanced codec encoded forged video. They defined a convolutional neural network (CNN) that inevitably captures the high dimensional features from so made image patches. A new absolute difference layer to lower down sequential redundancy between two video frames. But they only detected the forged frame but couldn't localize the particular area. To increase efficiency and accuracy of the algorithm, auto encoders played an important role. Dario D'Avino *et. al.* came up with an algorithm [86] that grounds on auto encoders and recurrent neural networks (RNN). A training stage permits the auto encoder to acquire an intrinsic prototype of the source. The huge reconstruction error supports in distinguishing the fake frames from the authentic ones.

2.3 REVIEW ON ATTACKS

The optimization and effectiveness of the algorithm for perceiving video counterfeits is checked by the accumulation of some attacks to the sequence. These attacks effect the internal parameters of the objects and hence generate the fake portion more tough to perceive. In method [87], L. D'Amiano, D. Cozzolino, G. Poggi and L. Verdoliva proposed an algorithm based on rotation-invariant features that are non-resilient. The dense-field matching over the whole video is done by modified fast algorithm for approx. nearest neighbours search and ad-hoc fast-post-processing was used for the region detection and to remove false matches. Omar Ismael *et. al.* reviewed [88] on different attacks to make a single umbrella for the other seekers to understand these attacks and their effect so as to develop a new methodology by keeping these effects in mind.

Lu Zheng *et. al.* proposed a [89] block-wise brightness variance descriptor (BBVD) scheme to detect the frame insertion in the video. They changed the number or frequency of frame insertion in the video sequence to test the efficiency of their algorithm. Shengda Chen *et. al.* [90] showed how object-based forged video are deteriorated throughout the compression method. They firstly decompressed the sequence and considered each frame in decompressed format as static frame. Qingzhong Liu *et. al.* [91] presented a novel method for disclosing counterfeits from the post-recompression attacks in JPEG. They used discrete transform features, collective learning used with the high feature dimensionality and avoid the general problem of regular classifiers i.e. overfitting under high descriptors dimensions.

Other authors [92] tested the forgery detector's efficiency by putting various attacks like contrast enhancement, compression, blurring attacks, frame rotation and frame cropping, random frame deletion and using increasing and decreasing he frame rate. Syaiful Andy *et. al.* come up with a technique [93] to perceive the replica of authentic frames in the sequence given by splitting the MJPEG video into frames with JPEG format. It was able to compute the hash value of frames and compare these values consecutively.

Other researchers [94] used the noise as a significant source to utilize them in detecting counterfeits parts of the frame. The MPEG video sequence with added compression noise brought the better results. The noise was extracted out spatially with the help of modified Huber Markov Random Field (HMRF) and transition probability matrices so obtained is used for classification of forged and authentic frames.

While other investigators utilized the similar concept of using the significant resources present in the video and are compatible with it i.e. noise, color and texture to find the forgery out of the video frames. The SIFT features are used as the main descriptors that are further fed to DWT which compresses the frames of the sequence the frame. They used optical flow as a pattern of deceptive object-motion, exteriors, and boundaries in a graphical scene produced by the relative motion amongst an observer and the scene.

2.4 MOTIVATION

Many methodologies have been followed to perceive several kinds of forgery in video. The above survey showed the progress of this investigation extent as well as state-of-the-art. Confidentiality of video is used in many critical areas like medical, army, courtrooms, journalistic photography, insurance claims, etc. Also, in today's scenario, photo and video

spoofs in our electronic-mail in-boxes, WhatsApp, Facebook or any other social media is easily noticeable which demanded a new field to detect such tampering. Video counterfeiting is becoming a challenge to every single individual. There is a vital and urgent requirement of bringing out more efficient video alteration recognition techniques. Digital video forensics can provide the authentication claim to such situations because of its power of revealing doctored content.

Talking of trail courtrooms, when the video is presented as an evidence of innocence, there is great requirement to assure the genuineness of the video for clean justice. Similarly, the video to be shown in news channels are also supposed to be true with no altered content. Multimedia data purchased from internet is demanded to be from the authenticated producer. Hence, Detection of tampering is today's-need.

Another motivation is to work through the field of Video Forensics by analysing its strengths and weaknesses and hence to enhance the parameters like True Positive Rate and accuracy further with most efficient techniques that would help other investigators or enthusiasts who are new to this field to understand it better.

2.5 RESEARCH OBJECTIVES

This thesis has the following objectives:

1. To implement an improved algorithm for object-based forged video content detection using Deep Convolutional Neural Networks.
2. To confirm the robustness of proposed algorithm on compression attack so as to achieve the results closer to that of un-compressed video.
3. To compare the evaluated parameters of proposed method with the existing results using REWIND and GRIP dataset.

CHAPTER 3
DEEP CONVOLUTIONAL NEURAL NETWORK

3.1 OVERVIEW

The concept of Deep Convolutional Neural Networks is a powerful thought in the field of artificial intelligence. Various large companies and organizations in this field are offering money to have automatically working robots which learns and recognize objects around the world and responds correspondingly. Computer Vision systems have travelled so far to achieve these successes in hand and neural networks in deep learning have helped to empower this further. A new machine learning area, called Convolutional Neural Networks, is a deep learning class and a part of artificial intelligence field with forward feed, is most commonly used these days for analyzation of computer vision systems. These are widely used for pattern recognition, identifying the object in the problem given or for classifying inputs. Hence deep learning allows various models to learn exactly how brain receives and perceives and understands the signals.

If we have a dataset that contains video-frames of dimension 114×114×3, it is really difficult to consider each channel of each pixel as an independent knowledgeable parameter because individual neuron will add many new extracted features to the model. This situation becomes more difficult as the size of image increases further. User's common way to handle this problem is down sample i.e. resize the images. But resizing images will lead to loss of information that may have proved profitable. With the introduction of deep learning techniques, a new solution to above problem has been presented. Convolutional Neural Networks takes full advantage of architectural encoded information within the image.

3.2 GENERALIZED STRUCTURE OF DCNN

Deep Convolutional Neural Networks (CNN) are encouraged from the biological operation of neural mapping inside brain. The connectivity neural design of DCNN resembles that of human visual cortex. Hence, a neural network can be considered as a system of artificial “neurons” that are interconnected and are able to exchange information whatever is fed to them like a neural organization of the brain. The elementary computational component of the brain is neuron where distinct neurons reciprocate to stimuli in limited space of visual field which are called as receptive fields. These fields further overlap and form a visual scene. In simple words,

whenever the data is input to the brain, the neuron inside it acquires and impulses conveys this information to succeeding neuron and this process further proceeds until the brain recognizes it. This flow of information from one branch to another is called training of neural network. The mathematical representation of single unit with 3 inputs [95] is given is Figure 3.1, where x_0, x_1, x_2 represents input, w_0, w_1, w_2 represents weights and b is bias.

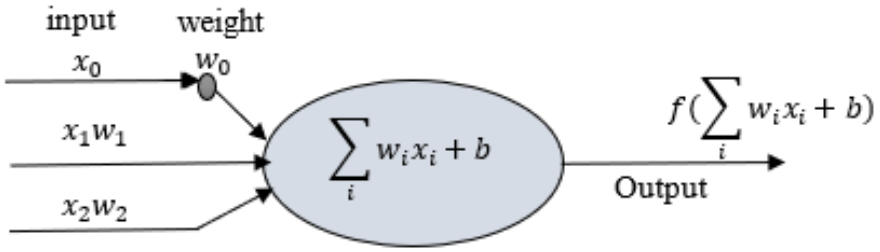


Figure 3.1 The mathematical prototype of single unit with 3 inputs [95]

To each neuron of the hidden layer, some weights and biases have been assigned that are shared at each point in the network. Weights are algebraic parameters that helps in determining how each neuron affects each other. Bias is a supplementary parameter which helps in adjusting the output. Bias is added to the input of each hidden layer and is not influenced by preceding hidden layer. The effect of weights and bias on the output is given by following eq. (3.1).

$$output = sum(weight \times input) + bias \tag{3.1}$$

Neurons of each layer can respond to various combinations of data input. The layers of the neural network are so built up that the first layer is able to detect a set of features in the input, the second layer detects features of features to learn more clearly and so on. The broad classification of layers in any convolutional neural network is input layer, convolutional layer, ReLU activation layer, pooling layer, fully connected, softmax and finally followed by output layer.

Each layer of the CNN or ConvNet converts the three-dimensional input volume to a three-dimensional output volume. Various pre-trained networks like VGG Net in combination with deep convolutional neural networks gives high performance output data when worked for detecting the doctored content in video. The general block diagram of a deep convolutional neural network showing each of its layers is shown in Figure 3.2.

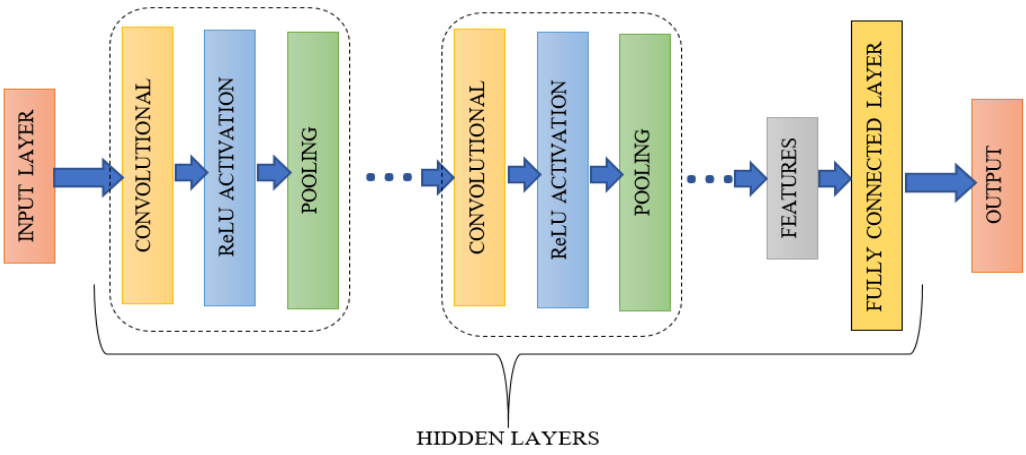


Figure 3.2. General Block Diagram of DCNN

There can be as many hidden layers as the researcher wants. More the number of hidden layers more will be the complexity of the algorithm but it also leads to higher accuracy. A typical CNN uses 5 to 25 distinct layers of pattern recognition [96]. The input is fed to the neural network where it undergoes through some hidden layers but the output may or may not be as desired. So, we need to adjust the weights of the hidden layers to detect features and do computations correctly. As we know, millions of pixels are spatially connected with each other form an image. Each pixel has its own importance. Also, label of image will have equal importance of the feature corresponding to that pixel is found to have its own importance.

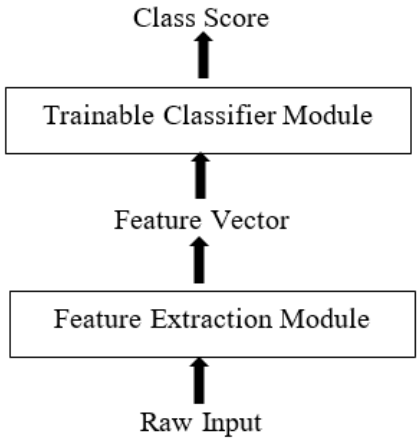


Figure 3.3 Traditional Classification using a Neural Network [96]

Whenever we have deal with high dimensional inputs like images or video frames at once, it becomes difficult to connect each neuron to all preceding neuron, so individual neuron is connected to a local region of input image called Local Receptive Fields. A smaller region of the input layer neurons is connected to the hidden layer neurons. These create a feature map

and uses the convolution to process the steps efficiently. Yann Le Cun [96] divided the process of classification using Neural Network in two parts and is shown in Figure 3.3. The first step is called feature extraction i.e. transforming the input in order to represent them by low-dimension vectors that are further useful in comparison and matching. These features are invariant w.r.t. to the distortions and transformations of the input patterns. Let us take an example that we need to classify a set of images that comprises of four different objects. In order to recognize these objects individually in each image using neural networks, we will label the images so as to get the training data for the network. This training dataset will allow the network to comprehend various features of objects in the picture and associate them with different corresponding categories. Now each layer from the network receives the information from the preceding layer and transforms and learns it and passes it on to the following layer. And the second step is classification which is a general-purpose procedure and is composed of trainable classifier. In a Convolutional Neural Networks, convolution layer act as feature extractor and are initialized by filter kernel weights pre-defined in training phase. Network trains by regulating weights to estimate accurate label of the given input. Figure 3.4 is a more descriptive and consistent way to see Convolutional Neural Network and its layers.

1. *Input Layer:* This is the layer which takes input of our proposed model. The sum of neurons present in this layer is equivalent to total features extracted from a video frame.

2. *Hidden Layer:* Second layer of the network is Hidden layer where input flows next. The number of hidden layers is not fixed. It varies with model and size of the data. Different numbers of neurons are associated with each hidden layer and usually bigger than that of features extracted. This layer contains the following sequence of layers:

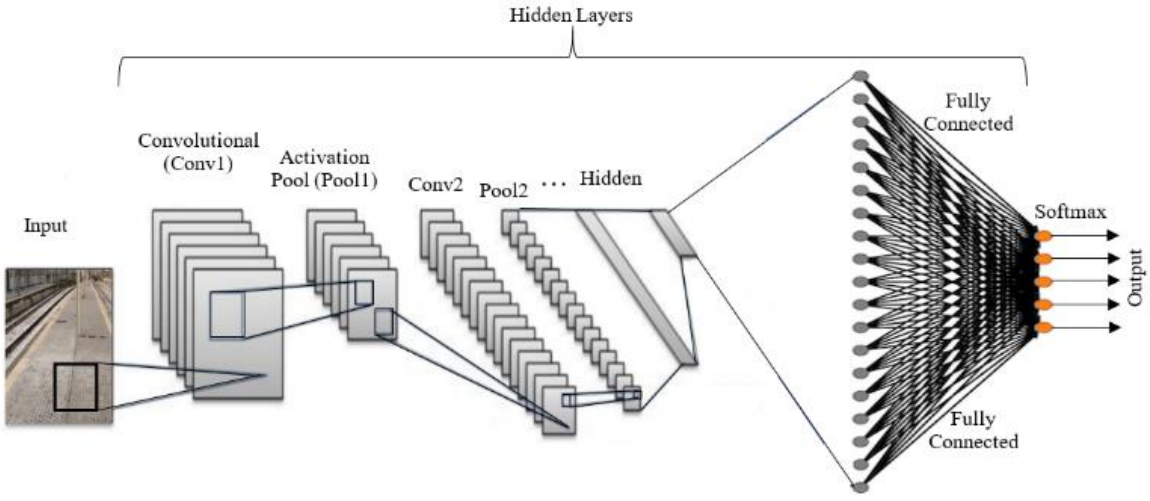


Figure 3.4 Basic division of a regular neural network layers

i. *Convolutional Layer*: This layer comprises of a set of filters with each one having a minor width, height and the depth equal to that of volume of input. The images are passed through these filters. The convolution starts from the top-left corner of the input and sliding to the right it performs dot product. This process is repeated until the bottom-right corner of the frame is reached. This process is shown in Figure 3.5 for one block of input. The whole output of the convolution of image with weight or filter matrix is shown in Figure 3.6. Mathematically, convolution of two variables can be done by following equation 3.2:

$$(x * y)[n] = \sum_{m=-\infty}^{\infty} x[m]\, y[n-m] = \sum_{m=-\infty}^{\infty} x[m-n]\, y[m] \tag{3.2}$$

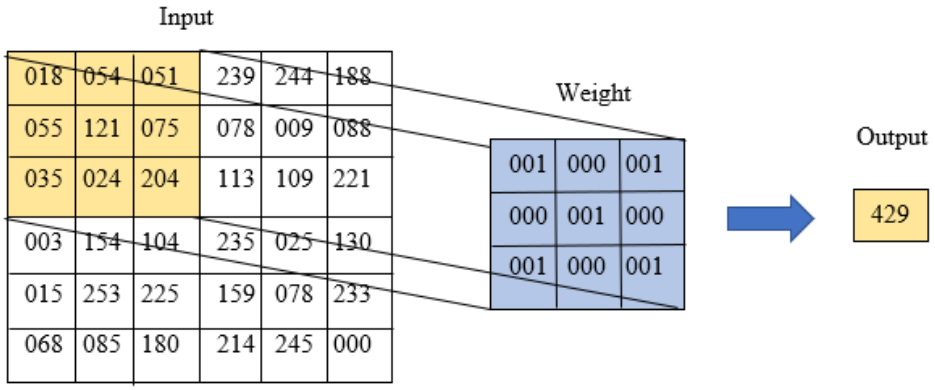


Figure 3.5 Graphical illustration of basic convolution process in CNN [97]

Taking an example of the image with 6×6 size. To extract the features, we decide the weight matrix and as for initialization, suppose this to be of size 3×3. This defined weight matrix will cover all the pixel at least once to give a convolved output. First of all, element-wise multiplication of weight matrix with the shown highlighted 3×3 fragment in Figure 3.5 will be computed and the result will be further added to get the value 429 shown above.

These pixel values are again utilized when the weight matrix completes its turn and slides to the right to perform the dot product again. The final convolved result is given in Figure 3.6 This is basically permitting the sharing of parameter in a CNN. Hence it is understood that the weight matrix so defined is behaving as a filter that extracts the specific data from the original input. Weights at one layer might extract the edge features and next layer weights gets deeper and may provide the information about the color present in the image. As the internal structure of CNN goes further deep, the

features extracted by the weight also becomes more complex. Thus, weights assist the network in estimating the output accurately.

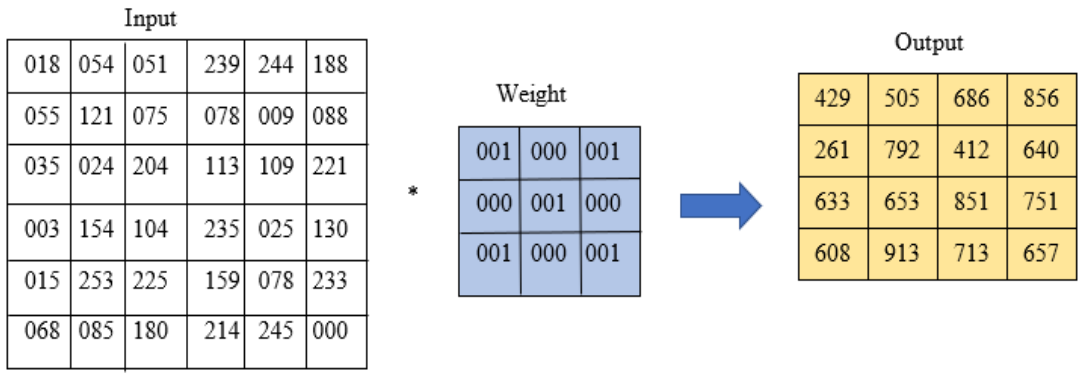


Figure 3.6 Example of convolution of image and weight matrix in CNN with stride one [97]

(a) Concept of Stride and Padding:

As the filter matrix slides, one pixel at a time to cover the whole image, this sliding parameter is to be defined by the user through stride. Stride controls how convolution of input with filter is going to happen. It is amount by which the filter is going to shift or hop. Normally, stride is increased for the less overlap between the receptive fields and to have smaller spatial extents. And as the value of stride increases, the size of image reduces with it. Hence, we pad zeros across the image to solve the problem. Figure 3.7 illustrates the convolutional results of padded image.

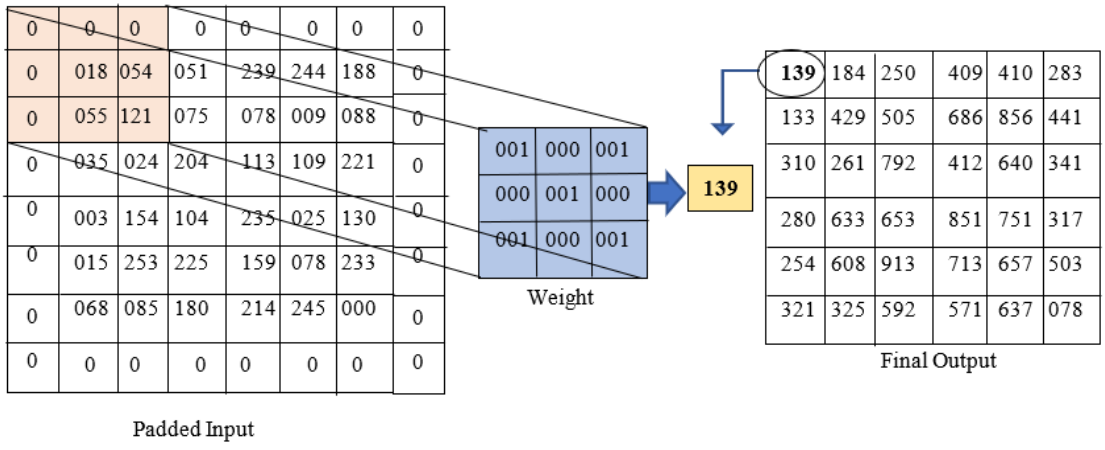


Figure 3.7 Example of convolution of padded input and filter in CNN with stride one [97]

Padding is done to keep the size of volume as it is in the start of the algorithm because we want to keep as much information as possible.

The size of padding can be computed from the following equation 3.3:

$$p = \frac{f_s - 1}{2} \tag{3.3}$$

The output size of the convolutional layer can be calculated from the equation 3.4:

$$O = \frac{i_s + f_s + p}{s} + 1 \tag{3.4}$$

where p is padding, f_s is size of filter, O is output size of convolutional layer, i_s is size of input and s is the stride.

ii. *ReLU Activation Layer:* This layer ensures element-wise activation of input by captivating the output from preceding layer and plotting it to the uppermost positive value. This is completed by using active functions like Rectified Linear Unit (ReLU). If out is the output value of the ReLU activation and in defines the input fed to it, then

$$out = (out)^+ = \max(0, in) \tag{3.5}$$

iii. *Max Pool Layer:* Pooling is another concept for activation with a difference that it reduces the dimensionality of ReLU output further. It aids in dropping memory necessity, fastens the computations and prevents over-fitting. Figure 3.8 shows the mathematical computation of Max-pool layer. If $out_{n,k}^m$ signifies neuron in the mth output activation map calculated over $(r \times r)$ section in mth input map $in_{n,k}^m$.

The output of this layer can be expressed as shown in equation 3.6:

$$out_{nk}^m = \max_{0 \le i,j < r} (in_{n.r+i,k.r+j}^m) \tag{3.6}$$

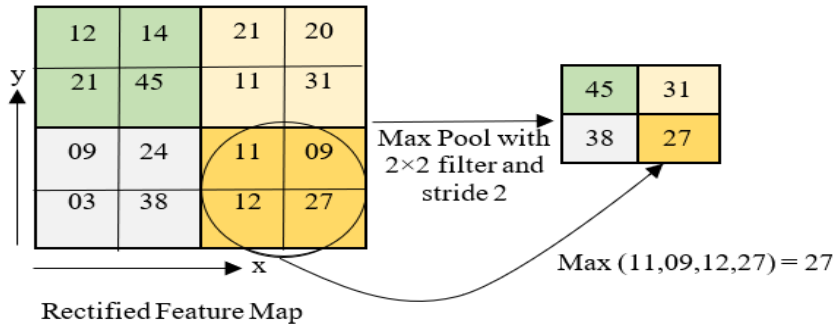


Figure 3.8 Basic operation of Max Pool Layer with 2×2 filters

iv. *Fully Connected Layer:* As the title suggests, each output neuron of the preceding layer $(x-1)$ is linked to following layer x. FC layer is consistent neural layer that captures the input from the preceding layers and calculates the class scores [98]. If N^{x-1} are the no. of neurons in preceding layer $(x-1)$, $w^x(m,n)$ signifies the weight from neuron m in $(x-1)$ layer to x layer and $b^x(n)$ is the bias of neuron n in layer x.The output of the n[th] neuron i.e. $(out)^x(n)$ in FC layer x is given by equation 3.7:

$$(out)^x(n) = f^n\left(\sum_{m=1}^{N^{x-1}} y^{x-1}(m).w^x(m,n) + b^x(n)\right) \tag{3.7}$$

3. *Output Layer:* The hidden layer further transfers the data to output layer and it passes through a logistic function called softmax [98]. It creates probability score of each classified data. Considering z classes in x training trials $\{(m1,n1),(m2,n2),\dots\}$ and m_i is the i_{th} training sample with y_i as the class then the probability $p(.)$ can be made on hypothesis $h_\alpha\ (m_i)$ as shown in equation 3.8.

$$h_\alpha\ (m_i) = \begin{pmatrix} p\left(n_i = 1 \,\middle|\, x_i; \alpha\right) \\ p\left(n_i = 2 \,\middle|\, x_i; \alpha\right) \\ \vdots \\ p\left(n_i = 3 \,\middle|\, x_i; \alpha\right) \end{pmatrix} = \frac{1}{\sum_{j=1}^{K} e^{\alpha^J{}_{j^{x_i}}}} \begin{pmatrix} e^{\alpha^J{}_{1^{x_i}}} \\ e^{\alpha^J{}_{2^{x_i}}} \\ . \\ . \\ . \\ e^{\alpha^J{}_{K^{x_i}}} \end{pmatrix} \tag{3.8}$$

The output layer comprises of loss function for error generation. Loss function in fully connected layer helps in calculating mean square loss. The gradient of error so computed is again back propagated to update the weights and biases if the training phase has more than one iteration.

After the above detailed study of internal architecture of CNN, the next step is to design a convolutional neural network specially for the detection of video forgery. Deciding the various parameters like input neurons, number of hidden layers, number of weights and biases for initialization, and the interfacing a different concept that detects and localize the forged region inside the frame is difficult process. Following section gives the basic steps for video forgery detection using Deep Convolutional Neural Networks.

3.3 DCNN FOR FORGERY DETECTION

This thesis mainly emphases its attention on an algorithm based on deep convolutional neural network that can detect the inter-frame duplication in the video and also localizes the externally added object in the video if any. The next section demonstrates the general algorithmic steps taken to detect the video forgery.

3.3.1 DCNN for Video Forgery Detection

The generalized steps for Video forgery detection using DCNN has been described in detail. Figure 3.9 shows the general flow diagram to detect forged frames and localize the forged region in those classified frames.

1. *Load Video and Extract Frames*- First step is to load the available video dataset. A video can be taken as collection of images varying with time. So, if it is the case of video, the video-frames are first extracted from it and then frames can be saved in any image format based on choice of the researchers. This format can be jpg, bmp, png or any other. The foremost necessity is to resize these frames into one similar dimension or enhancing the brightness of the images to bring out every little characteristic of the image. For example, if any frame has a dog inside it as an object, then defined DCNN must know about the features around the eyes, ears and hair of that dog.

2. *Deciding layers of Convolutional Network with required parameters*- Once the frames are extracted, we define Convolutional Network with all possible hidden layers depending on the accuracy desired. The layers can be increased up to the desire of the user but it should also be kept in mind that the more number of layers introduce complexity to the network. The DCNN model will take these frames as input and each layer of the model extracts some significant features. The work of each layer inside the DCNN has been properly discussed above.

3. *Train the Network*- In the training phase, the features of all the frames will be extracted and the network will learn accordingly. While the above process is called feature learning process.

4. *Testing phase*- Once the data is trained, its training variables can be saved in a file and used for with a pre-trained data. Now the test video can be uploaded. We again extract the frames of the test video before feeding it to the classifier. The classified output will

be either authentic frames or forged frames. This decision has been made on the basis of matching features of the frames from the test data with the trained data.

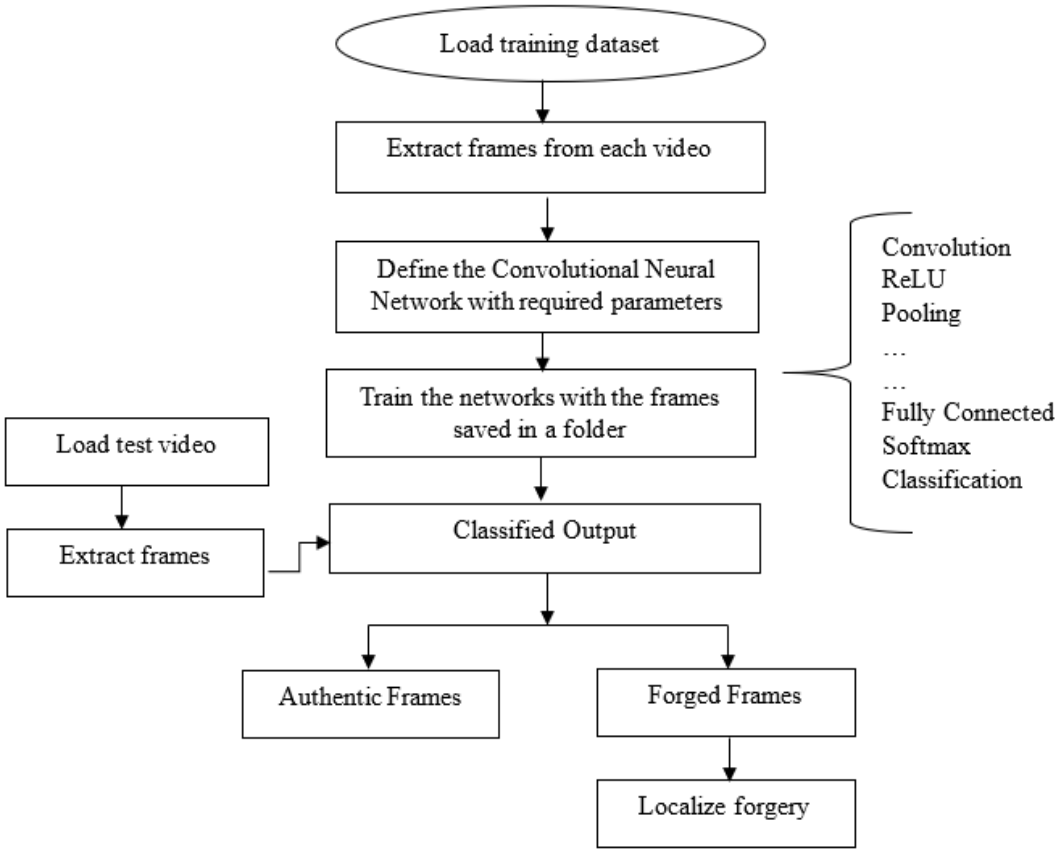


Figure 3.9 Generalized steps for detecting and locating video forgery using DCNN

5. *Classification*- Fully Connected (FC) layer acts as a classifier for all those extracted features. On the basis of this probability score, the network will classify the frames as forged or authentic. This layer sum weights from the last feature layers and to determine a precise output. SoftMax layer uses SoftMax function which finds out the probability of the huge set of values in the image matrix. This layer helps in distribution of a set of data category wise i.e. finds the probability distribution of K-possible values.

6. *Localizing the Forgery*- Once the classification is done, the forgery can be located from classified forged frames using some applicable algorithms.

Previously researchers apply their traditional methods on the entire video which actually comprises of both authentic and forged frames. Hence the algorithm also worked on the frames that weren't actually tampered with wastes the energy of the system. But Convolutional Neural Network allows to classify our video frames into authentic and forged and thus enables us to

perform the forgery localization on the classified forged frames only. Hence the consumption of energy will be less.

3.4 SUMMARY

Deep Convolutional Neural Networks (DCNN) are quite convincing and influential artificial neural network method. This chapter summarizes layers of DCNN and their working. A collection of various neurons in a particular order forms the network. Each individual neuron contains weights and biases for the accurate learning of the network. Once this data is learned by training the input, the network will be able to classify the input by allotting probability scores to each class. In case of detecting video forgery, there are only two classes: authentic and forged. These forged classified frames can be further processed to capture out the exact forged region. This chapter describes basic steps of video forgery detection with the help of a flowchart. It ended with a statement that DCNN reduces human efforts by applying forgery detection algorithm on only forged frames rather than entire video frames.

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 PROPOSED ALGORITHM FOR OBJECT FORGERY DETECTION IN VIDEO

The proposed algorithm works on the elementary relation between the video frames. The digital video sequence comprises of large volume of data in the spatial as well as temporal domain. The literature survey proves that the correlation of adjacent video-frames is quite high that will vary abruptly if some doctored content has been added to the original content of the video. The proposed algorithm can be broadly classified into:

i. *Training till Classification based upon Correlation theory:* In this step, the dataset with large set of video frames are input to the defined Convolutional Neural Network and are trained with maximum possible iterations depending upon the memory specifications of the machine. The last outcome of this step will be the spatial and temporal correlation-based Class-categorized data.

ii. *Testing the semantic segmented frames to localize forgery:* The doctored frames so classified by the first step will be segmented semantically and are fed to another network to train the forged frames with specific correlations and depending on the various abruptions observed, the region will be marked as tampered.

The graphical demonstration of the different stages of proposed algorithm has been shown in Figure 4.1 and detailed explanation of process used in proposed work has been explained in the next section.

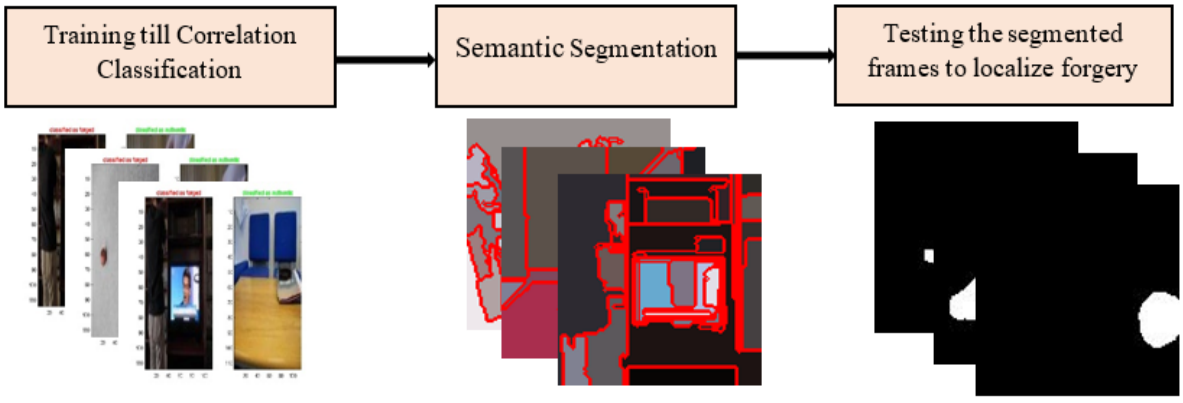


Figure 4.1 Broad Classification of proposed algorithm

Figure 4.2 shows the general steps of proposed algorithm for video forgery detection and localization using Deep Convolutional Neural Networks that uses the concept of semantic segmentation for localization of forged region.

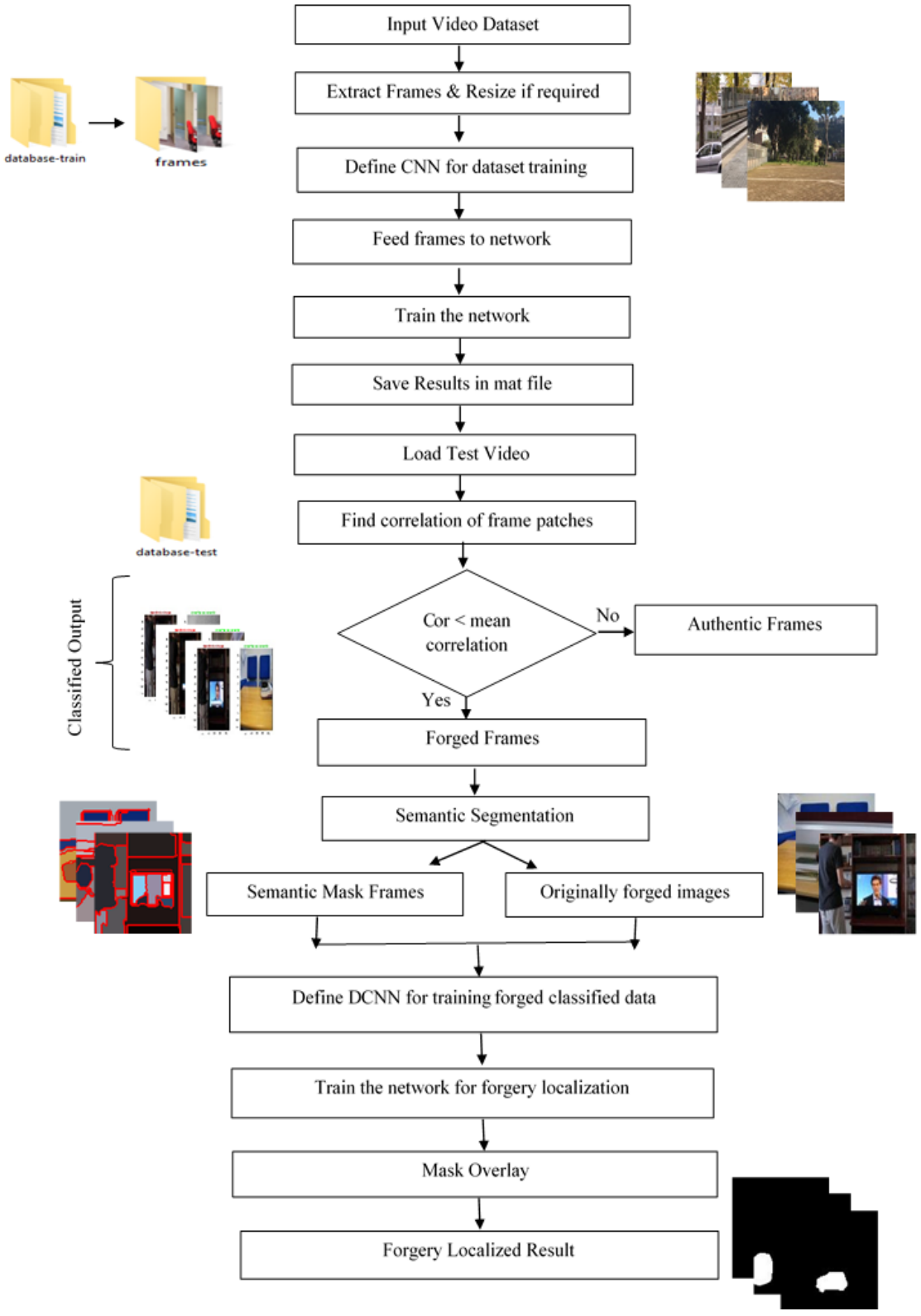


Figure 4.2 Flowchart of proposed algorithm

4.2 TRAINING AND CLASSIFICATION

Input Video Dataset and Extract Frames: The video sequence contains of set of consecutive frames. There is a similarity between these image frames with the solitary difference of moving objects. The resemblance can be from the analogous background.

These frames are gathered in a folder for feeding them to the Deep Convolutional Neural Network (DCNN). We need to assure that the frames of video sequences must have same aspect ratio hence we need to perform some of the required pre-processing steps. There can be as many pre-processing operations as the user wishes. Like resizing the frames to equal aspect-ratio, making dpi of each frame equal and if any of the video has been shot in darkness, then those are required to be enhanced to capture out the important features from the frame. For example, if any frame has a horse inside it as an object, then defined DCNN must know about the features around the eyes, ears and hair of that horse. We performed pre-processing operations to have equal dpi and equal dimension of the frames for fast computation. The DCNN model will take these frames as input and each layer of the model extracts some significant features whose complexity keeps on growing with each consecutive hidden layer.

4.2.1 Defining parameters and layers of Deep Convolutional Neural Network

Now, we define layers of the Deep Convolutional Neural Network (DCNN). The 12-layered CNN configuration is described in the following Figure 4.3. Input layer takes the video frames stored in a folder and grasp the raw pixels of the frames. Our video frame is 114×114×3 i.e. 114 wide, 114 in height, and contains 3 color RGB channels. Our input is Zero Centre normalized as it is always appropriate to pad input volume round the border with zeros to regulate the spatial extent of the output volume. Input normalization is significant as it guarantees that each pixel has analogous data distribution. Normalization is achieved by deducting the mean value from each pixel and dividing the outcome by standard deviation.

To each neuron of the hidden layer, some weights and biases have been assigned that are shared at each point in the feature map and hence each feature map has its own weight. As we know, weights are algebraic parameters that helps in determining how each neuron affects each other. These are given by:

$$\text{Weight} = D_p \times W_k \times H_k \tag{4.1}$$

Where D_p is preceding layer descriptor, W_k is Kernel Width and H_k is Kernel Height.

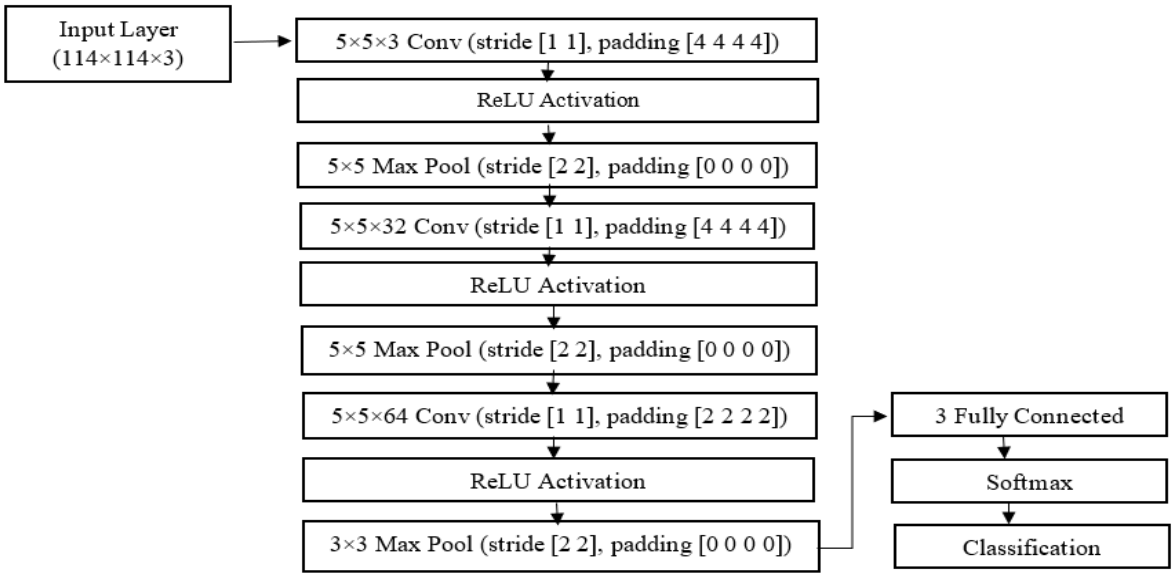


Figure 4.3 Convolutional Network Used for training dataset

32 high-pass filters with 5×5×3 size have been utilized for initialization of weights, in the first convolutional layer. Once we define the size of filters, we need to take care of the stride and padding. The first convolutional layer is defined with stride and padding [1 1], [4 4 4 4] respectively and this keep on changing with the following. Stride controls how convolution of input with filter is going to happen. It is amount by which the filter is going to shift or hop. In our case, the stride is [1 1], hence the elementary convolution will take place i.e. filter will shift by one value. Normally, stride is increased for the less overlap between the receptive fields and to have smaller spatial extents.

Padding is done to keep the size of volume as it is in the start of the algorithm because we want to keep as much information as possible. If we apply 5×5×3 filter to 114×114×3 input, the output volume will decrease and hence zero padding to that layer. The size of padding can be computed from the following equation 4.2:

$$p = \frac{f_s - 1}{2} \tag{4.2}$$

The output size of the convolutional layer can be calculated from the equation 4.3:

$$O = \frac{i_s + f_s + p}{s} + 1 \tag{4.3}$$

where p is padding, f_s is size of filter, O is output size of convolutional layer, i_s is size of input and s is the stride. The filtered matrix produce from this layer is next fed to ReLU activation and Max Pooling.

Max Pooling with [2 2] stride and [0 0 0 0] padding have been used in further layers to reduce the complexity of the feature matrix. If W and H denotes the width and height and s is the stride, then W_{fm} and H_{fm} are width and height of feature matrix respectively can be calculated by following equations:

$$W_{fm} = (W_{image} - W_{filter})/s + 1 \quad (4.4)$$

$$H_{fm} = (H_{image} - H_{filter})s + 1 \quad (4.5)$$

To monitor the continuous performance, some of the training options are defined in the network. We define SGDM (Stochastic Gradient Descent with Momentum) optimizer and set its value to 0.9 to see if the training improves. The Initial Learn Rate is set at 0.0. this valued decides the speed of training. If it is too low, the training speed also reduces. Learn Rate Schedule is defined to regulate learning rate throughout training period by dropping its value in accordance with default schedule. The Learn Rate Drop Factor is 0.5 and Learn Rate Drop Period is 10, L2 Regularization is 0.004 and Mini Batch Size is set at 100 with 1000 Max Epochs. Our next task is to classify the data.

4.2.2 Correlation Computation and Classification

The video are made of analogous frames with the solitary difference of some moving objects. So, it can be considered that there is similarity between the consecutive frames because of comparable background. This similarity can be considered of two types i.e. spatial correlation and temporal correlation. Spatial correlation is the similarity index between the two patches of two different frames. The matches between$n_i$non-overlapping blocks of first frame and n_j overlapping blocks of second frame are obtained and the correlation between these patches is captured in a matrix $n_i \times n_j$ called symmetric matrix where n is the total number of blocks and i^{th} and j^{th} are position of block. This can be measured by finding correlation between two consecutive frames. The correlation between the two arrays X and Y can be measured by using following eq. 4.6.

$$Cor(X,Y) = \frac{\sum_m \sum_n (X_{mn} - X_{mean})(Y_{mn} - Y_{mean})}{\sqrt{(\sum_m \sum_n (X_{mn} - X_{mean})^2)(\sum_m \sum_n (Y_{mn} - Y_{mean})^2)}} \quad (4.6)$$

The above criteria have been explained in following proposed algorithm for the classification of the frames as authentic and forged.

Table 4.1 Algorithm for classification of frames based on spatial and temporal correlation

1. Read video sequence $v(x, y, t)$ using VideoReader
2. Extract frames img using readFrame
3. % Let NF: length of file list that contains frames
4. % Let Cor_{mean_t}: threshold for mean temporal correlation and feed it to neural network
5. % Let Cor_{mean_s}: threshold for mean spatial correlation and feed it to neural network
6. % Let Cor_t: temporal correlation of each frame
7. % Let Cor_s: spatial correlation of individual frame
8. For $i = 1 : NF$
9. Construct Cor_t i.e. temporal correlation
10. For $j = 1 : (NF - 1)$
11. Construct Cor_s i.e. spatial correlation
12. If $(Cor_s < Cor_{mean_s})$ && $(Cor_t < Cor_{mean_t})$
13. Display error: msgbox ('Forged Frames', 'error')
14. Save forged frames in another folder: imwrite (img2(n)) %Classification

Whereas temporal correlation symbolizes the correlations between X_n and Y_n frames in a given sequence can be calculated by following eq. 4.7.

$$R_{xy}(m) = E\{X_{n+m} \; Y_{*n}\} = E\{X_n \; Y_{*n-m}\} \tag{4.7}$$

This similarity index so calculated for both spatial and temporal locations of patches and frames respectively will be utilized for categorizing the authentic and forged frames. The correlation coefficient is considered as a degree of similarity and the decision is made on the basis of following expression:

$$if \begin{cases} Cor > Cor_{avg}, \text{frame is authentic} \\ Cor < Cor_{avg}, \text{frame is forged} \end{cases} \tag{4.8}$$

4.2.3 Semantic Segmentation

Once the frames are classified, the forged frames can be fed to another neural network for testing the localization of the counterfeit region. The frames will be now first segmented semantically i.e. into meaningful fragments and the network will be trained with particular

color and any abnormality in the pasted region will be considered as the region with counterfeits. Segmentation of images means performing partition of a frame into numerous fragments with the knowledge of which fragment represents what object in the frame. Segmenting an image semantically means understanding it at pixel level i.e. assigning each pixel an object class.

Hence, the objects can be easily detected and classified. Every object in the frame will be first assigned a particular color for semantic segmentation and the neural network will be made to train with these segmented frames.

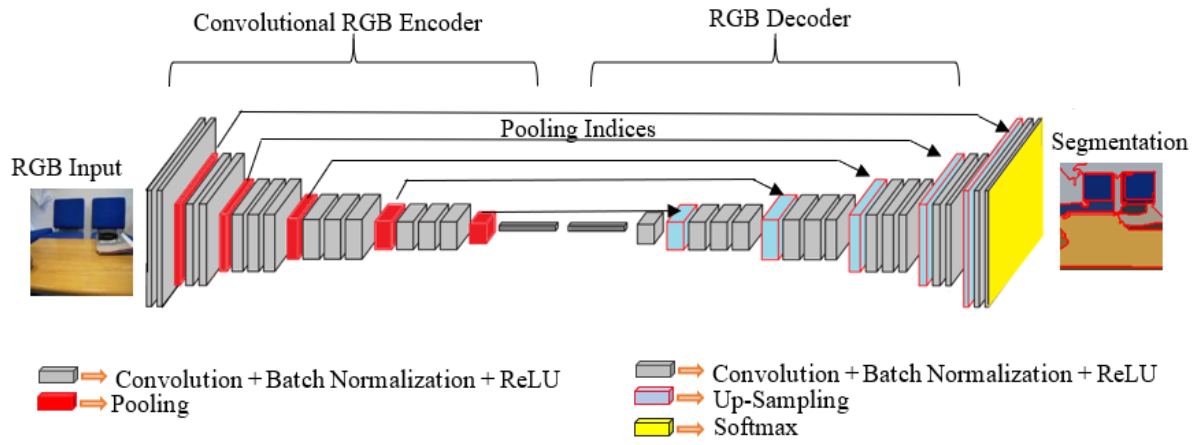


Figure 4.4 Video frame on the left side and its semantic segmented frame on the right side

VGG-16, a MATLAB supported model helps in achieving this segmentation with 90% above accuracy by using set of convolutional layers with large number of minor receptive fields in the preceding layers. Basic division of its model is RGB encoder and Decoder. Convolutional RGB encoder progressively decreases the dimension spatially with further pooling layers and RGB decoder does exactly the opposite i.e. gradually mends the object particulars spatially. We defined a 90-layered neural network for training in which we added batch-normalization after every 1st layer of Convolution. The input layer will be normalized by regulating activations and it correspondingly permits each layer to automatically learn slightly additional info without depending on other layers. To acquire higher steadiness of network, it normalizes the yield of a preceding activation layer by deducting the mean value and isolating it by the standard deviation.

In this CNN, first 45 layers comprises of set of Convolution, Batch normalization, ReLU and

Max Pool layers. The next 45 layers contain the set decoder layers that allows the network to understand the two categories of frames i.e. ‘authentic’ and ‘forged.’

4.3 FINAL TESTING AND FORGERY LOCALIZATION

The frames that were classified forged in the first training phase along with the corresponding semantically segmented frames are trained again through the model so that the CNN learns respective allotment of the color. Any region that is copied from some other image will have some abnormalities. The forged classified frame and its semantic segmentation will be overlayed using Deep CNN and the abnormalities so produced will create a binary segmented mask that will further help in distinguishing the forged region. The better and brief graphical illustration of the forgery localization phase has been given in Figure 4.5.

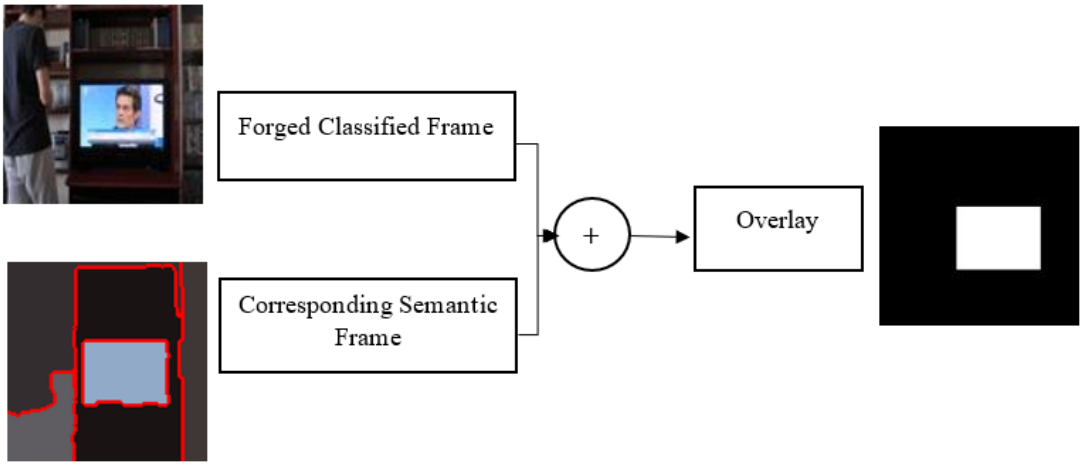


Figure 4.5 Generalized pictorial representation of localizing forgery in the proposed algorithm

4.4 VIDEO DATABASE

The neural network is trained with millions of images that helps it to learn features of objects within the frame. The proposed algorithm also uses the pre-defined VGG network that has been trained with millions of images to semantically segment the data. The dataset used in thesis for video forgery analysis are provided by Reverse Engineering of Audio-Visual Content Data (REWIND) [4] and Image Processing Research Group (GRIP) [3].

4.4.1 REWIND Dataset

This video dataset comprises of 20 video sequences with 10 forged and 10 authentic forming a total of 6950 frames. Individual video sequence with 320×240-pixel resolution has a frame

rate of 30 fps. Authentic video sequences have been captured with low-end devices hence compressed either in MJPEG or H.264 codecs. Doctored video sequences are uncompressed 24-bit RGB. All the sequences with same standard specifications are saved in uncompressed format i.e. YUV files. Some authentic video are given by Surrey University Library for Forensic Analysis (SULFA) database [16] and the attached mat files assist as ground truth. The generalized description about the dataset provided by the Research Group itself has been given in the following Table 4.2.

Table 4.2 Description of REWIND Dataset [4]

Video	Resolution	Frame Rate
Video_01	320×240	30 fps
Video_02	320×240	30 fps
Video_03	320×240	30 fps
Video_04	320×240	30 fps
Video_05	320×240	30 fps
Video_06	320×240	30 fps
Video_07	320×240	30 fps
Video_08	320×240	30 fps
Video_09	320×240	30 fps
Video_10	320×240	30 fps

4.4.2 GRIP Dataset

This video dataset comprises of total. 40 video sequence which can be divided as:

1. 10 authentic video sequences
2. 10 forged video
3. 10 YouTube compressed video
4. 10 binary masks that serves as Ground truth.

Each video sequence has 1280×720 resolution. These are captured from different source cameras the description of which is given in Table 4.3. Each video sequence contains different number of frames. The exact description about the dataset provided by GRIP Research Group itself has been given in the following Table 4.3.

Table 4.3 Description of GRIP Dataset [3]

Video	Resolution	Source Camera
V1_Tank	1280×720	Nokia Lumia520
V2_Man	1280×720	iPhone7
V3_Cat	1280×720	Huawei P7mini
V4_Helicopter	1280×720	iPhone5
V5_Hen	1280×720	Huawei P9plus
V6_Lion	1280×720	Samsung
V7_UFO	1280×720	MotoG
V8_Tree	1280×720	Huawei P8lite
V9_Girl	1280×720	Samsung J5
V10_Dog	1280×720	Nokia Lumia 520

4.5 MACHINE CONFIGURATION

The algorithm is executed on software MATLAB 2018a. our algorithm requires numerous toolboxes to run on. The description of each toolbox is given as follows:

Computer Vision System toolbox: This toolbox offers various algorithms and functions for designing and simulating computer vision systems.

Neural Network Toolbox: It offers pre-trained models to train and simulate all deep neural networks. It helps in performing classification and clustering various systems.

Parallel Computing Toolbox: This toolbox allows the user to utilize multicore desktops by performing work on locally running engines and hence, single machine can use high processing specifications. In other words, same code can be executed in parallel on a cluster of GPUs.

Neural Network Toolbox Model for VGG-16 Network: It is a pre-trained CNN with approx. 1.2 million of images from the ImageNet. This toolbox is capable of classifying images into 1000 classes. For example, car, bus, keyboard, mouse, cat, dog and many other classes. The prototype has made to learn quite rich feature descriptions for a huge variety of images.

The basic machine specifications, on which the proposed algorithm has been executed, is given in the following Table 4.4:

Table 4.4 Configuration of Machine used for implementation of algorithm

Sr. No.	Parameter	Machine Configuration
1.	Windows	10
2.	Processor	Intel Core-i7-7700T CPU @ 2.9 GHz
3.	OS	64-bit
4.	System Type	X64-based processor
5.	RAM	16.0 GB
6.	GPU	NVIDIA GeForce GTX 1070
7.	MATLAB version	18a

4.6 EXPERIMENTAL RESULTS

The algorithm is tested for two different datasets and it is able to recognize and localize the forged region very accurately and also evalutes the results for YT compressed video given with GRIP dataset to confirm the robustness of the proposed method. The algorithm is firstly checked for the whole trainable data and hence classifying each frames of every video. These results for two category classification shows a total of 936 frames as forged frames out of total 3176 test frames for REWIND dataset. And 839 forged frames out of 2950 test frames for GRIP dataset.

4.6.1 Classification and Segmentation Results

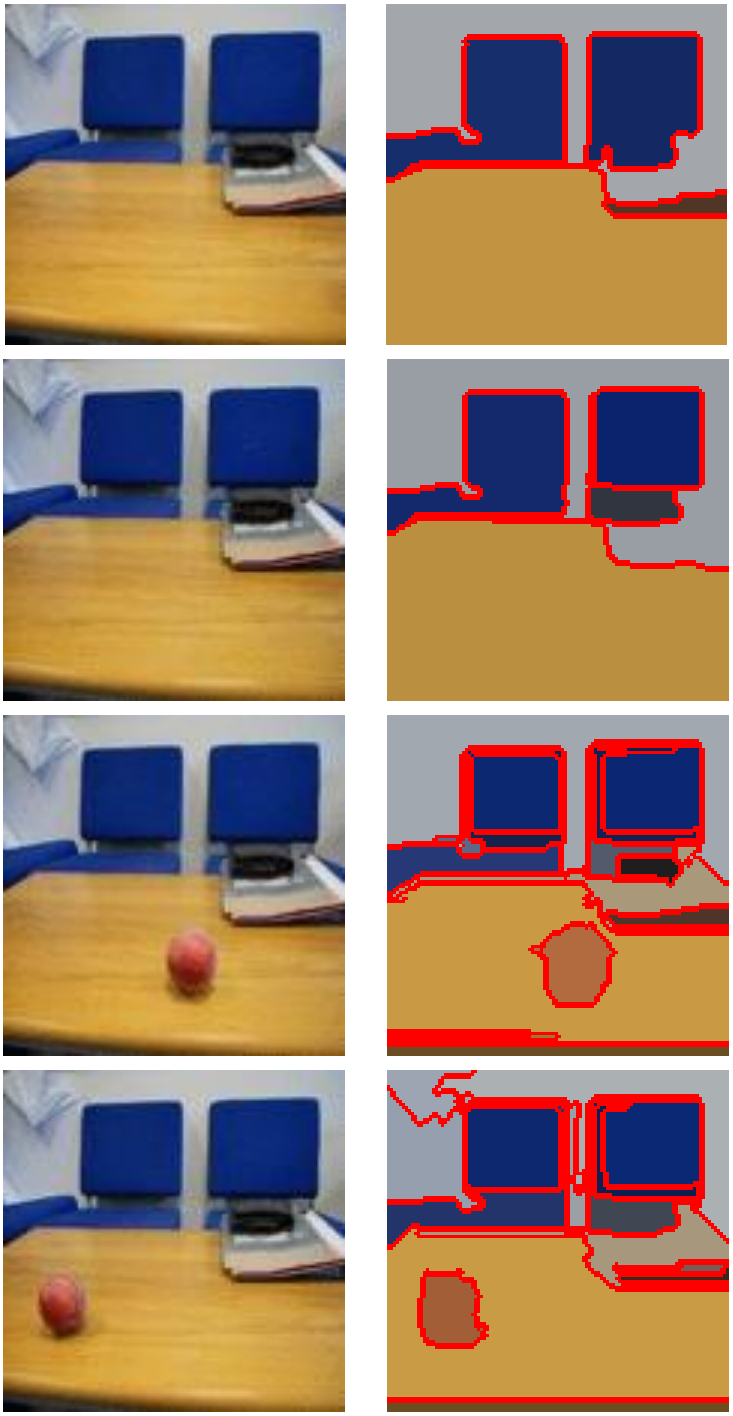
The suggested algorithm successfully classifies the input frames into two different categories:

1. ‘Authentic’
2. ‘Forged’

As it is clear from the above given explanation of proposed algorithm in Table 4.1, the authentic frames will be distinguished from the forged ones by the sudden changes in the both spatial and temporal correlation factor. Following results shows the truthful proofs of classification results of the proposed methods. These categorized frames are then saved in a different folder as forged and are fed to the newly defined neural network for performing segmentation semantically. Figure 4.6 shows the correlation based classified results from the input video frames whereas the next Figure 4.7 shows the segmented results of the so classified forged frames.

Figure 4.6 Correlation based classified results. The image on the left-side shows forged frame with title in red color and right-side image is authentic frame with green title

These frames are further saved in the different folders and hence forged frames are extracted separately for processing out the tampered region in those frames. The classified frames are now semantically segmented i.e. each region is assigned with different color to read out abnormalities. The results for segmented frames are given in the following Figure 4.7.



(a) Forged frames from Video_02 (b) Segmented frames from Video_02

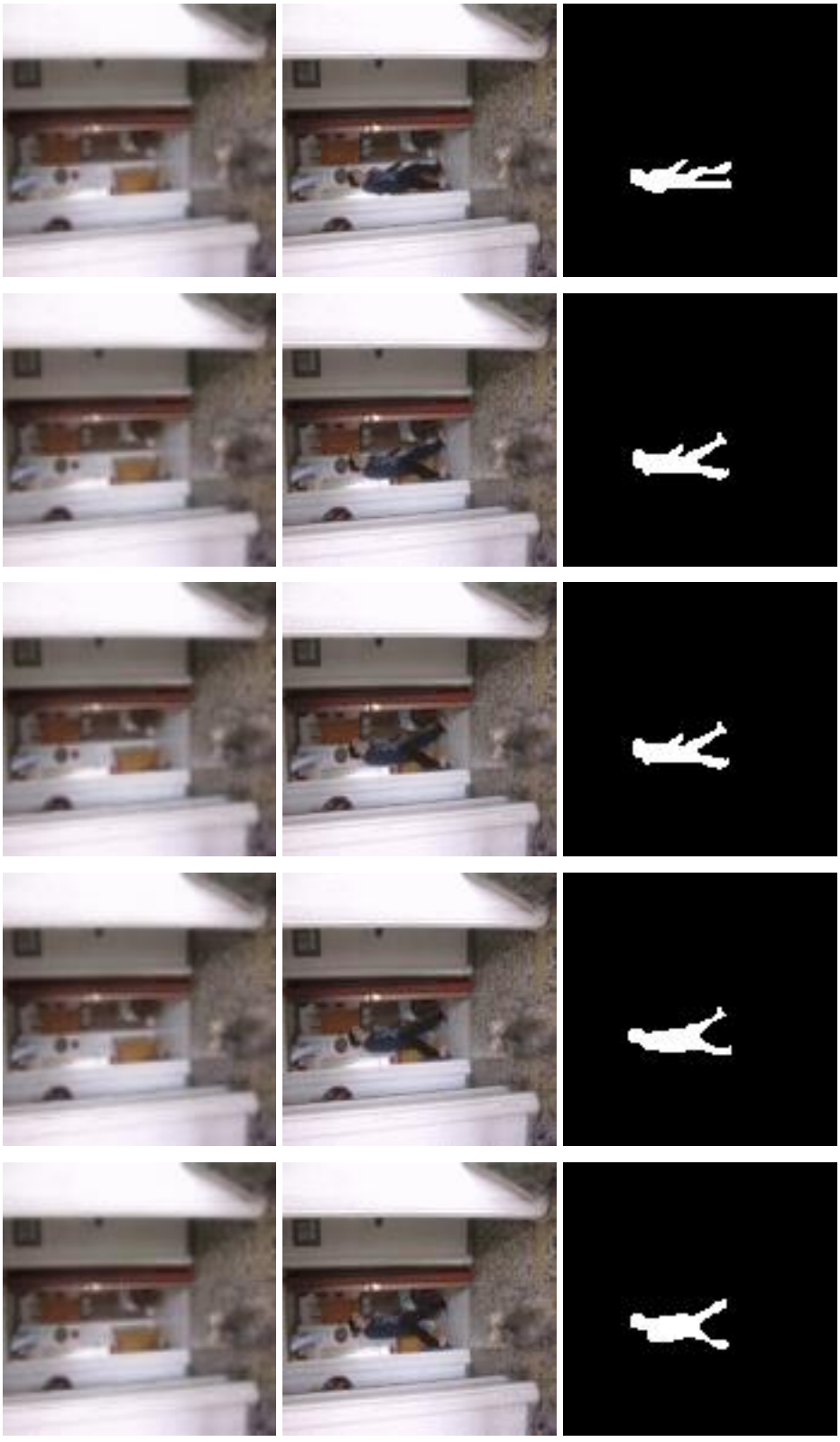
Figure 4.7 Correlation based classified results. The image on the left-side shows forged frame with title in red color and right-side image is authentic frame with green title

4.6.2 Results on REWIND dataset

The forgery detected results of video 5 and 10 from REWIND dataset are given in Figure 4.8 and 4.9:

(b) Original Frames (a) Forged Frames (c) Final Results

Figure 4.8 Results from video 5 from dataset (a) shows the authentic frames of the video (b) forged frames where the TV screen inside the frame has been in-distinguishably changed with some other (c) shows final forgery detected results with forged region shown specifically by white color

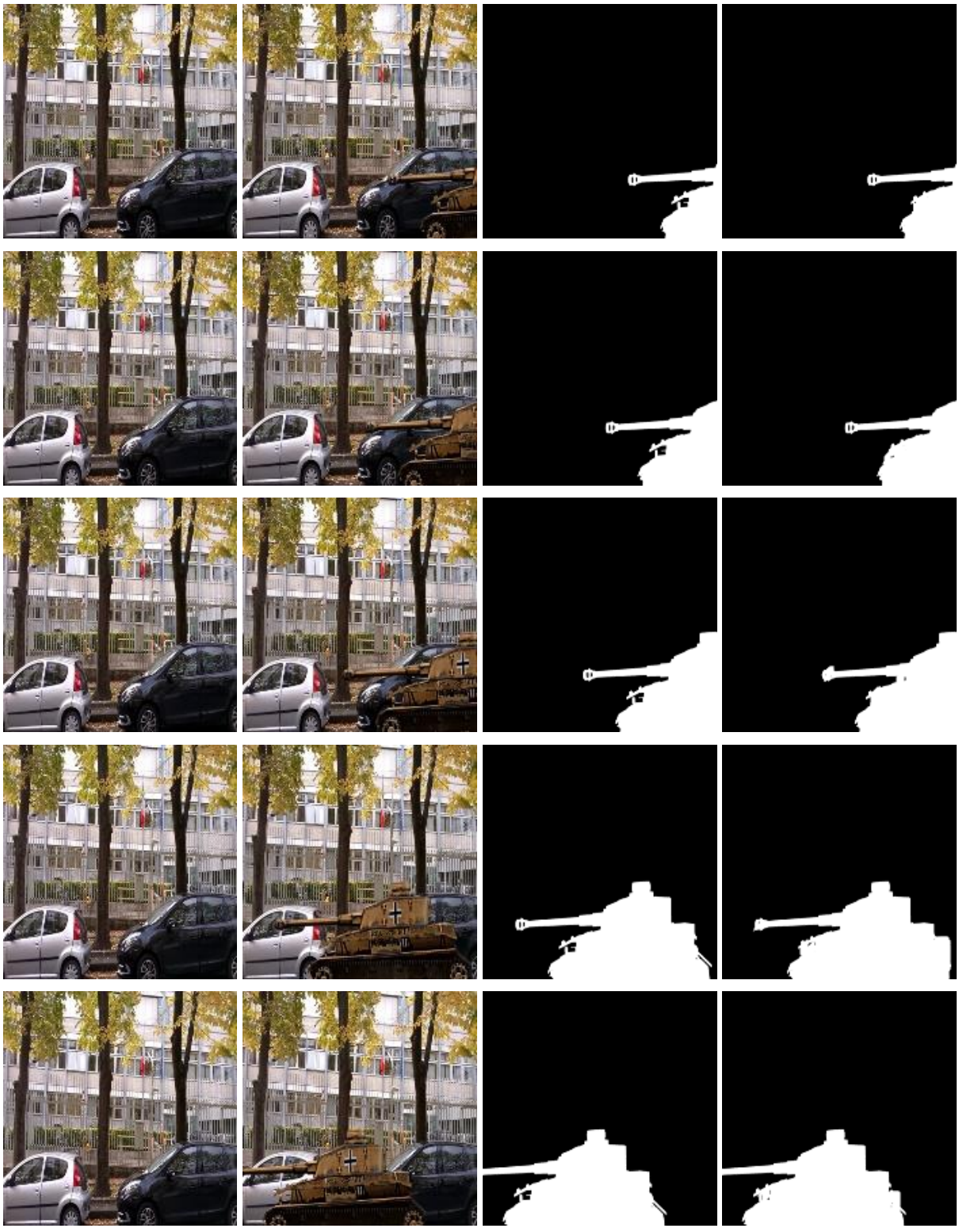


(c) Original Frames (a) Forged Frames (b) Final Results

Figure 4.9 Results from video 10 from dataset (a) shows the authentic frames of the video (b) forged frames contains the walking person (c) final results with forged region shown specifically by white color

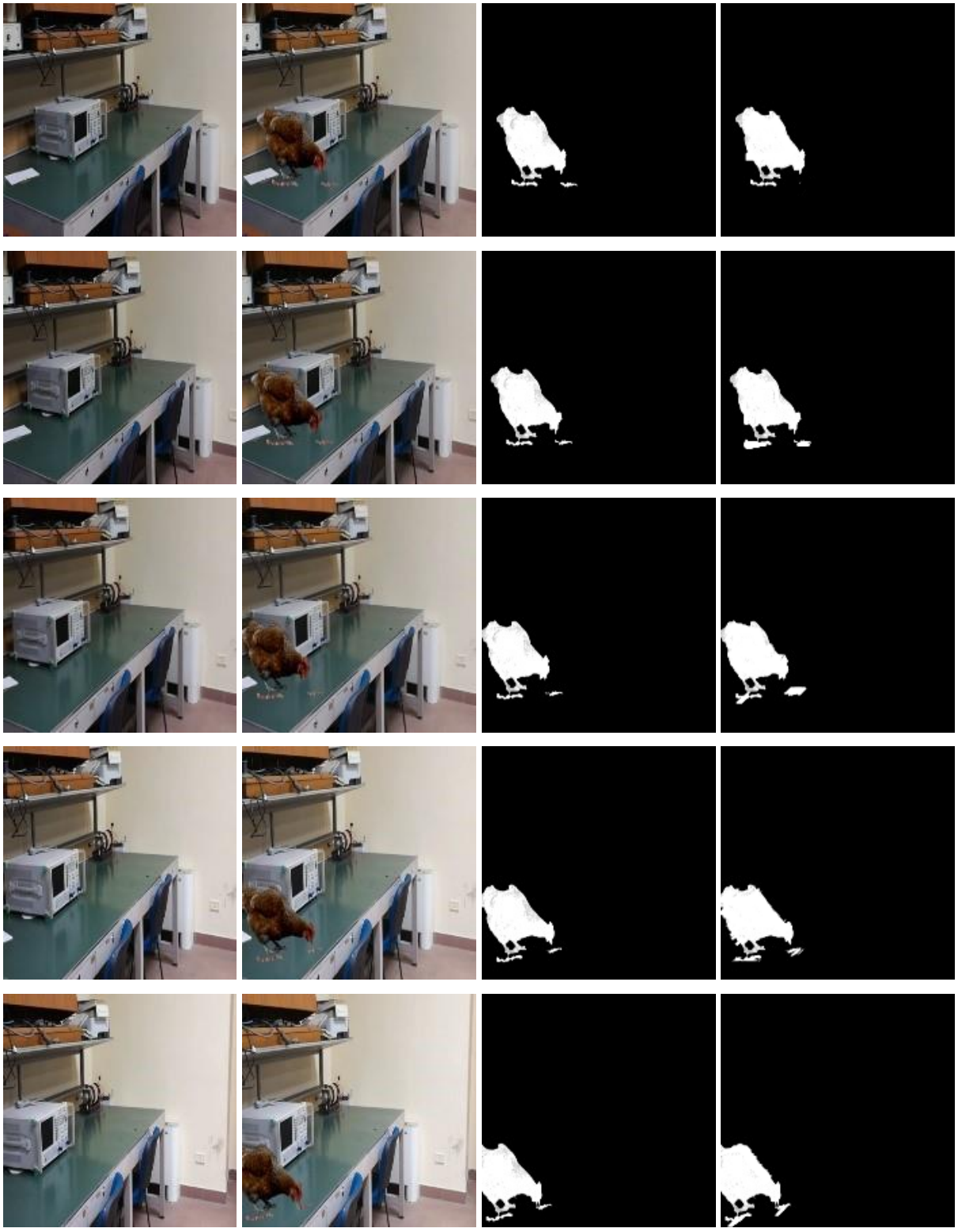
4.6.3 Results on GRIP Dataset

The results of video 1 and video 5 from Grip dataset are given in Fig. 4.10 and 4.11 below:



(a) Original Frames (b) Forged Frames (c) Ground Truth (d) Final Results

Figure 4.10 Results from TANK video 1 from dataset. (a) shows the authentic frames of the video (b) forged frames (c) ground truth and (d) Results with forged region shown specifically as white region



(a) Original Frames (b) Forged Frames (c) Ground Truth (d) Final Results

Figure 4.11 Results from HEN video5 from dataset. (a) shows the authentic frames of the video (b) forged frames (c) shows the ground truth and (d) final results with forged region shown specifically by white color

In Figure 4.8, it is clear that the frames of the authentic video have been forged by adding external object from some other source and pasted inside in the TV screen shown in the frame. The forgery has been done so in-distinguishably that it is really difficult to tell which frame is original and which is not. The proposed algorithm not only reveals the forged frame but also localizes the final forged region which is shown specifically by white color.

The video frames shown in Figure 4.9 are provided by REWIND research group. The original frames of the trustworthy video have been altered by adding the motion of a person. The person in forged video moves across the whole room whereas the original video doesn't have any such thing. Purely the alteration is in-distinguishable. The proposed algorithm classifies the forged frame and authentic frame successfully and also localizes the forged region which is exposed explicitly by white color.

Both the results in Figure 4.8 and 4.9 are obtained from the REWIND dataset. The nest two Figures 4.10 and 4.11 have been taken from GRIP dataset. The result discussion of these video is given as follows.

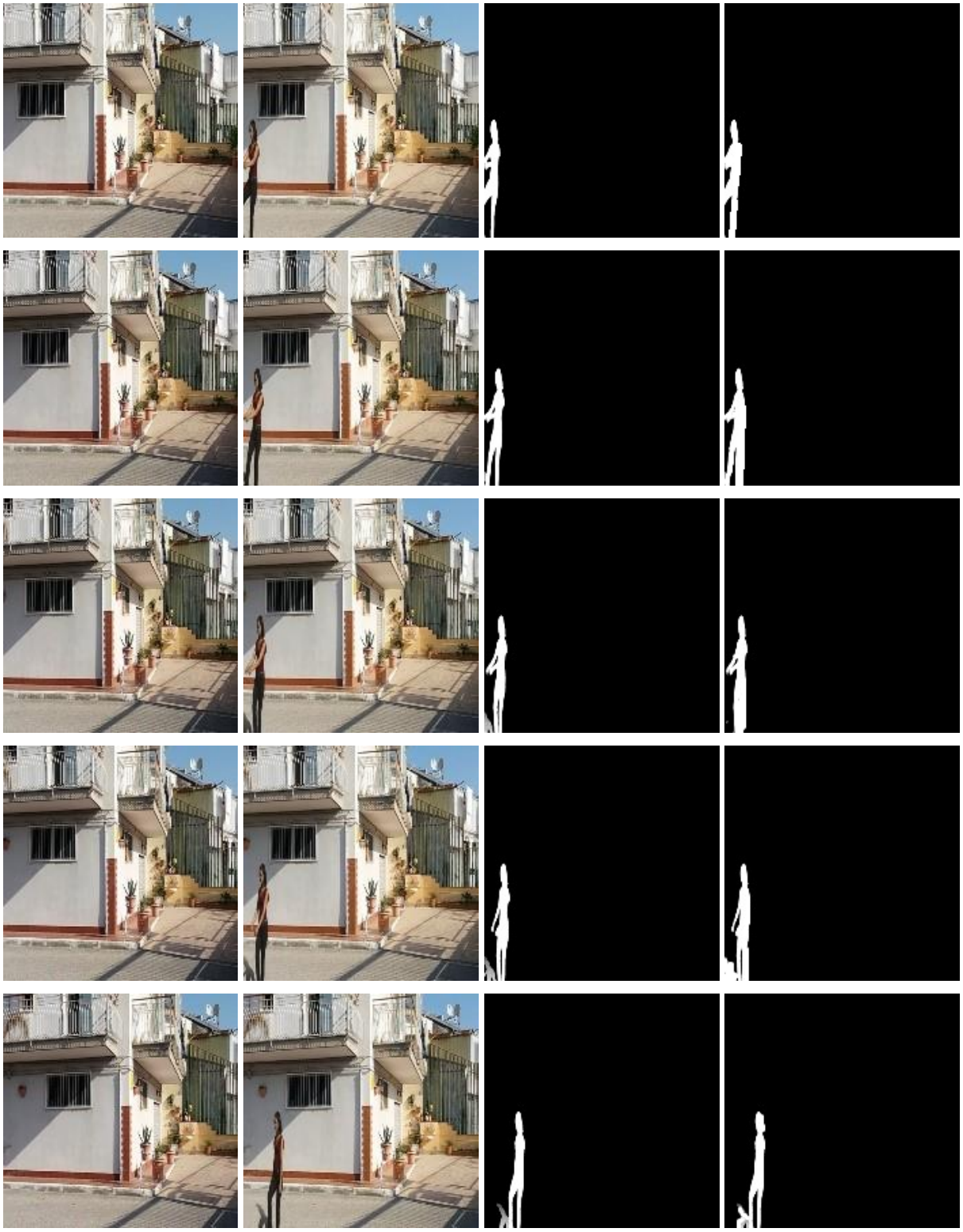
The Figure 4.10 shows the video frames from the GRIP dataset. The authentic frames contain only cars but these frames are altered with the insertion of an object tank from some other source. Figure also shows the ground truth frames provided by GRIP research group. The proposed algorithm when executed for this video gives exquisite results revealing the exact shape of the forged object with white color.

The Figure 4.11 displays the frames of video sequence from the GRIP dataset. The authentic frames do not contain any object like hen on the table. But these frames are altered by pasting hen. Figure 4.11 also demonstrates the ground truth frames provided by GRIP research group. The proposed algorithm when executed for this video gives results revealing the exact shape of the forged object with white color. The results so obtained matches precisely with the ground truth provided.

4.6.4 Compression Attack Results

GRIP dataset also provides YouTube Compressed forged video along with ground truth of each video sequence. The propose algorithm has been tested on these video sequences to check the robustness and come up with following results.

Some of the frames form the YT (YouTube) compressed forged video sequence along with their corresponding results have been demonstrated in Figure 4.12, 4.13 and 4.14.

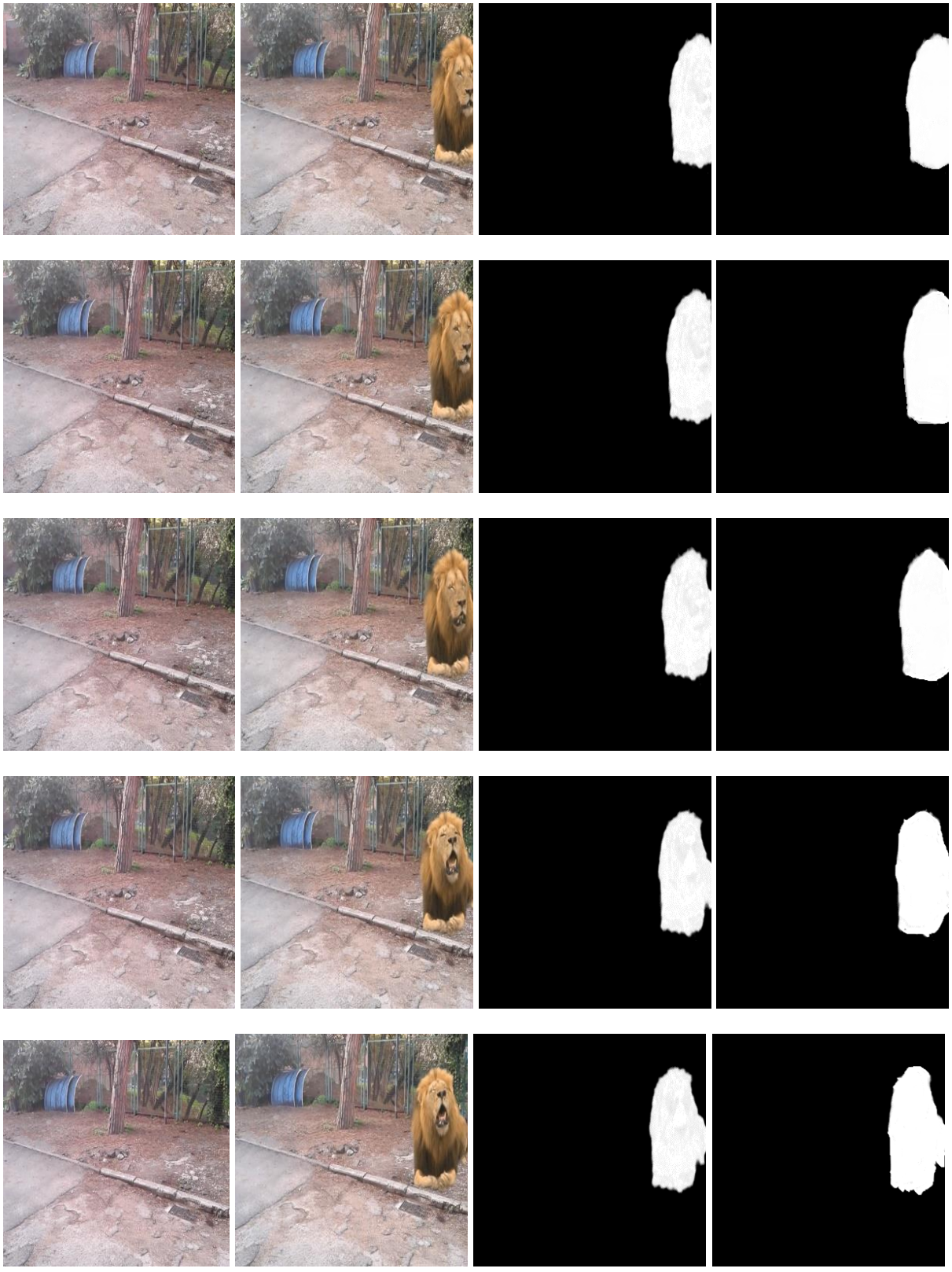


(a) Original Frames (b) Forged Frames (c) Ground Truth (d) Final Results

Figure 4.12 Results of You Tube compressed GIRL video 9 of dataset. (a) shows the authentic frames of the video (b) forged frames where the girl is inserted (c) shows the ground truth and (d) Results with forged region shown specifically by white color

(a) Original Frames (b) Forged Frames (c) Ground Truth (d) Final Results

Figure 4.13 Results of You Tube compressed HELICOPTER video 4 of dataset. (a) shows the authentic frames of the video (b) forged frames (c) ground truth and (d) Results with forged region shown specifically by white color



(b) Original Frames (c) Forged Frames (c) Ground Truth (d) Final Results

Figure 4.14 Results of You Tube compressed LION video 6 of dataset. (a) shows the authentic frames of the video (b) forged frames (c) ground truth and (d) Results with forged region shown specifically by white color

The Figure 4.12 shows the YT compressed video frames from the GRIP dataset. The authentic frames have been modified by pasting a girl. Figure also shows the ground truth frames provided by GRIP research group. The results of proposed algorithm for this video gives results exposing the precise shape of the forged object with white color.

The Figure 4.13 displays the frames form YT compressed video sequence provided by GRIP research group. The trustworthy frames have been modified the addition of external object helicopter. The proposed algorithm when executed for this video sequence shows the results exposing the precise shape of the forged object with white color that exactly matches with the ground truth provided in the dataset.

The Figure 4.14 shows the YT compressed video frames from the GRIP dataset. The original frames have been altered by pasting an external object lion. The results of proposed algorithm for this video gives results exposing the precise shape of the forged object with white color and also matches with ground truth.

4.6.5 Comparison CPU vs GPU

We tested our algorithm on two machines with different specifications to check the variation in the computational complexity. Specifications of both the machines are described in following Table 4.5. in the first machine, the algorithm execution time was quite high whereas it decreases when we shifted from machine 1 to machine 2. We used multiple GPUs using parallel processing toolbox with machine 2 and worked with high memory and the latest operating system, we attained results in fewer time. Moreover, the number of iterations in the training phase can be increased up to 1000 or 2000 or even more with CNN whereas it was difficult to go even up to 50 while working on machine 1 with CPU only.

Table 4.5. Comparison of machine configuration

Sr. No.	Parameter	Machine 1	Machine 2
1.	Windows	10	10
2.	Processor	Intel Core-i5-6200U CPU 2.9 GHz	Intel Core-i7-7700T CPU 2.9 GHz
3.	OS	64-bit	64-bit
4.	System	X64-based processor	X64-based processor
5.	RAM	4.0 GB	16.0 GB
6.	GPU	Intel	NVIDIA GeForce GTX 1070
7.	MATLAB	18a	18a

Figure 4.15 shows the maximum utilization of CPU as these are intended for extra general computing jobs whereas GPUs are considered to compute the similar commands parallelly. Deep Neural Networks are organized in such a way that all alike neurons of individual layer accomplish the similar computation in the end.

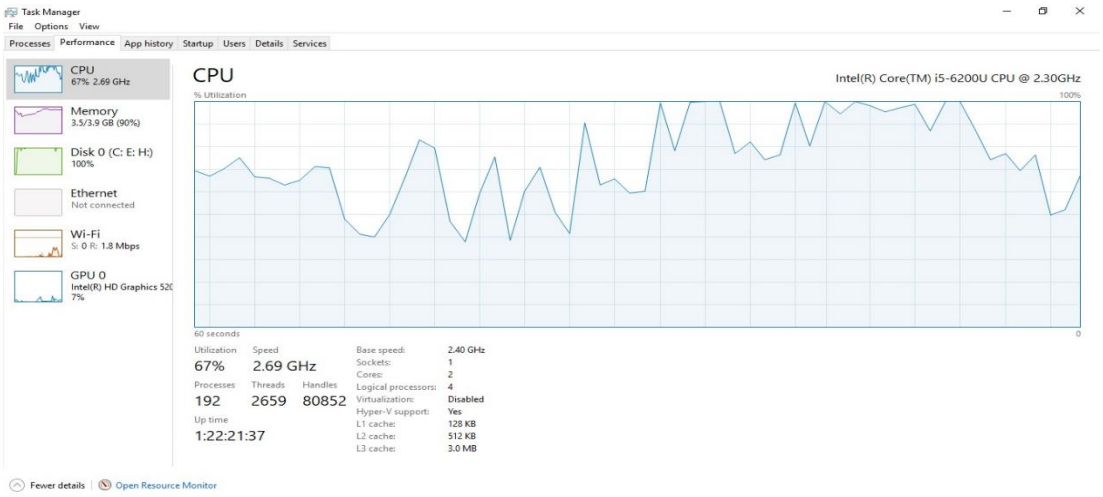


Figure 4.15 CPU properties of Machine 1

Hence, DNN structure suits particularly sound with computation that require GPU. Additionally, in applications like Convolution Neural Networks for image and video processing, the abilities of GPU graphics can be exploited to have high speed computations. Figure 4.15 and 4.16 show the comparison of CPU and GPU's usage during the time of code execution. In neural network-based problems, memory is the biggest issue as it stores the input that is fed to it and also the weight parameters related to the input till the end of the execution. We used NVIDIA GeForce GTX 1070 CUDA graphics and exploit its working power to achieve the execution results as fast as possible.

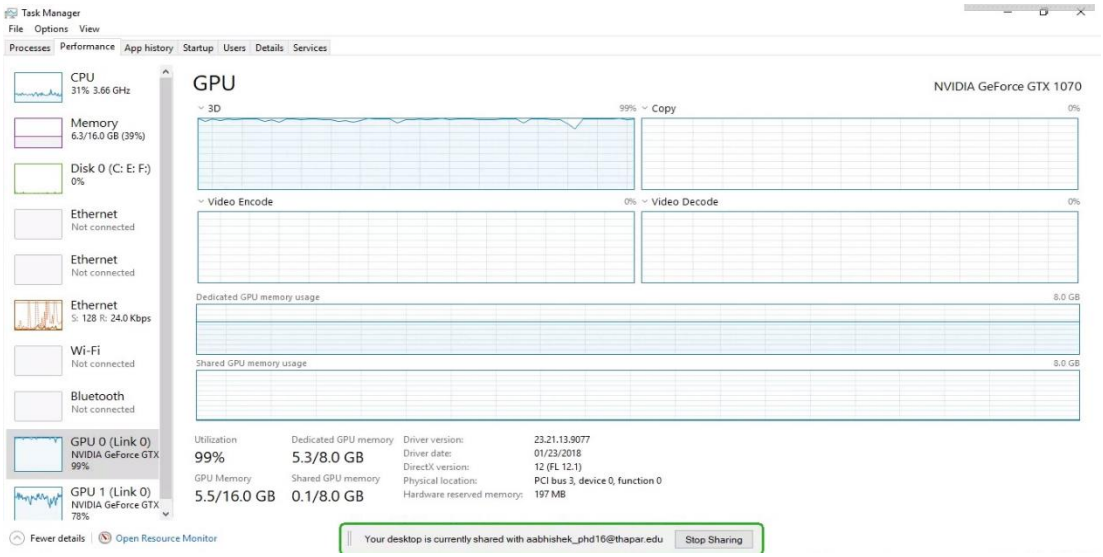


Figure 4.16 GPU properties of Machine 2 when two GPUs were being used

4.6.6 Performance Analysis

The training progress of the proposed network reaches up to 100% accuracy. The graph in Figure 4.17 shows curve accomplishing 100% accuracy from 60% and the curve below shows the loss that touches to zero before the completion of execution of iteration.

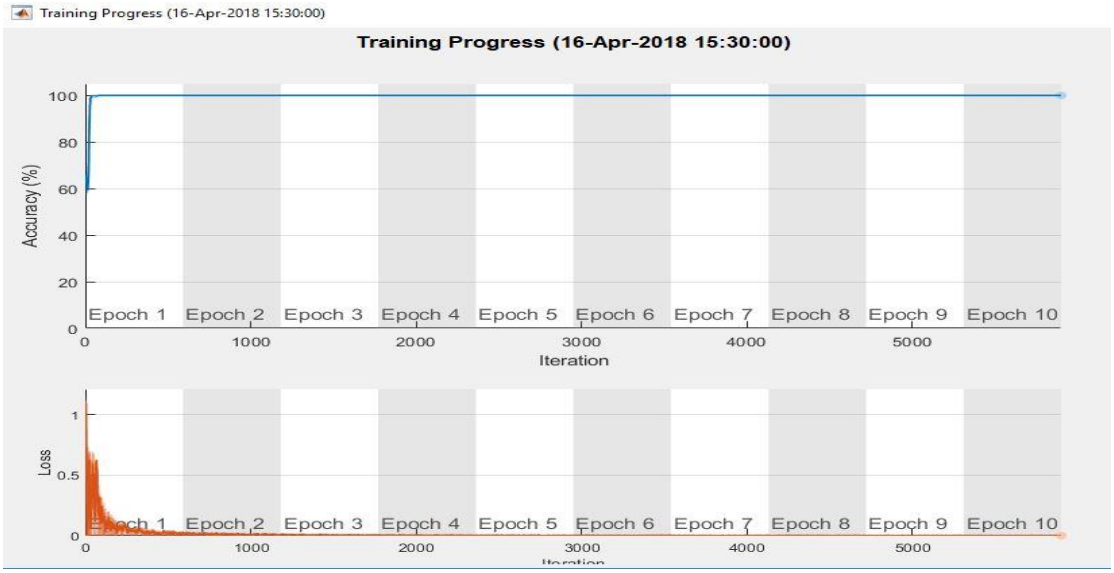


Figure 4.17 Learning progress of algorithm during training of semantic segmented frames.

To analyse the performance of the classification algorithm, we plotted a confusion matrix that gives a recovering impression about the errors that our classification model is making. If n is the total number of samples and TP represents the True Positive parameter i.e. forged frames acknowledged as forged, FP represents the False Positive parameter i.e. genuine frames acknowledged as forged, TN represents the True negative parameter i.e. genuine frames acknowledged as genuine, FN represents the False Negative parameter i.e. forged frames acknowledged as genuine, then True Positive Rate and False Positive Rate can be computed from following eq. 4.9 and 4.10 respectively.

$$TPR = {TP}/{(TP + FN)} \tag{4.9}$$

$$FNR = {FN}/{(TP + FN)} \tag{4.10}$$

The first plot in Figure 4.18 illustrates the True Positive Rate with 100% accuracy in predicted the true class and second plot illustrates False Positive Rate with 100% Positively predicted value and zero value false discovery rate.

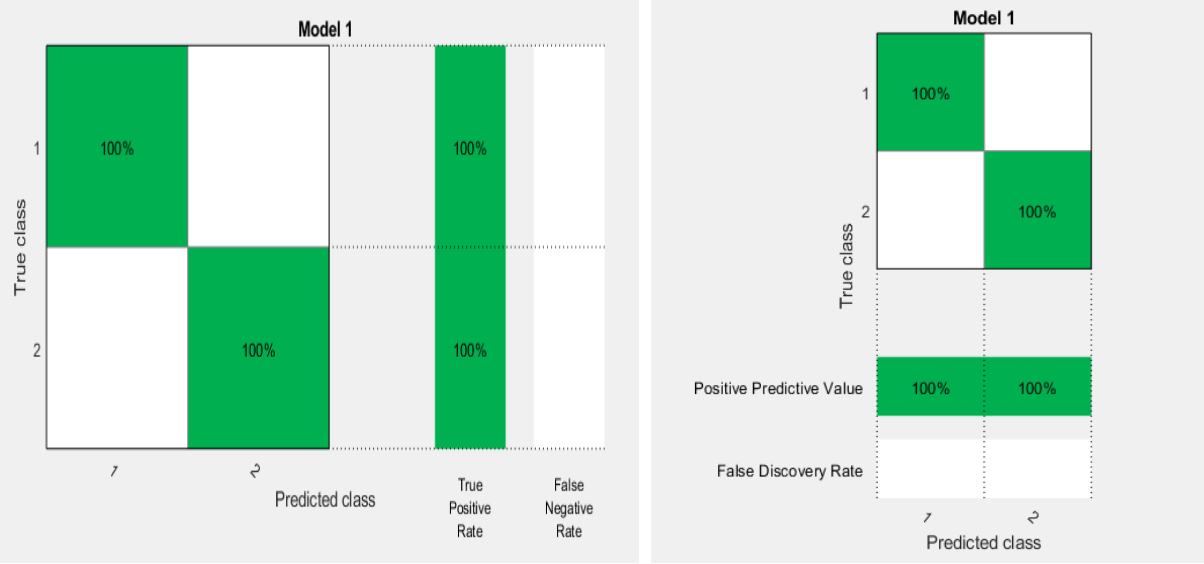


Figure 4.18 Graphs showing TPR, FPR and positive predictive rate using Confusion Matrix plot

The most common way to picture the performance of the classifier is ROC curve, also known as Receiver Operating Characteristics. ROC curve is an essential tool for analytical evaluation of test phase. ROC Curves shows how the classifier separates true and false values and also identifies the threshold that separates them. The accuracy can be computed by following equation 4.11.

$$\text{Accuracy} = {(\text{TP} + \text{TN})}\Big/{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{4.11}$$

We plot Sensitivity i.e. TPR against specificity i.e. FPR and is shown in third plot of Fig. 4.19.

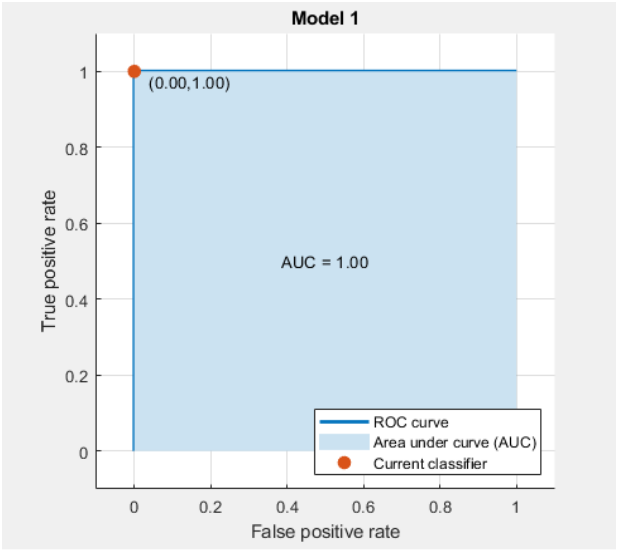


Figure 4.19 Receiver Operating Characteristics (ROC) plot for the proposed algorithm

Receiver Operating Characteristics with 100% AUC shown in Figure 4.19 has been obtained for GRIP video database. AUC or Area under Curve is used to summarize the performance of the classifier. It averages the performance over the range of scores of classifiers starting from low false negatives to high true positives.

The implemented results of proposed algorithm with 100% accuracy for GRIP dataset shows its superiority. This is also demonstrated in Table 4.6 by comparing the implemented results with other existing algorithms.

Table 4.6. Comparison of implemented accuracy with other algorithms

Algorithm	Detection Accuracy (%)
Pun [6]	90.8
A.V. [36]	89.7
Wang [45]	70.0
Lichao [52]	92.6
Proposed	100

The bar graph in Figure 4.20 corresponds to the results shown in Table 4.6. This shows the implemented results using Deep Convolutional Neural Networks detects the inter-frame video forgery with high accuracy in contrast to other traditional methods.

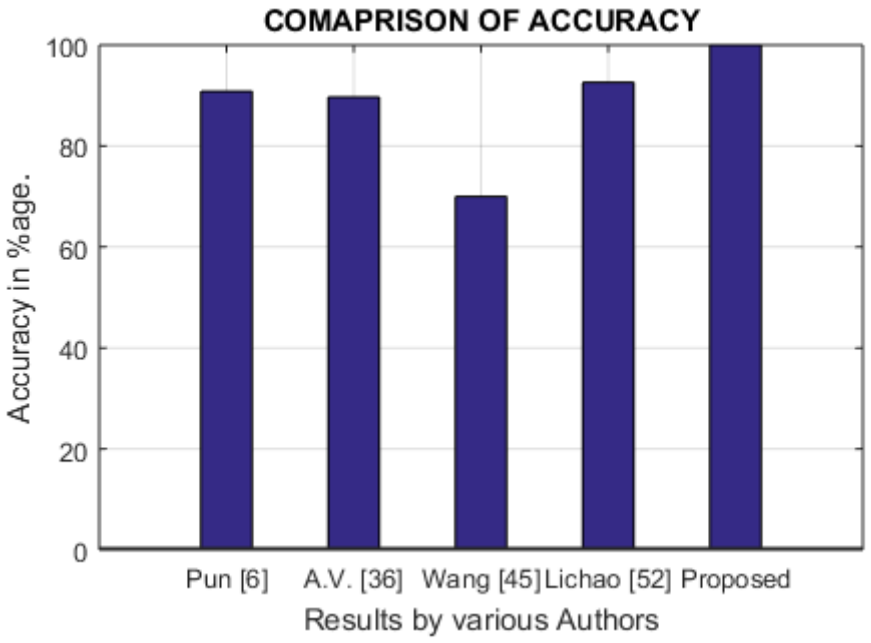


Figure 4.20 Bar Graph showing comparison of implemented detection accuracy with other algorithms

The compared approaches shown in Figure 4.20 comprise of the video forgery detection using HOG features in A.V. [36] (2012), the other method suggested by Wang [45] (2007), similar traditional approach-based algorithm given by Pun [6] (2015) and Mirror SIFT based technique in Lichao [52] (2017).

The Lichao [52] used the similar dataset to detect video forgery using traditional method named Mirror-SIFT and came up the following parametric results. The proposed algorithm using Deep CNN gives more efficient and superior results. The comparison of both the algorithms is shown in Table 4.7.

Table 4.7. Comparison of implemented results with Lichao

Lichao [52]	Implemented Results
True Positive (%)=95.2	True Positive (%)=100
False Negative (%)=10.0	False Negative (%)=0.0

The bar graph in Figure 4.21 corresponds to the results shown in Table 4.7. This proves the Deep Convolutional Neural Networks based approach for the detection of the video graphic forgery gives high true positive rate when compared to other approaches.

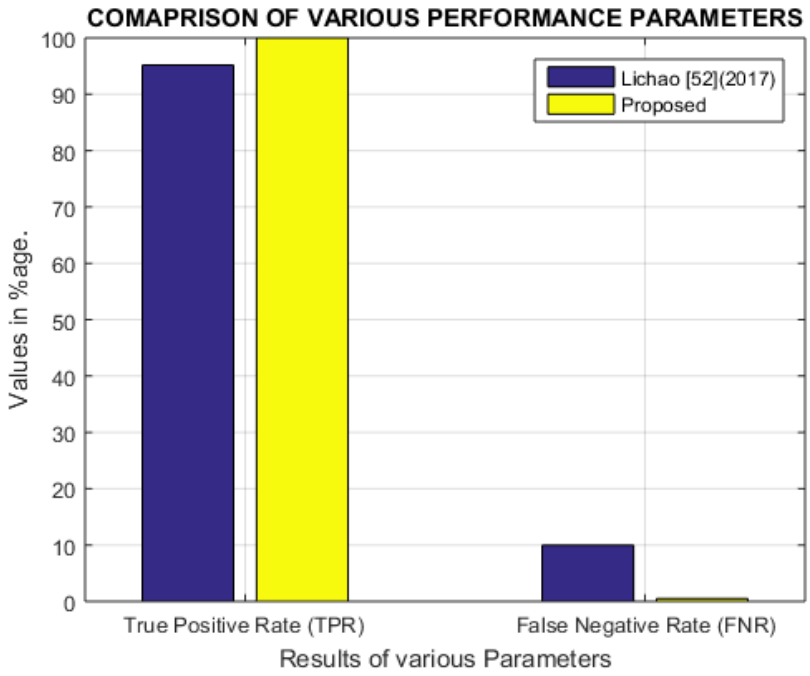


Figure 4.21 Bar Graph showing comparison of performance parameters with Lichao

4.7 SUMMARY

An optimized method for the exposure of inter-frame tampering in the video by means of Deep Convolutional Neural Network (DCNN) has been proposed that classifies the forged frames on the basis of correlation between the frames and the observed abnormalities. The decoders used for batch normalization of input improves the training swiftness. The robustness of the proposed algorithm is also tested on various You Tube compressed video provided by GRIP research group. Experimental results with high detection rate obtained on REWIND and GRIP video dataset shows the superiority of the proposed algorithm. The comparison of implemented results with pre-existing conventional methods proves high detection rate and highly accurate outcomes.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Digital forensics field has developed fast in the previous span in response to the growth of visual data alterations to commit a crime. From the wide-ranging examination of literature, it has been observed that procedures that could determine the counterfeits in visual data are deficient. Different sophisticated tools and progressive manipulation techniques have made forgery detection difficult. Many methods had been proposed in the previous years and their gaps had been covered by other researchers and this process seems to be never ending. Digital video forensics is still a research area at its infancy.

A deep convolutional neural network-based approach has been introduced for detection of forged content in the video. The trained dataset is fed to classifier that categorize the frames as authentic and forged on the basis of abruption in spatial and temporal correlation among the frames. The forged frames are again trained with the corresponding semantic segmented masks that highlights the abnormalities and hence framing out the forged region. The implemented results give 100% TPR and zero false negative with 100% detection accuracy. The proposed algorithm proves to be highly efficient when compared with existing techniques.

Mostly, the commercial video capturing devices store the images in compressed formats. In the world of social networking, people like to upload everything they capture on internet. You Tube (YT) platform is the fast-growing platform when it comes to interaction with world. But sometimes the video is altered or its properties like brightness, picture quality or background are modified before uploading and hence the compression occurs. This compression is further increased when these are finally uploaded and re-saved. Hence, the video content suffers much more compression then it actually had. The proposed algorithm is capable of detecting the forged content in You Tube (YT) compressed video with an accuracy reaching up to 100% for GRIP dataset and 98.99% for REWIND dataset.

Another statement that can be made on reviewing the above proposed work is that machines with low memory are less efficient to work on for deep neural network-based problems. Whereas the machines with highly memory efficient and NVIDIA CUDA graphics can train thousands on frames with more than 5000 iterations in a short interval of time. Also, CPU-

based calculations are the simplest and easily available option. But CPU-based computations are only good for less complicated structures using pre-trained networks as they tend to show less speed when it comes to high-end problems. But a GPU-based machine allows to work on single MATLAB software without any additional hardware. Also working with multiple GPUs with inbuilt parallel processing toolbox in MATLAB can further speed up the processing.

For a large resolution frame, individual neuron adds to extracted features and hence making it more complex. The most common way to handle this problem is down sample i.e. resizing frames but it can further lead to loss of information that may have proved profitable. With the introduction of deep learning techniques, a new solution to above problem has been presented. DCNN takes full advantage of architectural encoded information within the image. Hence, DCNN gives the best performance and lessen the continuous interaction.

Since the alteration in video is attractively becoming popular, an effort was put to comprehend and design new technology to localize these tampering. Encouraging outcomes have been attained using correlation consistencies to recognize any graphical replication.

5.2 FUTURE SCOPE

The area of video forgery detection is a fast expanding research field that promises a momentous improvement in detecting counterfeits despite all the restrictions of existing approaches. This field will keep on putting its best efforts to detect these doctored contents. The technology today has allowed alterations in such a manner that were merely incredible a decade ago. The upcoming technology may allow the manipulations that are far away from today's imagination. The future work for our implemented algorithm is that:

1. Cloud-based GPU calculations are proving much more beneficial in terms of cost and time. For this, no new hardware is required to be bought. MATLAB code can be written using the local GPUs and be extended for cloud-based resources just by changing few settings.
2. Training a Convolutional Neural Network is quite an enormous task when it comes to time consumption. Gradual training by small to larger network can help in reliable performance.
3. The implemented results have been obtained by semi-supervised training for localizing the forged content. The future vision for this can be obtaining similar results using complete unsupervised learning with less complicated networks.

REFERENCES

[1] Thyagarajan, KS *Still Image and Video Compression with MATLAB.* New Jersey: John Wiley & Sons, 2011.

[2] Kaur R and Kaur J (2016). Video Forgery detection using Hybrid techniques, *International Journal of Advanced Research in Computer and Communication Engineering*, 5(12), 112-117.

[3] REWIND (REVerse Engg. of audio-Visual content Data) - Video: copy-move forgeries dataset. Available at https://sites.google.com/site/rewindpolimi/downloads/datasets (Accessed on 10th September 2017).

[4] Bhanu Bhavya MP and Kumar Arun MN (2017). Copy-Move Forgery Detection Using Segmentation, *11th International Conference on Intelligent Systems and Control (ISCO)*, 224-228.

[5] Papinwar Sonal R (2016). Forgery Detection in Video Using Watermarking: A Review, *IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents*, 3654, 40-51.

[6] Pun CM, Yuan XC and Bi XL (2015). Image Forgery Detection Using Adaptive Over segmentation and Feature Point Matching, *IEEE Transactions on Information Forensics and Security*, 10, 1705-1716.

[7] Singh RD and Aggarwal N (2017). Video content authentication techniques: a comprehensive survey, *Multimedia Systems*, *24*, 211-240.

[8] GRIP – Image Processing Research Group. Available at http://www.grip.unina.it/web-download.html (Accessed on 10th January 2018).

[9] Before and After effects. https://www.pinterest.com/mariotamashiro/before-and-after/ (Accessed on 6th May 2018).

[10] Dutta U and Sharma C (2013). Analysis of Copy-Move Image Forgery Detection, International Journal of Advanced Research in Computer Science and Electronics Engineering, 2(8), 607-609.

[11] Christlein V *et al.* (2012). An Evaluation of Popular Copy-Move Forgery Detection Approaches, *IEEE Transactions on Information Forensics and Security*, *7*, 1841-1854.

[12] Ng T, Chang S and Sun Q (2004). Blind detection of photomontage using higher order statistics, *IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512), 5*, V-V.

[13] Amerini I *et al.* (2011). A SIFT-Based Forensic Method for Copy–Move Attack Detection and Transformation Recovery, *IEEE Transactions on Information Forensics and Security, 6*, 1099-1110.

[14] Goodfellow IJ, Bengio Y and Courville AC (2015). Deep Learning, *Scholarpedia, 10*, 32832.

[15] Krizhevsky A, Sutskever I and Hinton GE (2012). ImageNet Classification with Deep Convolutional Neural Networks, *Neural Information Processing Systems* (*NIPS)*, 1-9.

[16] Zaremba W, Sutskever I and Vinyals O (2014). Recurrent Neural Network Regularization, *Computing Research Repository* (CoRR*), abs/1409.2329*.

[17] Widrow B and Lehr MA (1990). 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation.

[18] Ponlatha S and Sabeenian RS (2013). Comparison of Video Compression Standards, *International Journal of Computer and Electrical Engineering, 5(6), 549-554.*

[19] Bidokhti A and Ghaemmaghami S (2015). Detection of regional copy/move forgery in MPEG videos using optical flow. *The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 13-17.

[20] Kobayashi M, Okabe T and Sato Y (2009). Detecting Video Forgeries Based on Noise Characteristics, *Pacific-Rim Symposium on Image and Video Technology (*PSIVT*),* 306-313.

[21] Kang X et al. (2015). Forensics and counter anti-forensics of video inter-frame forgery, *Multimedia Tools and Applications, 75*, 13833-13853.

[22] Lin C and Tsay J (2014). A passive approach for effective detection and localization of region-level video forgery with spatio-temporal coherence analysis, *Digital Investigation, 11*, 120-140.

[23] Zhang K et al. (2014). Fast Visual Tracking via Dense Spatio-temporal Context Learning. *European Conference on Computer Vision*- ECCV, 127-141.

[24] Wang W and Farid H (2007). Exposing Digital Forgeries in Interlaced and Deinterlaced Video, *IEEE Transactions on Information Forensics and Security, 2(3)*, 438-449.

[25] Kiran BR, Thomas DM and Parakkal R (2018). An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos, *Computer Vision and Pattern Recognition, 4*, 1-15.

[26] Lin G, Chang J and Chuang C (2011). Detecting frame duplication based on spatial and temporal analyses, *2011 6th International Conference on Computer Science & Education (ICCSE)*, 1396-1399.

[27] Wang W and Farid H (2007). Exposing digital forgeries in video by detecting duplication, *9th workshop of Multimedia and Security*, 35-42.

[28] Mathai M, Rajan D and Emmanuel S (2016). Video forgery detection and localization using normalized cross-correlation of moment features, *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 149-152.

[29] Hyun D *et al.* (2013). Detection of Upscale-Crop and Partial Manipulation in Surveillance Video Based on Sensor Pattern Noise, *Sensors Basel*, 3(9), 12605-12630.

[30] Xu J *et al.* (2016). A novel video inter-frame forgery detection method based on histogram intersection, *IEEE/CIC International Conference on Communications in China (ICCC)*, 1-6.

[31] Andy S and Haikal A (2017). Simple duplicate frame detection of MJPEG codec for video forensic, *2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 321-324.

[32] Kobayashi M, Okabe T and Sato Y (2010). Detecting Forgery from Static-Scene Video Based on Inconsistency in Noise Level Functions, *IEEE Transactions on Information Forensics and Security, 5*, 883-892.

[33] Hsu C et al. (2008). Video forgery detection using correlation of noise residue, *IEEE 10th Workshop on Multimedia Signal Processing*, 170-174.

[34] Pandey RC, Singh SK and Shukla KK (2016). Passive Forensics in Image and Video Using Noise Features: A Review, *Digital Investigation*, 19, 1–28.

[35] Kaur M and Walia S (2016). Forgery Detection Using Noise Estimation and HOG Feature Extraction, *International Journal of Multimedia and Ubiquitous Engineering*, 11(4), 37-48.

[36] Subramanyam AV and Emmanuel S (2012). Video forgery detection using HOG features and compression properties, *IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, 89–94.

[37] Goodwin J and Chetty G (2011). Blind video tamper detection based on fusion of source features, *IEEE International Conference on Digital image computing techniques and applications (DICTA)*, 608–613.

[38] Khammar MR (2012). Evaluation of different block matching algorithms to motion estimation, *International Journal of VLSI and Embedded Systems-IJVES,* 3(3), 148-153.

[39] Conotter V, O'Brien JF and Farid H (2012). Exposing digital forgeries in ballistic motion, *IEEE Transactions on Information Forensics and Security,* 7(1), 283–296.

[40] Shih TL *et al.* (2011). Video Motion Interpolation for Special Effect Applications, *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews,* 41(5), 720-732.

[41] Stamm MC *et al.* (2012). Temporal Forensics and Anti-Forensics for Motion Compensated Video, *IEEE Transactions on information forensics and security,* 7(4), 1315-1327.

[42] Wang W and Farid H (2006). Exposing digital forgeries in video by detecting double MPEG compression, *Proceedings of 8th Workshop on Multimedia and Security (MM&Sec*), 37–47.

[43] Su Y *et al.* (2011). A frame tampering detection algorithm for MPEG videos, *6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, 2, 461–464.

[44] Chen J *et al.* (2016). Detecting double MPEG compression with the same quantiser scale based on MBM feature, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2064-2068.

[45] Wang W and Farid H (2009). Exposing digital forgeries in video by detecting double quantization, *11th ACM workshop on multimedia and security*, 39–48.

[46] Chetty G (2010). Blind and passive digital video tamper detection based on multimodal fusion, *14th WSEAS International Conference on Communications*, 109–117.

[47] Sundaram MA and Nandini C (2015). CBFD: Coherence Based Forgery Detection technique in image forensics analysis, *International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, 192-197.

[48] Zhong J *et al.* (2017) A new block-based method for copy move forgery detection under image geometric transforms, *Multimedia Tools and Applications*, 76(13), 4887-14903.

[49] Yin H *et al.* (2012). A Novel Large-Scale Digital Forensics Service Platform for Internet Videos, IEEE *Transactions on Multimedia, 14*, 178-186.

[50] Zhang J, Su Y and Zhang M (2009). Exposing digital video forgery by ghost shadow artifact, *1st ACM Workshop on Multimedia in Forensics (MiFor*), 49–54.

[51] Ma R, Chen J and Su Z (2010). MI-SIFT: Mirror and inversion invariant generalization for SIFT descriptor, *ACM international conference on image and video retrieval, ACM,* 228–235.

[52] Su L and Li C (2017). A novel passive forgery detection algorithm for video region duplication, *Multidim Syst Sign Process,* 1-18.

[53] Amerini I *et al.* (2011). A SIFT-Based Forensic Method for Copy–Move Attack Detection and Transformation Recovery, *IEEE Transactions on Information Forensics and Security, 6(3),* 1099-1110.

[54] Gao B, & Jin Y (2010). Detection of Image copy-move tamper Using SURF in digital forensics, *Asia-Pacific conference on information network and digital content security,* 58–62.

[55] Shahroudnejad A and Rahmati M (2016). Copy-move forgery detection in digital images using Affine-SIFT, *2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS),* 1-5.

[56] Liu L *et al.* (2014). Improved SIFT-Based Copy-Move Detection Using BFSN Clustering and CFA Features, *Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing,* 626-629.

[57] Hashmi MF, Hambarde AR and Keskar A (2013). Copy move forgery detection using DWT and SIFT features, *13th International Conference on Intellient Systems Design and Applications,* 188-193.

[58] Agarwal VR and Mane VM (2016). Reflective SIFT for improving the detection of copy-move image forgery, *Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN),* 84-88.

[59] Panchal1 PM, Panchal SR and Shah SK (2013). A Comparison of SIFT and SURF, *International Journal of Innovative Research in Computer and Communication Engineering,* 1(2), 323-327.

[60] Pun C, Yuan X and Bi X (2015). Image Forgery Detection Using Adaptive Over segmentation and Feature Point Matching, *IEEE Transactions on Information Forensics and Security, 10,* 1705-1716.

[61] Malviya AV and Ladhake SA (2015). Copy Move forgery detection using low complexity feature extraction, *UP Section Conference on Electrical Computer and Electronics (UPCON),* 1-5.

[62] Zhang K *et al.* (2014). Fast visual tracking via dense spatio-temporal context learning, European Conference on Computer vision (ECCV), 127–141.

[63] Lin GS *et al.* (2011). Detecting frame duplication based on spatial and temporal analyses, *6th IEEE International Conference on Computer Science and Education (ICCSE)*, 1396–1399.

[64] Xu J *et al.* (2016). A novel video inter-frame forgery detection method based on histogram intersection. *IEEE/CIC International Conference on Communications in China (ICCC)*, 1-6.

[65] Lin GS and Chang JF (2012). Detection of Frame Duplication Forgery in Videos Based on Spatial and Temporal Analysis, *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI), 26, 1-18.

[66] Sharma S and Dhavale SV (2016). A review of passive forensic techniques for detection of copy-move attacks on digital videos, 3*rd International Conference on Advanced Computing and Communication Systems (ICACCS), 1*, 1-6.

[67] Chetty G, Biswas M and Singh R (2010). Digital Video Tamper Detection Based on Multimodal Fusion of Residue Features, *Fourth International Conference on Network and System Security*, 606-613.

[68] Jun W, Lee Y and Jun B (2015). Duplicate video detection for large-scale multimedia, *Multimedia Tools and Applications, 75*, 15665-15678.

[69] Hyun DK *et al.* (2013). Detection of upscale-crop and partial manipulation in surveillance video based on sensor pattern noise, *Sensors*, 12605–12631.

[70] Jaiswal S and Dhavale S (2013). Video Forensics in Temporal Domain using Machine Learning Techniques, *International Journal Computer Network and Information Security*, 9, 58-67.

[71] Bhandare A *et al.* (2016). Applications of Convolutional Neural Networks, *International Journal of Computer Science and Information Technologies (IJCSIT)*, 7(5), 2206-2215.

[72] Ciresan DC *et al.* (2011). Flexible, High Performance Convolutional Neural Networks for Image Classification, *Twenty-Second International Joint Conference on Artificial Intelligence*, 2, 1237-1242.

[73] Gopi E. S *et al.* (2006). Digital Image Forgery Detection using Artificial Neural Network and Auto Regressive Coefficients, *Canadian Conference on Electrical and Computer Engineering* (2006), 194-197.

[74] Rao Y and Ni J (2016). A deep learning approach to detection of splicing and copy-move forgeries in images, *IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-6.

[75] Gironi A *et al.* (2014). A video forensic technique for detecting frame deletion and insertion, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6226-6230.

[76] He P *et al.* (2017). Frame-wise detection of relocated I-frames in double compressed H.264 videos based on convolutional neural network, J. *Visual Communication and Image Representation, 48*, 149-158.

[77] Ying Z *et al.* (2016). Image Region Forgery Detection: A Deep Learning Approach, *Singapore* Cyber Security R&D Conference *(SG-CRC*).

[78] Chen J *et al.* (2015). Median Filtering Forensics Based on Convolutional Neural Networks, *IEEE Signal Processing Letters,* 22, 1849-1853.

[79] Gopi ES (2007). Digital image forgery detection using artificial neural network and independent component analysis, *Applied Mathematics and Computation*, 194(2), 540-543.

[80] Long C *et al.* (2017). A C3D-based Convolutional Neural Network for Frame Dropping Detection in a Single Video Shot, *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1898-1906.

[81] Shelhamer E, Long J and Darrell T (2015). Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 640-651.

[82] Sizyakin R *et al. (2017).* Defect detection on videos using neural network, MATEC Web of Conferences, 132.

[83] Niklaus S, Mai L and Liu F (2017). Video Frame Interpolation via Adaptive Separable Convolution, *IEEE International Conference on Computer Vision (ICCV)*, 261-270.

[84] Xu L *et al.* (2014). Deep Convolutional Neural Network for Image Deconvolution, *27th International Conference on Neural Information Processing Systems,* 1790-1798.

[85] Yao Y *et al.* (2017). Deep Learning for Detection of Object-Based Forgery in Advanced Video, *Proc. Symmetry,* 1-10

[86] D'Avino D *et al.* (2017). Autoencoder with recurrent neural networks for video forgery detection, *Proc. IS&T Electronic Imaging: Media Watermarking, Security, and Forensics, abs/1708.0875.*

[87] D'Amiano L *et al.* (2015). Video forgery detection and localization based on 3D patchmatch, *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1-6.

[88] Sanjary A and Ismael O (2015). Detection of Video Forgery: A Review of Literature, *Journal of Theoretical and Applied Information Technology*, 74(2), 207-220.

[89] Zheng L, Sun T and Shi YQ (2014). Inter-frame Video Forgery Detection Based on Block-Wise Brightness Variance Descriptor, *International Workshop on Digital-forensics and Watermarking (*IWDW).

[90] Tan S, Chen S and Li B (2015). Automatic Detection of Object-Based Forgery in Advanced Video, *IEEE Transactions on Circuits and Systems for Video Technology, 26*, 2138-2151.

[91] Liu Q *et al.* (2016). Exposing Inpainting Forgery in JPEG Images under Recompression Attacks, *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 164-169.

[92] Li M and Monga V (2012). Robust Video Hashing via Multilinear Subspace Projections, *IEEE Transactions on Image Processing, 21*, 4397-4409.

[93] Andy S and Haikal A (2017). Simple duplicate frame detection of MJPEG codec for video forensic, *2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 321-324.

[94] Ravi H *et al.* (2014). Compression noise-based video forgery detection, *2014 IEEE International Conference on Image Processing (ICIP)*, 5352-5356.

[95] Hijazi SL and Kumar R (2015). Using Convolutional Neural Networks for Image Recognition, *cadence*, 1-12.

[96] LeCunn Y *et al.* (1998). Gradient-based Learning applied to Document Recognition, *Proceedings of IEEE*, 86(11), 2278-2324.

[97] Architecture of Convolutional Neural Networks (CNNs) demystified. Available at https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/ (Accessed on 16th April 2018).

[98] Al-Waisy AS *et al.* (2017). A multi-biometric iris recognition system based on a deep learning approach, *Pattern Analysis and Applications*, 1-20.

LIST OF PUBLICATIONS

[1] Kaur H and Jindal N (2018). Image and Video Forensics - A Collaborative Survey, *Communicated. (SCI Indexed).*

[2] Kaur H *et. al.* (2018). Video Interframe Replication Detection using Deep Neural Networks, *Communicated. (SCOPUS Indexed).*

[3] Kaur H and Jindal N (2018). Deep Convolutional Neural Network for Graphics Forgery Recognition in Video, *Communicated. (SCI Indexed).*