Image and Video Forgery Identification and Localization from Multimedia Forensics Perspective

A thesis submitted to University of Hyderabad in partial fulfilment

for the degree of

Doctor of Philosophy

by

Raghavendra Gowda D

Reg. No. 18MCPC02

SCHOOL OF COMPUTER AND INFORMATION SCIENCES

UNIVERSITY OF HYDERABAD

HYDERABAD -500046

Telangana

India

March, 2023

CERTIFICATE

This is to certify that the thesis entitled **"Image and Video Forgery Identification and Localization from Multimedia Forensics Perspective"** submitted by **Raghavendra Gowda D** bearing **Reg. No. 18MCPC02** in partial fulfilment of the requirements for the award of **Doctor of Philosophy** in **Computer Science** is a bonafide work carried out by him under my supervision and guidance. This thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for award of any degree or diploma. Parts of this thesis have been published online in the following publications:

1. Gowda Raghavendra and Digambar Pawar. **Deep learning-based forgery identification and localization in videos**. *Signal, Image and Video Processing* (2022): Page 1-8. https://doi.org/10.1007/s11760-022-02433-7. Indexed: Science Citation Index Expanded (SCIE), SCOPUS, UGC-CARE List (India). This publication is reported as part of Chapter 4.

2. Gowda, R., Pawar, D and Barman, B. **Unethical human action recognition using deep learning based hybrid model for video forensics**, *Multimedia Tools and Applications* (2023): Page 1-26. https://doi.org/10.1007/s11042-023-14508-9. Indexed: Science Citation Index Expanded (SCIE), SCOPUS, UGC-CARE List (India). This publication is reported as part of Chapter 5.

and

has made presentations in the following conferences.

1. Gowda Raghavendra and Digambar Pawar. **Porn Image Forensics: Image Classification, Forgery Detection, and Localization**. In *International Conference on Computing in Engineering & Technology* (2022) (pp.359-371). https://doi.org/10.1007/978-981-19-2719-5-34. Indexed: SCOPUS. This publication is reported as part of Chapter 3.

2. Gowda Raghavendra, Digambar Pawar and Tetali S. R. **Multimedia Forensics-An approach to detect and analyze Human faces in multimedia files**. In 2019 Fifth *International Conference on Image Information Processing (ICIIP)* (2019) (pp.274-279). https://doi.org/10.1109/ICIIP47207.2019.8985694. Indexed: IEEE Xplore. This publication is reported as part of Chapter 5.

Further, the student has passed the following courses towards fulfilment of coursework requirement for Ph.D.

	Course Code	**Name**	**Credits**	**Pass/Fail**
1	CS402	Algorithms	4	Pass
2	CS800	Research Methods in Computer Science	4	Pass
3	CS803	Data structures and Programming Lab	2	Pass
4	CS858	Ethical Hacking and Computer Forensics	3	Pass

Dr. Digambar Pawar
Supervisor
SCIS, University of Hyderabad
Hyderabad-500 046, India

Prof. Atul Negi
Dean of School
SCIS, University of Hyderabad
Hyderabad-500 046, India

DECLARATION

I, **Raghavendra Gowda D**, hereby declare that this thesis entitled **"Image and Video Forgery Identification and Localization from Multimedia Forensics Perspective"** submitted by me under the supervision of **Dr. Digambar Pawar**, is a bonafide research work and is free from plagiarism. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma. I hereby agree that my thesis can be deposited in Shodganga/INFLIBNET.

A report on plagiarism statistics from the University Librarian is enclosed.

Date:

Signature of the Student

(Raghavendra Gowda D)

Reg. No.: 18MCPC02

//Countersigned//

Signature of the Supervisor

(Dr. Digambar Pawar)

Dedicated To My Family & Teachers

Acknowledgements

I am deeply indebted to **Dr. Digambar Pawar**, my thesis supervisor for his meticulous guidance, wisdom, and support during the course of my Ph.D. at University of Hyderabad. His consistent encouragement and positive reinforcement have made it a gratifying experience. I have learned from him the virtues of rigour and professionalism. I am grateful to him for keeping me focused at the times when it mattered a lot. I express deep respect and gratitude to my supervisor for astute guidance and constant inspiration, without which this work would have never been possible.

I thank my doctoral review committee members, Prof. Chakravarthy Bhagvati and Prof. Saifullah Abdul MD for their encouragement, insightful comments, and hard questions which strengthened my knowledge.

It is my privilege to thank **Prof. Atul Negi**, Dean, SCIS, University of Hyderabad for his considerate support and encouragement throughout the tenure of my research work and for extending the facilities to pursue my research.

I express my sincere gratitude to every member of SCIS family for providing a smooth and enriching environment so that I enjoyed the tenure at the institute. I thank every member at my lab for providing a lovely experience to me.

I, especially, thank my family members and friends for their support and inspiration during the Ph.D work. At the end, I am grateful to **University of Hyderabad**, for making it a memorable experience.

Abstract

In the computing world, multimedia forensics is an exciting and challenging field, which is basically a branch of digital forensics. With the rapid increase in the use of digital technology, crimes today are committed using contemporary techniques that do not involve physical contact. As a result, forensic specialists are unable to examine and analyze the data at the crime scene. A change in the investigation techniques is necessary to achieve effective investigation of crimes involving advanced technology. This thesis focuses on image and video forgery analysis from a multimedia forensics perspective.

Forgeries of digital images compromise the authenticity and integrity of the images. We focus on two frequently used image forgery attacks i.e., copy-move and image splicing. The use of deep learning-based approaches in image recognition tasks inspired us to develop a model that can detect and locate manipulated regions in an image. We propose an LSTM-CNN based hybrid model for the generation of binary masks and detect the forged region with an improved SIFT algorithm. The SIFT algorithm helps the model invariant to detect and localize forged objects. Then generate the bounding box around the forged region to classify the image tampering as copy-move or image splicing.

Video forgery has become an easy and on-going task for the users of smart devices since easily available software tools made the task of a naive user effortless for video forgery. The impact of video forgery is critical when it is used to defame a personality and hide (or forge) important information to prove innocent in a crime scene and escape from legal action. The traces left behind after the forgery can be used to distinguish between genuine and manipulated videos. Using passive approaches, we can detect any unauthorized manipulation, whether it's done within a frame (intra-frame level) or between

frames (inter-frame level). The investigation of inter-frame video forgeries is the main emphasis of our work. We propose, a deep learning based 3 Dimensional Convolutional Neural Network (3DCNN) model for detecting video inter-frame forgery and localize the same using multi-scale structural similarity (MS-SSIM) index measurement algorithm. The proposed model outperforms existing models in various post-processing operations and compression rates.

Video forensics faces new obstacles in recognizing unethical human actions in video surveillance systems, human-computer interactions, etc. that requires multiple activity recognition systems. Existing deep learning methods solve the problems of unethical human action recognition which are effective in learning low-level temporal and spatial features but struggle from learning high-level features that affect the feature learning capability of the model. Due to this problem, deep learning methods suffer from poor performance and learning ability. We propose, a deep learning-based hybrid model for unethical human action recognition using two-stream inflated 3D ConvNet(I3D) and spatio-temporal attention (STA) modules. The I3D model improves the performance of 3DCNN architecture and STA increases the learning capability of the model. The proposed model is compared with the existing models using unique and multi-action datasets to show better performance capability.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Forensic science refers to the scientific methods used to obtain proof or evidence that is proven true in criminal investigations. The term forensics derives from the Latin word 'forum', which means 'main square', an ancient location where public court hearings were held.

The definition of digital forensic science was first introduced at Digital Forensics Research Workshop (DFRWS) in 2001 by academic researchers as "The use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations"[6].

The NIST (National Institute of Standards and Technology) definition of digital forensics is: "Digital forensics, also known as computer forensics and network forensics, is the application of science to the identification, collection, examination, and analysis of data while maintaining the information's integrity and a strict chain of custody"[7]. With the fast evolution of digital technology, digital forensics has branched out into new fields such as computer forensics, multimedia forensics, network forensics, disk forensics, mobile forensics, cloud forensics, IoT forensics, and so on. Our focus of research is on multimedia forensics. The emergence of multimedia forensics as a subset of digital forensics is due to the widespread use of social networks (Facebook, Instagram, Twitter, and so on) that share approximately 1.2 billion multimedia contents per day, which we now refer to as Bigdata. Multimedia forensics allows testing the multimedia contents to authenticate and identify any forgeries that have occurred in the files.

1.1 Multimedia Forensics

Multimedia forensics is one of the exciting and challenging fields in the computing world which is basically the branch of digital forensics. Multimedia forensics deals with audio, image, and video analysis using various methods to authenticate and test the integrity of a digital source for the purpose of detecting forgery. Due to the widespread adoption of mobile devices, lower storage costs and faster transfer speeds, online customers are generating massive amount of data. The effect of rapid increase in the usage of digital technology, crimes today is adopted with modern techniques that involve no physical communication. This has surpassed the forensics specialists' abilities to successfully examine and analyze the data at a crime scene. There is a need to change the investigation techniques to achieve effective investigation of crimes involving advanced technology.

Multimedia forensics is evidence based strategy that will be used to investigate and analyze the extracted digital evidence from multimedia files to combat Cybercrimes or any other incident involving data misuse. The purpose of analyzing digital evidence is to maintain the integrity and authenticity. "In the forensics domain, multimedia forensics is concerned with evaluating digital multimedia elements such as images, videos, and audio to produce digitally legal evidence"[6]. Multimedia forensics strategies include: 1. Revealing the historical background of digital content, 2. Validating the content's integrity, 3. Source device identification, 4.Retrieving data from multimedia signals, 5. Recognizing unethical human actions from the videos, 6. Identifying forgeries in the image/video files, etc.

1.1.1 Multimedia Forensics Authentication Process

Multimedia forensics strives to analyze the multimedia content such as video, image or audio to generate forensic evidences. Our goal is to detect and localize the forgeries in images and videos from the multimedia forensics perspective. Two approaches are used in image/video forensics analysis to determine the genuineness files: a) active authentication and b) passive authentication.

(a) Active authentication: In this technique, a known authentication code is embedded into the generated image/video by the source device for assessing the integrity of the files at the receiver end. The active authentication techniques require a watermark or a digital signature as an authentication

code. The authentication technique faces certain disadvantages as the authentication technique needs to be embedded with extra code at the time of generating the multimedia contents using a hardware device and many of the multimedia files found on the Internet are not included with a watermark or a digital signature, such files need additional techniques to verify the authenticity and integrity.

(b) Passive authentication: In these methods, no extra code like a digital watermark or digital signature is embedded within the image/video for assessing authenticity and integrity. The passive or blind authentication methods work by considering the traces or clues that are left during the creation of digital forgeries mainly the statistical characteristics of the image/video that are disturbed during the forgery operations.

1.2 Digital Image Forensics

1.2.1 Definition

Digital image forensics is a scientific field that identifies, validates, analyses, and interprets images as a shred of eventual digital evidence. Image manipulation changes the information of the original image and creates forgery images that are not easily identified by human eyes. In digital image forgery, original images are manipulated to create forged images. Digital image manipulation can be used for a variety of purposes, including entertainment, hiding evidence of image tampering, disseminating false information, generating child pornography, and producing fake image evidence in a court of law, etc. The process of digital image forgery results in the loss of authenticity and integrity of the images. Creating a forgery typically involves some processing steps, which leave statistical traces that can be utilized in the image forensic analysis process. Image tampering can be carried out through a variety of different image processing operations, such as compression, adding noise, scaling, filtering, rotation, upsampling, downsampling, resizing, cropping, retouching, and blurring [8]. Original images have certain characteristics like noise variation, brightness/contrast, smoothness etc. The modification of an image content results in the alteration of these characteristics, which causes inconsistency in the image. These inconsistencies obtained from the image characteristics can be calculated in order to detect image forgery. Figure 1.1 shows a typical image tampering example.

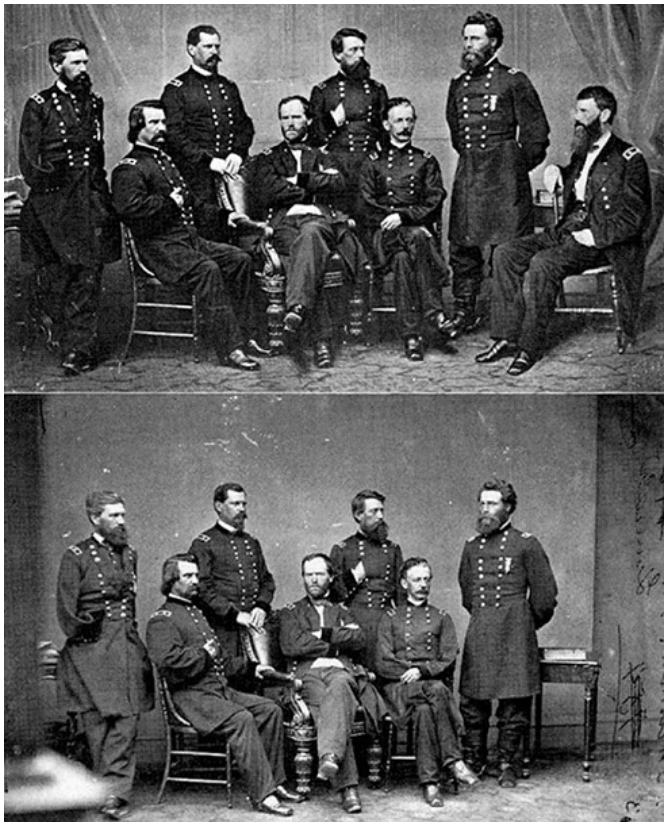


Figure 1.1: Typical digital image tampering.[1].

The image of General Sherman with his generals was captured by Matthew Brady at Circa in 1865. The photograph of General Francis P. Blair (sitting far right) was inserted (above image), as he was not present in the original shot (below image). Digital image forgery detection techniques are being developed to prove the identity and integrity of such image tampering.

1.2.2 Image Forgery Types and Characteristics

The tampering of images changes a region or multiple regions of an image to generate fake content and hide the facts of the original image. The common tampering operations are: deleting, adding, and modifying the contents present in the image. There is a variety of image manipulation types that are carried out on the images and videos. The various image forgery attacks are copy-move, resampling, noise variations, image splicing, retouching, JPEG compression, etc.[2] The most commonly used image forgeries are shown in the Figure 1.2

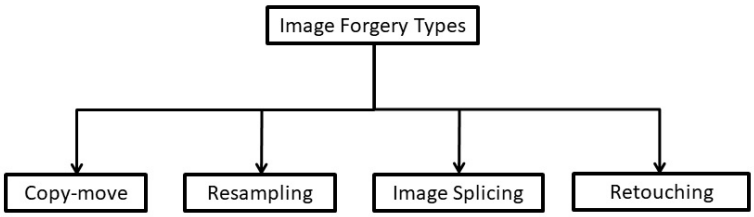


Figure 1.2: Common image forgery types.

The focus of our research is on two frequently used image forgery attacks i.e., copy-move and image-splicing (examples are shown in Figure 1.3).

Figure 1.3: Tampered images: from left to right are the examples showing manipulations of image-splicing(Changing a person) and Copy-move(Copying the fountain)[2].

The act of copying and pasting certain segments of an image onto other parts of the same image, commonly known as the copy-move attack, is the most frequently employed method of image tampering. The primary goal is to detect and locate such types of copy and move.
Resampling forgery is carried out by modifying the geometrical transformation of the digital image such as resizing, rotation and stretching, etc. However, the transformations applied in an image basically leave some traces that are not present in original images. The techniques for detecting the traces that are left during resampling are discussed by Peng et al.[9], showing better results in terms of accuracy.

Image splicing "is generally used as, a substitute for cut-paste in which a composite image is made by cutting and joining the multiple images"[10]. It denotes the region duplication between two images. Image splicing forgery is achieved by merging two or more foreign images to change the original image meaning and generate a forged image. If the image is altered with a malevolent intent, the doctored images can lead to serious social and legal problems.
Retouching the image basically will not completely change the original image; rather it is an enhancement or reducing certain features in the original image. Retouching image forgery is considered as a post-processing operation where only a few properties and characteristics of the image are modified.
The need for image manipulation types can be both beneficial and detrimental. Image manipulation can also be used for commercial objectives, such as creating realistic visual effects in movies, glamorizing an image with image filters for entertainment, or sharing creative ideas. Today image tampering is extremely effective and difficult to detect because of the advanced software tools which are freely available over the Internet. Image tampering operations are not limited to single operations due to advance software tools, images are manipulated in multiple locations using different manipulation types to create realistic views.

1.2.3 Gaps in existing image forgery (copy-move and image-splicing) detection techniques.

Digital image forensic plays an important role in analyzing, maintaining the authenticity and integrity of the digital images. 70% of images that are found over the Internet are forged images and are distributed for some fraudulent gain or to misrepresent the information among the society. There are many research challenges that need to be focused in image forgery (copy-move and image-splicing) detection.

1. Major challenges to be addressed include the authentication, validation, computation complexity, robustness of the post-processing operations, and dimensionality of the features. Localizing forgery in images that have been manipulated do not exhibit visual hints which is also a difficult and time-consuming operation to identify and locate forgeries.

2. There is a need for deep forensic analysis in images which is essential to rebuild trust lost in multimedia contents by using forgery detection techniques.

3. The major problem in digital image forensics is, the accurate detection and localization of image forgeries (copy-move and image-splicing).

4. Many handcrafted feature extraction approaches developed in the literature for the detection of copy-move and image splicing forgery have shown better results on single tampered locations and cannot locate multiple tampered locations with different attacks like copy-move and image splicing etc.

5. To be effective in practical situations, techniques designed to identify image tampering must be capable of detecting any type of modified images, rather than concentrating solely on a particular format type.

6. In image forgery (copy move), there is more importance for identifying effective features and utilizing feature-matching approaches to locate correlation regions (sections).

7. In digital image forensics, detecting copy-move and image-splicing forgeries is a major challenge using deep learning-based methods.

1.3 Digital Video Forensics

In the fast-progressing era of digital technology, video forgery has become an easy and on-going task for users of smart devices. Easily available software tools and smartphones made the task of a naive user effortless for video tampering that may happen in the field of entertainment, crime, social media, medical, political world or intentionally damaging the credibility of an individual. A video forgery can have a significant impact when applied to disparage an individual, cover essential data to prove innocence at crime scenes or escape from legal action. Digital video forensics is a branch of digital forensics that aims to provide tools and techniques that support digital video authentication and integrity verification. Digital video forensics is divided into three categories: a. Source device identification, b. Differentiating forged (or tampered) videos from the original, and c. Video forgery detection. Our main focus of the research is on video forgery detection.

1.3.1 Definition

Video forensic analysis involves scientific investigation, comparison, and/or assessment of video files that are considered as proof in the court of law. The

detection of forgery in videos aims to locate artifacts of tampering thereby evaluating the authenticity and verifying the contents of the video files. Videos that are manipulated affect the authenticity and integrity of the video files because the statistical characteristics of the videos are modified and abnormalities are introduced. The statistical analysis when deeply carried out ensures easy identification of the artifacts or footprints left during the tampering process.
Similar to image forgery detection methods, video forgery detection methods are also classified into active and passive (blind) methods. Passive video forgery detection methods depend on the traces of tampering left during the forgery operations. Active video forgery detection methods are straightforward methods that depend on the advanced information (digital watermark or electronic signature) embedded in the initial stages of video file creation. This results in the reduced quality of the video file and requires specialized hardware to process. In addition, the existing active video forgery detection methods address limited solutions for video forgery operations. Passive video forgery detection methods are improved techniques that can be used to detect various video forgery operations accurately without relying on previously embedded information.

1.3.2 Video Forgery Types and Characteristics

A video is composed of a series of sequential image frames that appear to be identical to one another. The tools and techniques used in digital video forensics prove the authenticity and integrity of the videos. Video tampering is classified into 1. Temporal tampering, 2. Spatial tampering, and 3. Spatial-temporal tampering.
Temporal tempering: Temporal tampering is performed on the sequence of frames. These attacks are mainly affecting time sequence of visual information. Common attacks are frame removal, addition, shuffling, and duplication. Temporal tampering can be at the frame, scene, and shot level
Spatial tampering: Tampering is performed on the frame's content (x-y axis), which shows the changes in the content of the video. The operations in spatial tampering are morphing, cropping, inpainting, replacement, modifying, content addition, and removal. Spatial tampering can be performed at the pixel level or block level. In both cases, the contents of video frames are modified.
Spatial temporal tampering: In this type of tampering, spatial and temporal tampering are both involved. In the same video, frame sequences and visual contents are modified. It is done at the scene level. This tampering involves manipulating

both the visual information along with the time sequences.
The video tampering that occurs in the spatial or the spatial-temporal domain is categorized as intra-frame video forgery and the video tampering that occurs in the temporal domain is categorized as inter-frame video forgery. Fig.1.4 depicts various categories of video forgery operations. The main focus of inter-

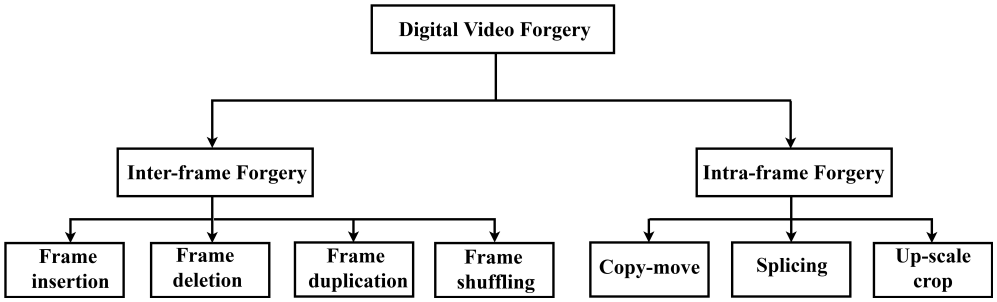


Figure 1.4: Categories of video forgery [1].

frame video forgery is tampering with the sequence of frames within a video. In this type of forgery, the order of the frames is manipulated by means of inserting, deleting, duplicating, and shuffling. Inter-frame video forgeries exploit the temporal-correlations of the video frames and the properties of the traces that are left during forgery operations determine the detection techniques employed. Inter-frame video forgery distorts the sequence of frames in four different ways as shown in Fig.1.5:

1. *Frame insertion:* In this type, a set of frames from foreign videos are inserted at random locations for false actions or evidence. In Fig.1.5b, frames a, b, and c are inserted replacing frames at locations 4, 5, and 6.

2. *Frame deletion:* In frame deletion, a set of frames are intentionally removed from certain locations to prove false evidence in the court. In Fig.1.5c, frames 4, 5, and 6 are dropped from the series of frames.

3. *Frame duplication:* In the case of frame duplication, some frames of the same video are replicated in different locations. Frame duplication is one form of frame mirroring. In Fig.1.5d, frames 7, 8, and 9 are replicated and inserted after frame 3.

4. *Frame shuffling:* To modify the information in the original video, the series of the video frames are modified or scrambled. In Fig.1.5e, frames 3, 4, and 5 are shuffled to change the order of frame sequence.

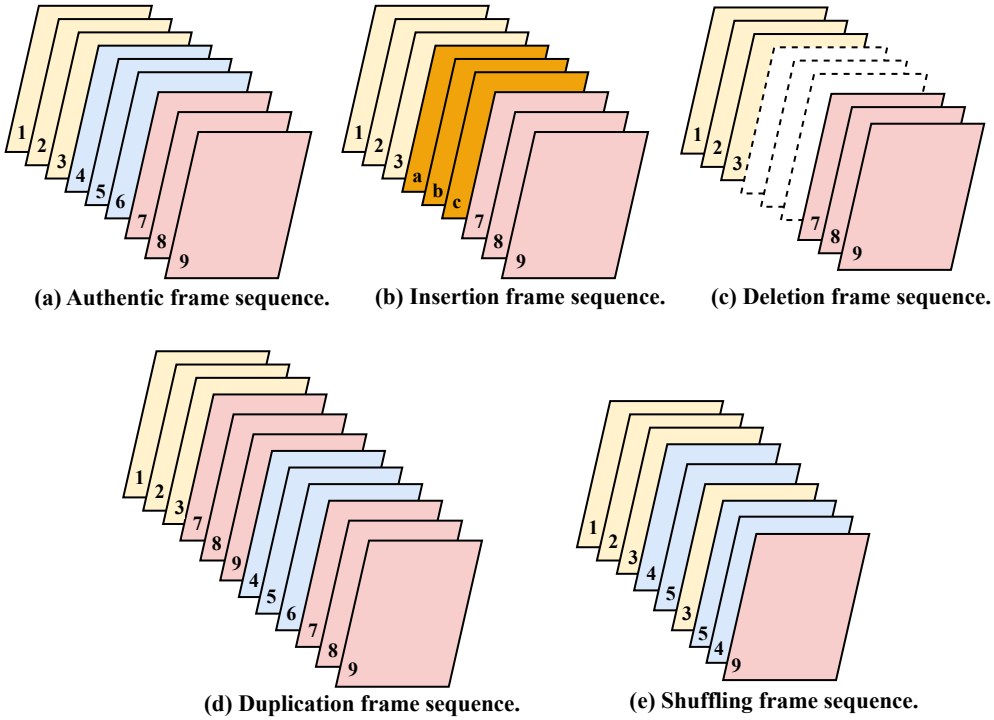


Figure 1.5: Types of inter-frame forgery in the video.

The video intra-frame forgery occurs in the spatial or in the spatial-temporal domain and individual frames are analyzed at an instant of time. Intra-frame video forgery is classified into various categories as 1. Pixel-level, 2. Object-level, and 3. Frame-level. In pixel-level forgery, the visible contents are modified using copy-move, splicing, and re-sampling techniques. In Object-level forgery, a part or object of the video frame is copied from one position to another position within a frame. In frame-level forgery, the manipulation of the frame takes place on the whole frame.

The general forgery operations in intra-frame video forgeries are copy-paste/copy-move, splicing, or by up-scale cropping. In copy-paste, a small section of the frame is copied and pasted into another location of the same video. Copy-move forgery involves copying and moving a particular part of a frame to another part. Splicing forgery in videos involves inserting foreign content or copying video frame contents from another video and pasting them into the current video frames. Splicing video forgery detection is a difficult and challenging task compared to copy-move/copy-paste. The upscale-crop forgery enlarges and crops an original video frame to hide some important content of a crime scene. Considering the

relevance types of video forgeries in cybercrimes, our research work focuses on detecting and localizing inter-frame video forgeries (particularly frame insertion and frame deletion). We assume that frame replication and shuffling do not play a significant role in cybercrimes related to video files.

1.3.3 Gaps in existing video forgery (inter-frame) detection and localization techniques.

Digital videos submitted as evidence in the court are very hard to be trusted, due to the rapid increase in advanced technology, low-cost gadgets, and easy availability of tampering tools. Any layman can manipulate the videos and produce forged evidence in a court of law. There are various challenges in video inter-frame forgery detection and localization that need to be analyzed and addressed.

1. Traditional methods to evaluate inter-frame forged videos (high quality and lengthy) have demonstrated poor performance and lower efficiency.

2. Deep learning (DL) based 2DCNN models are best in extracting spatial features but suffer from temporal feature extraction and result in high computational costs when used to detect inter-frame video forgery.

3. The techniques developed so far for detecting inter-frame video forgery works well on fixed GoP structure and fails to detect forgery which has variable and multiple GoP structure.

4. To identify and localize video processing and manipulation, universal video forgery detection and localization is essential.

5. Techniques that detect forgeries in static backgrounds will not work in dynamic background videos, and methods that detect forgeries in slow-action videos are not able to detect forgeries in fast-action videos.

6. Methods which are designed for the detection of a single type of inter-frame video forgery are not capable to detect forgeries involving multiple types.

7. Many of the video inter-frame forgery detection techniques have limited the number of frame counts in the operation of frame insertion, deletion, or duplication.

8. The existing techniques for detecting and localizing video forgery work only on minimum video length and low-resolution videos.

9. The inter-frame video forgery techniques suffer from the availability of standard datasets to carry out the comparative analysis of forgery operations.

10. To detect and locate forgeries in videos efficiently, the researchers need to address the robustness and computational complexity of the forgery detection and localization techniques.

1.4 Objectives of the Research

The goals of the research work are as following:

1. To develop an image forgery recognition model from the digital image forensics perspective.

2. To identify and locate forgeries in the inter-frame video files (insertion and deletion forgery).

3. To recognize unethical human actions in the videos from digital video forensics perspective.

1.5 Scope and problem definition

Multimedia forensics uses the scientific approaches to analyze the multimedia contents to prove the authenticity and integrity of the files. The major focus is on the analysis of forgery in image and video contents. Multimedia forensics tools are essential due to increase in the multimedia contents generated from various advanced multimedia software tools, digital devices, and social media websites. The growth of massive data have challenged the forensic investigators to analyze and process the data effectively. Many of the approaches for multimedia forgery detection techniques focus on unique forgery detection techniques but in real world multiple image/video forgery operations are carried out to hide the details of forgery from the human visual system. So, there is a need of universal forgery detection techniques.
The proposed models are beneficial to the law enforcement agency in developing forensic tools to process image/video forgery analysis on huge amount of data.

It also helps the forensic analyst to investigate the forgery in crimes related to unethical human action, and pornographic images. Our proposed models are tested on benchmark datasets which are collection of numerous images/videos of multiple types of forgery operations. To develop the proposed model, a fusion of recent approaches in deep learning are utilized.

1.6 Contribution of the thesis

The contributions of this work can be organized into four aspects. The first approach is to analyze images that are forged with easily available software. There are image forensic tools which help the forensic investigator for analyzing forged images, but are not convincing in detecting the type of forgery and forged region. A pre-trained hybrid LSTM-CNN based model is proposed for the generation of binary masks and detect the forged region with improved SIFT algorithm. The bounding box around the forged region classifies the image tampering as copy-move or image splicing.

The image manipulation due to specific intentions like cyber bullying, extortion etc. has increased rapidly in recent times. The need of pornographic image analysis is essential to rebuild the trust on the images by using pornographic forgery detection technique proposed in the second aspect. A three steps process is used in the analysis of porn image forensics. A pre-trained ResNet50 model is used to classify porn images. To detect porn forgery an image, we have used LSTM-CNN model. Finally, the forged object is classified as copy-move or image splicing using template matching with SIFT algorithm.

In the third aspect, we have proposed inter-frame video forgery detection and localization with respect to frame insertion and deletion forgeries. To confirm the integrity and authenticity of the video contents, inter-frame video forgery detection and localization are essential for the forensic investigator from digital video forensics perspective. A deep learning based 3DCNN (3 Dimensional Convolutional Neural Network) model is designed for detecting video inter-frame forgery and to localize the forgery, we have used multi-scale structural SIMilarity (MS-SSIM) index measurement algorithm.

In the fourth aspect, we discuss deep analysis of video files that has become a prerequisite in human action recognition methods concerning to cyber-crime investigation and prevention. A Deep Learning based hybrid model is proposed for unethical human action recognition using two-stream inflated 3D ConvNet (I3D)

and spatio-temporal attention (STA). The I3D model improves the performance of the 3D CNN architecture by inflating 2D convolution kernels into 3D kernels and STA increases the learning capability by giving attention to each frame's spatial and temporal information.

1.7 Outline of the thesis

This thesis focuses on image and video forgery analysis from a multimedia forensics perspective. The thesis is divided into six chapters including an introductory chapter and a concluding chapter. The following is the outline of each of these chapters:
Chapter 1 is the introductory chapter. This chapter provides the necessary background and the motivation for the work reported in this thesis. The chapter begins with an introduction to multimedia forensics and the role of the authentication process (forgery of images and videos). The impact of digital forensics on images and videos is discussed with image/video forgery types. The research gaps in image forgeries (copy-move and image-splicing) and video forgeries (insertion and deletion) were identified from the literature and defined the scope along with the problem definition. Finally, the Chapter concludes with the contributions and outline of the thesis.

Chapter 2 discusses the background and literature survey of image and video forgery detection techniques, as well as the challenges and limitations of forgery detection. Image forgery detection techniques for copy-move and image-splicing are discussed with respect to handcrafted and deep learning-based methods with their limitations. Inter-frame video forgery detection techniques for forgeries such as frame insertion and deletion are discussed from feature engineering and deep learning-based methods.

In Chapter 3, we propose the image forgery detection and localization approach to copy-move and image-splicing forgeries. The objective of our work is to analyze images that are intentionally forged with copy-move and image-splicing to cover the crime scene. The forensics tools available today may help the forensic investigator to analyze the forged images, but they lack in identifying the type and region of forgery[11]. Our objective in this Chapter is to categorize the forged images (copy-move or image-splicing) and locate the areas that were forged. We propose a pre-trained LSTM-CNN[12] based hybrid model to identify forgeries (copy-move and image-splicing) in complex images. The output of this model is

a binary mask. The template matching along with the SIFT algorithm[13] uses the binary mask and the input image to generate the forged object. Then, we generate bounding boxes to show the classification of forgery. The proposed model classified image forgery on benchmark datasets resulting in better performance compared to existing models.

In Chapter 4, we proposed a model for inter-frame video forgery (insertion and deletion) detection and localization from a cybercrime analysis perspective. Extracting efficient features from videos is a significant challenge. In the literature, there are various techniques proposed for the detection of inter-frame video forgeries[1, 14]. We proposed a deep learning-based 3DCNN model[15, 16] to extract high-dimensional features and detect inter-frame video forgeries. Localization of inter-frame forgeries (insertion and deletion) in the video is carried out by spatial-temporal analysis using a multi-scale structural similarity(MS-SSIM)[17, 18] index measurement algorithm for which the original source is not available. The proposed model learns more relevant characteristics to detect video inter-frame forgeries with high classification accuracy and outperforms the existing models in both still and background moving videos without limiting post-processing operations, compression rates, and video length.

In Chapter 5, we explore the recognition of unethical human actions from a video forensic perspective using a hybrid model (deep learning based). We addressed the problem of complex unethical human action recognition by improving the high-level feature learning capability using the fusion of spatio-temporal attention(STA)[4] and two-stream inflated 3D ConvNet(I3D)[3]. The I3D improves the performance of 3DCNN architecture by inflating 2D convolution kernels into 3D kernels and STA increases the learning capability by giving attention to each frame's spatial and temporal information. From the video forensics perspective, the proposed model is unique and is the first experimental demonstration of STA+I3D hybrid model for intelligent unethical human action recognition of complex video files.

Chapter 6 summarizes the contributions made to address the aforementioned three problems. Furthermore, the chapter suggests some research directions for future work.

1.8 Summary

The discussion on digital forensics began in the 1970s when the US Federal Rules on digital evidence were initially introduced. In the mid-to-late 1980s, actual digital forensics investigations commenced as federal agents had to devise methods to examine computers for digital evidence. Arch. group for digital forensic research was started in Utica, NY (August 2001) and named DFRWS (Digital Forensic Research Workshop).
The term Multimedia forensics first appeared in early 2000. By utilizing multimedia forensics techniques, it is possible to verify authenticity, integrity verification/tempering detection, enhancement/restoration, interpretation and content analysis, and source identification. In this chapter, we introduce the concept of digital forensics and multimedia forensics. The impact of digital image/video forensics is discussed along with image/video forgery types. The research gaps in image and video forensics were identified from the literature and defined the scope and the problem definition. Finally, we include the contributions made and gave the outline of the thesis. In chapter 2, we discuss the background and the related work with respect to this research.

Chapter 2

Background and Literature Survey

In Chapter 1, we discussed about multimedia forensics in regard to image and video forgery and then identified the gaps in image forgery and video forgery (inter-frame) detection. In Chapter 2, we go into more detail about its background and related research. Numerous survey papers have been published in recent years on image and video forgery detection. Syed tufael nabi et al.[19], carried out a comprehensive survey on image and video forgery detection techniques, their challenges and future directions which gives a clear understanding to carry out further research in this area. Akhtar et al.[14], provided a systematic and detailed explanation of passive video forgery (intra and inter-frame) detection techniques and elaborated the research work carried out to date with pros and cons. The problems of the proposed methods and datasets are systematically discussed along with future research directions. Judith A. Redi et al.[8] discussed about a collection of tools that can be used to examine the sources of image forgeries and verify the authenticity of the devices used to capture images. Also, the authors have emphasized on the major challenges that are yet to be overcome by the community of researchers working in the field of digital image forensics (i.e., 1. robustness of the existing forgery detection tools and 2. lack of publicly available standard datasets).
According to Qureshi et al.[20], image forensics is a major research topic in security applications, focusing on detecting and authenticating image forgery. In addition, the paper gives a comprehensive overview of different methods for detecting forgery in images, complementing the limitations of existing approaches.

Xiang Lin et.al (X. Lin et al., 2018) provided a brief review on recent advances in passive digital image security forensics and categorized image forensics approaches based on various kinds of traces (1. traces left in image acquisition, 2. traces left in image storage and 3. traces left in image editing).

2.1 Image Forgery detection techniques

The goal of digital image forensics is to identify authentic and forged images. It attempts to restore trust in images by employing forgery detection techniques. Forensic analysis of digital images is critical and is a major topic in multimedia security research. Digital image forgery detection techniques mainly focus on detecting forgery by using various properties of the tampering process. Passive or blind digital image forgery detection techniques are categorized into five types: 1. Pixel-based, 2. Compression-based, 3. Camera-based, 4. Physics-based, and 5. Geometric-based[20].
In practice, pixel-based techniques are generally used and these are grouped into four techniques: copy-move, image-splicing, resampling, and retouching. We limit our study to pixel-based image forgery (copy-move and image-splicing) detection techniques. In pixel-based techniques, the pixels are taken into consideration while detecting the tampering, since the statistical irregularities are occurred during manipulating the images at pixel level. These techniques work on the analysis of inter-pixel correlations that exits from the tampering process directly or indirectly. The copy-move and image-splicing forgery detection techniques can be divided into handcrafted-based techniques and deep learning-based techniques.

2.1.1 Handcrafted based techniques

Image forgery (Copy-move) detection techniques
In the process of copy-move forgery, an area of the image is duplicated and then inserted into a different location within the same image. During the forgery operation of copy-move, correlation within the copied region exists. This correlation artefact are used in various detection methods for detecting copy-move image forgery. The general structure of the copy-move forgery detection is shown in Fig. 2.1. The copy-move forgery detection techniques can be classified into block-based, key point-based, and hybrid-based techniques. The pre-processing stage is common in block-based and key-point based approaches. In block-based

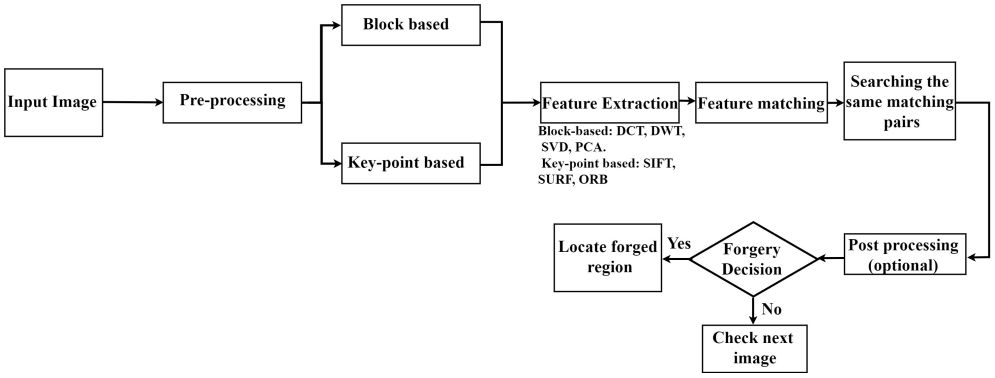


Figure 2.1: General structure of copy-move forgery detection.

techniques, the forged image is divided into overlapping or non-overlapping blocks. This division is followed by feature extraction from each block and these features are matched within the block pairs. During the feature mapping stage, suitable data structures are employed to arrange or organize block-based features, and the determination of forgery is based on the matching of feature pairs of neighboring blocks. Various matching techniques such as radix sort, lexicographical sorting, k-d tree, hash value, and Euclidean distance can be employed in block-based techniques [21].

Local features such as corners, edges, and blobs are extracted from tampered images using key-point-based techniques. A group of descriptors is utilized to enhance the reliability of each feature. Then, each descriptor is matched with another descriptor to find forged regions. In key-point-based techniques, matching techniques are explored using best-bin-first, 2-nearest neighbors (2NN), generalized 2NN (g2NN), Broad First Search Neighbors (BFSN), clustering etc. Further filtering is included for removing spurious matches and an optional post-processing step is carried out that follows a transformation pattern.

The block-based or keypoint-based approaches are used to extract the features, and pairs of similar feature points were matched and then filtered. The two distinct regions are categorized as duplicates if they are densely matched pairings. Both approaches use diverse feature extraction techniques. For block-based method, one can use Discrete cosine transform (DCT)[22], local binary pattern (LBP)[23], Discrete wavelet transform (DWT)[24], Singular value decomposition (SVD)[25], and Principal component analysis (PCA)[26], etc., while the keypoint-based method uses scale-invariant feature transform(SIFT)[27], Mirror-reflection Invariant Feature Transformation (MIFT)[28], speed up robust

features (SURF)[29], and ORB[30] etc. to extract features from an image.
A large number of copy-move forgery detection methods have been reported in the literature. From the block-based copy-move image forgery detection techniques, Wang et al. [31] applied discrete cosine transform (DCT) as a feature extraction technique with package clustering algorithm as feature matching method. The model can locate irregular and tampered regions, but resist in adding white Gaussian noises and suffers from detecting contrast ratio, luminance and color intensity forged images. The DCT and cellular automata were used as a feature extraction technique for copy-move image forgery detection and Kd-tree based nearest neighbor searching algorithm was used as an feature matching technique by Gulnawaz and Qadir [32]. This approach works well even when copy-move forged image is heavily affected by post-processing attacks. The overall detection accuracy achieved is better but, the time complexity issue needs to be addressed in the proposed model. The block-based techniques are good in detecting forged regions with higher accuracy but are affected by high computational complexity and are not robust in detecting various post-processing attacks in copy-move forgery.
In Key point-based techniques, image key points are detected and matched in the whole image for the detection of duplicated regions. These techniques are applied to entire images with high entropy regions (high disorder and low energy). The key point-based techniques are classified into Scale Invariant Feature Transform (SIFT) and Speed Up Robust Features (SURF). Scale Invariant Feature Transform (SIFT) is invariant to forged images which are geometrical and illumination transform.
The authentic image region and the copied region are correlated in copy-move image forgery. This correlation is used as a basis to detect the forgery. With the SIFT algorithm, the precision of detection can be improved and robustness against post-image processing can be increased. SIFT algorithm is the best feature-based matching algorithm that can match post-processing operations like rotation, noise, and scale variation [27]. The SURF algorithm is fast and robust since it works on similarity invariant representation for comparing forged images. The feature descriptor used by SURF is 64-D for every key point and compared to SIFT, SURF is easy to compute. SURF image feature points detector has appeared as an alternative to SIFT. Its main advantage is, it is fastest in computation while keeping a high descriptive power (including repeatability, robustness and distinctiveness)[31]. SURF techniques are better than SIFT in rotation invariant, blur and warp transform. SIFT is better than SURF

in case of different scaling of forged images. Key point-based approaches are computationally efficient when it comes to image compression, Gaussian noise, illumination, and rotation as compared to block-based techniques. However, Key point-based approaches do not work well in smooth background areas that are used to hide small smooth regions, which cannot be extracted effectively resulting in low detection accuracy.

Hybrid-based methods are used to solve the differences of block-based and key-point-based techniques[21]. To make the copy-move forgery detection more efficient and accurate, two or.. or more methods of block-based and key-point based methods are combined to have a hybrid approach. The hybrid-based techniques overcome the limitation of key point-based and block-based methods. Zheng H et al.[33], have proposed an efficient algorithm to detect image forgery (copy-move) using moments of Zernike and SIFT with g2NN feature matching. Results from the experiment revealed that the technique is reliable in detecting smooth regions and post-processing operations when compared to key point-based and block-based techniques. To improve the detection of copy-move images that are forged with varying scale and JPEG compression, feature extraction techniques like Fourier-Mellin and SIFT techniques are used and feature matching methods like g2NN and Patch match algorithm are used by K Bihari et al.[34]. The model results in the less computational time and better performance.

Image-splicing forgery detection

Image-splicing forgery is the image tampering operation of cutting and pasting objects from one or more images generating a new forged image. To make the image-splicing forgery indistinguishable some post-processing operations are carried out on a spliced region of the forgery image. Image-splicing is also called image composition and there is no region duplication in image-splicing. In contrast to copy-move, locating forged regions is a difficult task in image splicing. The detection of image-splicing forgery depends on statistical clues like discontinuity in edges, inconsistency in lighting, disturbances in image bi-spectrum, resampling features, compression etc. The discontinuity in the edges arises when a replicated portion leaves irregular sides in the altered region. The two images don't have the same illumination since there might be some illumination variances between the tampered area and the rest of the image.

Johnson and Farid[35] presented a technique that measures the difference in the image illumination path. A change in the computed path is an indication that the image has been tampered. Poly spectral analysis, sometimes referred to as

bi-spectral analysis or bi-coherence is introduced as a result of the non-linearity of the signal co-relations in signals that can be used to formulate the issue of detecting image splicing forgeries. When used for image splicing detection, the bi-coherence magnitude and phase characteristics showed enhanced detection accuracy[36].
In a digital composite image (image splicing), lighting conditions are often difficult to match. Lighting conditions can reveal traces of digital tampering that can be utilized to detect forgeries in image splicing. Johnson, M. K et al[35] discussed how to estimate the direction of a point light source from a single image to detect image tampering.
The resampling feature is best suited for capturing compression, resampling, and shearing artifacts. Due to interpolation, resampling induces periodic correlations among the pixels. Forgeries involving image splicing and resampling are detected by resampling detection algorithms, while cloning and region removal are detected by copy-move detection algorithms.
Hilbert-Huang Transform (HHT) method was used to exploit the non-stationarity and non-linearity of image-splicing forgery. Furthermore, the moments of wavelet sub-bands were calculated as features to detect image-splicing forgery with high detection accuracy[20]. For image-splicing forgery detection, He et al.[37] used Markov features. Based on DCT's Transition Probability Matrix, these features are derived. The technique achieves detection accuracy of over 91%. An enhanced Markov model is applied in the Block Discrete Cosine Transform (BDCT) domain as well as in Discrete Meyer Wavelet Transform (DMWT) domain for feature extraction[38]. To classify the spliced image from an authentic image, the role of discriminative features for SVM classifier is applied. The experimental results show that the proposed method perform better than existing methods.

2.1.2 Methods based on Deep learning

Early detection methods of image forgery were limited to detecting a single type of forgery since the image forgery type used was unique and the clues left after forgery are also unique. A feature extraction technique and a classification method used were unique in a single type of image forgery detection[10]. The handcrafted feature extraction methods used for detecting single-type image forgery, in which the input image is preprocessed and features are extracted, then

thresholding criteria is applied on the features extracted from the image to map and classify the image as forgery or authentic. In a realistic scenario, a single image is altered by various image tampering operations using advanced software tools. In these forgeries, it is very difficult to detect and locate the traces that are left during multiple forgery operations. Thus, using handcrafted feature extraction methods, it is difficult to detect such multiple manipulations[39]. So, a universal image forgery detection approach is essential that can automatically learn traces left by image manipulation. The use of advanced deep learning (DL) techniques aid to solve the problems of computer vision that motivated to apply deep learning models to detect image forgeries. Deep learning (DL) models have outperformed traditional methods in feature extraction and detection accuracy. The features that are extracted from a picture determine how well the classification system performs. Higher the quality of the features extracted, the higher will be the accuracy. Forgery detection and localization in images are the two important areas of research in digital forensics which have received a lot of attention.

Given an adequate amount of input data, DL-based models can extract features (both complex and abstract) from the manipulated image. However, deep neural network (DNN) training is a challenging task and demands a large quantity of data and high processing capability.

Recently, deep learning architecture is popularly used to solve the problems of hand-engineering-based methods and has shown significant results in many complex cognitive tasks. The major benefit of deep learning is, it can learn complex and massive amount of data. Convolutional neural network (CNN) is one of the most popular deep learning networks for handling image classification using convolutional layers. Feature maps are created for an input image using convolutional filters. Feature maps holds the important features present in the input image. Deep convolutional neural networks are applied in image classification, image forensic, image hashing retrieval, etc., and have shown better performance than the traditional methods.

In [40], convolutional neural networks (CNNs) are used to automatically learn hierarchical representations from RGB color input images for detecting forgeries in image-splicing and copy-move. The experiments carried out on several publicly available datasets demonstrated that the proposed CNN model performed better and more accurate when compared to the other existing models. Wu Y et al. [41], introduced end-to-end deep neural network-based forgery masks and used

a CNN to extract block-like features from an image to detect copy-move forgeries. The experimental results demonstrated that the proposed method achieves better forgery detection performance than classic approaches relying on different features and matching schemes. Junlin Quyang et al.[42] proposed a copy-move forgery detection method based on a convolutional neural network that uses an existing trained model using a large database such as ImageNet. Even though the authors obtained good experimental results, the method is not robust to real scenario of copy-move image forgery detection.
To eliminate the laborious feature engineering process, a Convolutional neural network was used as a way of learning directly from the available training data [43]. The proposed method detects photographic splicing and locates forgery regions yielding a classification accuracy of more than 95%. The convolutional neural network-based detection methods explored the differences of image attributed between un-tampered and tampered regions in an image. The ringed residual U-Net (RRU-Net) with CNN was used for image splicing forgery detection [44] to strengthen the learning procedure. The results of the method were promising as compared to other splicing forgery detection methods. Semantic segmentation in deep learning models has shown good performance results by learning hierarchical features of different objects in an image[45]. But, semantic segmentation techniques segment only the meaningful objects within the images and could not segment objects of different image manipulation operations. To overcome the problems of deep learning with low robustness and complexity in image-splicing forgery detection, a multiscale lightweight image-splicing forgery detection model was proposed [46]. The MobilenetV2 model is used as the backbone network for improving the image-splicing forgery detection performance through skip connections. In comparison to other detection techniques, the experimental outcomes indicate that the suggested model exhibits commendable performance. The model needs to ensure better accuracy and capability to handle the complexity of the datasets.

2.2 Video Forgery Detection Techniques

Passive forgery detection techniques do not use the video metadata, instead they use the traces of forged content that are left during the generation of the fake video. By extracting features from video frames, passive inter-frame forgery detection techniques can be used to verify the authenticity of video files. Video

editing will dispense some traces of forgery that can be used to verify its authenticity. These traces can be: 1. More prediction errors, 2. High frame intensity values between temporal and spatial correlation, 3. Motion residues and noise, 4. Optical flow abnormalities, 5. Motion vectors, 6. Frame quality, 7. Variation of prediction footprint (VPF), and 8. Motion-compressed edge artifact (MCEA). Figure1.5 shows four different types of tampering attacks that can be detected by passive inter-frame video forgery detection methods. The typical process for creating video forgeries involves, breaking the video into individual frames, and then manipulating them through actions such as deletion, insertion, and replication. Finally, recompressing the altered video. There have been many different methods suggested in the literature for identifying inter-frame forgery in video sequences [14]. Most of the techniques are based on manual feature extraction to identify the artifacts of forgery and these features are sensitive to post-processing operations (light inconsistencies, noise, compression, blurring, etc.)[47]. Many authors have earlier suggested various methods to detect digital inter-frame video forgery for video forensics. We divide these methods into two groups as, 1. Feature engineering-based video inter-frame forgery detection methods and 2. Deep learning-based video inter-frame forgery detection methods.

2.2.1 Feature engineering-based video inter-frame forgery detection methods

Stamm et al. [48], presented an approach for detecting frame deletion or insertion operated by an increase in motion prediction error in P-frame and this method was used to design as an anti-forensic technique to make digital forgeries (deletion or insertion) undetectable by forensic techniques. The proposed model fails to locate deleted frames and only works with fixed group of pictures (GoP) sizes. A novel detection scheme for video inter-frame forgery is suggested by Chao et al.[49]. The insertion forgeries were detected using local feature-based (window-based) and logarithmic search (binary search) based methods. Optical flows and double adaptive thresholds were used frame-by-frame to identify deletion forgeries. The model performs well in insertion forgery detection as compared to deletion forgery detection. Wang et al. [50], proposed optical flow analysis to authenticate and identify inter-frame forgeries (insertion, deletion, and duplication) in digital videos. The proposed technique was robust to some degree on MPEG compression but had limits with respect to computational cost.
A double encoding detection method called the Variation of Prediction Footprint

(VPF) was used by Gironi et al.[51] to detect frame insertion. The proposed method cannot detect frame alterations when the attacker deletes an entire GoP and it was also unable to locate accurately the deletion and insertion forgeries. Sitara et al. [52], identified the abnormalities in video inter-frame forgeries (insertion, deletion, shuffling, and duplication) using inconsistency in velocity field and VPF, achieving a detection accuracy of 92.3%. The model was not tested on videos with different quantization scale and on videos with a moving background. Inter-frame forgery detection in MPEG-2 and H.264 encoding was proposed by Kingara et al. [53] by utilizing the footprints from motion, brightness gradient features, optical flow coefficients and prediction residual. The performance of the method suffers when used on videos with strong light source, dynamic videos and variable GoPs. Based on the similarity analysis, a passive-blind approach for detecting video forgery (inter-frame) was proposed by Zhao et al.[54]. In this, HSV(Hue-Saturation-Value) color histogram matching method and the SURF(Speeded Up Robust Features) technique was used for feature extraction. These two features are combined with Fast library for approximate nearest neighbors(FLANN) algorithm for detecting forgery. However, the method requires source videos to be correctly obtained. This technique was not capable of detecting complex inter-frame forgeries of videos containing many video shots.

V Kumar et al.[55], proposed inter-frame video forgery detection using correlation coefficient distance between the video frames by calculating minimum distance score and used dual-threshold to differentiate the type of forgery (insertion or deletion). The proposed model could achieve 97% accuracy, but had a limitation with the number of videos to process.

Video frame tampering is a common forgery operation today, where the frames are inserted or removed to modify the information in the videos. To detect forgery artifacts generated during the compression were explored by Xiao Jin et al.[56] based on high-frequency features of reconstructed DCT coefficients. The proposed algorithm detects tampering of frames in videos and locates the frame tampering point in the series of video frames. The problem of noises in illumination and jitterness occurring in real-time videos was explored by Han Pu et al.[57] using a robust optical flow algorithm that uses noises from jitterness and severe brightness of inter-frame forgery videos. The proposed optical-flow algorithm extracts features that have changes in the frame texture and the videos having more jitterness were detected with motion entropy of variable thresholds. The performance results of the method were compared with existing methods on 200

videos obtaining accurate and robust results. The feature engineering-based analysis methods work well on videos with less number of frames and cannot handle the videos when the number of frames are more. In addition, these methods are computationally complex and consume a lot of duration in processing, affecting the performance evaluation metrics.

2.2.2 Deep learning-based video inter-frame forgery detection methods

Deep neural networks have the ability to extract complex high-dimension features from the input image patches and make efficient representations to identify videos that are tampered with. Deep learning encourages researchers to use the models of machine learning and deep learning to investigate forgeries in the field of digital video forensics. Deep learning is the subset of machine learning that can extract the features automatically without an external feature engineering process. Many deep learning-based algorithms are proposed for forensic analysis in the image and video forgery till date[19]. The deep-learning-based methods, particularly CNNs have exceptionally achieved better results in many computer vision problems, specifically in large-scale image classification. Recently, efficient CNN models are used for detecting video inter-frame forgeries. Long C et al.[58] were the first to use CNN model for detecting frame duplication forgery using a coarse-to-fine deep convolutional neural network. In this model, when a video contains multiple sequences of duplicated frames, the model's performance degrades. Kaur et al. [59], proposed a highly efficient method to detect an inter-frame forgery in videos using in-depth CNN that utilizes spatial and temporal correlation between the frames and identifies the abnormalities within the frames. CNN models using a transfer learning approach were suggested by Xuan Hau et al. [60] for detecting video forgeries in inter-frame category. However, the approach was unable to extract the temporal features accurately. The authors proposed a Video Inter-Frame Forgery Dataset (VIFFD) that includes insertion, deletion, and duplication forgeries for experimental analysis. Neetu Singla et al. [61], proposed a feature engineering-based machine learning approach for detecting frame deletion forgery. In this, to predict the authenticity of video shots, three machine learning models were used: Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Convolution Neural Network (CNN). The results of this approach reveal that the CNN model is more accurate than SVM and MLP in distinguishing genuine and fake sequences. Mamtora et al.[62], used LSTM framework to identify and localize

spatial and temporal alterations in video. The experimental results were validated using the REWIND dataset, which included 10 forged and 10 authentic videos. The model's efficacy was shown at the pixel, frame, and video levels. Jhonston el al.[63] used Convolutional Neural Network as a framework for detecting features from H.264 video sequence. The features extracted were used for localizing the key frames of tampered areas in the forged video with better accuracy. However, the model proposed was limited to a single type of video forgery with a fixed GoP size and static background.

2.2.3 Summary

In this Chapter, we have explained the background details of image and video forgery detection techniques and their related works that exist in the literature. Image forgery detection techniques for copy-move and image-splicing are discussed with respect to handcrafted and deep learning-based methods with their limitations. Video forgery detection techniques for inter-frame forgery (insertion and deletion) approaches are discussed from feature engineering (feature extraction) and deep learning-based methods. The first objective of our work will be discussed in Chapter 3, i.e., detecting and locating forgery in images with respect to copy-move and image-splicing forgeries.

Chapter 3

LSTM-CNN based hybrid model for image forgery detection and localization

In Chapter 2, we discussed the background and literature survey of image and video forgery detection techniques and identified the challenges and limitations of forgery detection. In this Chapter, we design a model for image forgery (copy-move and image-splicing) detection and localization. Multimedia forensics is essential to analyze multimedia contents to produce authentic, and integrity reports as evidence in the court. One of the objectives of our work is to analyze images that are intentionally forged with cloning and cut-paste to cover the crime scene. The tampering of images is done using advanced and freely available software tools. The main idea is, forged images shouldn't be distinguishable from normal human eyes and generate fake content to cover up historical facts. Image tampering techniques can be beneficial for creating visual effects in movies, glamorizing an image with image filters for entertainment, or sharing creative ideas. At the same time, forged images can be used to humiliate the female gender by creating fake porn images and can also be used to provoke individual emotions causing serious problems. The forensics tools available today may help the forensic investigator to analyze the forged images, but they lack in identifying the type and region of forgery.

Digital image forensics aim is to detect forgeries in images and to build the trust lost in images and videos due to sophisticated image forgery tools. Image tampering refers to a distinct forgery type that alters or distorts a region or more

than one region of the image by applying multiple forgery operations. The most prevalent image forgery operations are image-splicing and copy-move. In copy-move, particular portions from the image are copied and pasted onto other parts of the current image. Finding more than two comparable areas within a single image is the main task of the copy-move image forgery detection technique.
Image-splicing "is a substitute for cut-paste in which a composite image is made by cutting and joining the multiple images"[64]. The detection of image-splicing forgery is more complex compared to copy-move forgery since, objects copied in the same image have similar contours with size, texture, transitions, etc. But, in image-splicing, external image content is added with different sizes, transitions, textures, etc., which makes it more difficult to detect the forgery. Various image-splicing forgery detection methods have been put forth by different researchers in the literature[11]. Image forgery detection techniques based on Handcrafted feature extraction are generally used but they are limited in detecting multiple forgery regions and are generally hard to detect the location and type of image forgery.
Recently, deep learning-based approaches have outperformed handcrafted feature extraction techniques, as they are better at extracting features and have shown high accuracy in forgery detection. The image forgery detection accuracy is higher when the quality of the image feature extraction is higher. Detecting image forgery and localizing the forged region in images have received more emphasis in the digital image forensics area. Among the various deep learning-based methods Convolutional Neural Networks (CNN)[65] and Recurrent Neural Networks (RNN)[66] have shown better results in pattern recognition of images. Semantic segmentation in deep learning models has also shown good results by learning the hierarchical features of various objects within an image. But, this technique segment only the meaningful objects within the images and could not segment objects of different image manipulation operations[45]. The artifacts generated during the image tampering operations like compression, resampling, and shearing are better captured by resampling features. The interpolation that occurs during forgery operations induces some periodic correlations among the pixels. The CNN approach is highly effective in producing spatial feature maps for distinct regions within the image and resampling features can be useful for capturing certain forgery artifacts. The spatial feature maps and resampling features can be combined and utilized to locate tampered regions.
To identify forgery in images, many conventional methods have been devised over time. Sami Bourouis et al.[10] contributed to the taxonomy and a review of the

current developments in the area of multimedia forensics. In addition, the authors investigated the way discrepancies affect the characteristics of the forgery image. A doubly stochastic model (DSM) based technique was proposed by Dua et al.[67] to differentiate original images from the copy-move and image splicing attacks and localize tampered regions. To localize a copy-move attack, features of each block are identified using phase congruency, and to localize a splicing attack block-wise correlation maps of dequantized DCT coefficients were used. Due to multiple techniques, the complexity and overhead of the model was increased. To identify fake image regions, Anuj R et al.[68] suggested an improved SURF and template matching technique. The advanced SURF technique detects copy-move-based image forgeries, and template matching identifies the spliced and cloned portions of the input image. When images are scaled differently or have smooth background areas, the SURF approach fails.

Deep learning-based image forgery detection techniques perform better than traditional image forgery detection techniques and have been applied in various categories like image classification, digital image forensics, object detection in images, etc. Junlin Quyang et al.[42] proposed a convolutional neural network for detecting copy-move forgery. The model uses the parameters from the existing trained model on ImageNet database, then the network structure is adjusted using copy-move training samples to test the forgery images. The model is not robust to real-world forgery images. The effectiveness of forgery detection in image-splicing and resampling forgeries is achieved by using resampling detection algorithms. Mohammed et al.[69] demonstrated better detection accuracy when combining resampling and copy-move algorithms. The algorithms of resampling are effective in detecting resampling and image-splicing forgeries. Peng et al.[70], used CNN as a dual filtering network structure for extracting the resampling features from the images. The proposed network was proved effective in capturing resampling artifacts for classifying the images but is limited only to uncompressed images. For the detection and localization of image manipulations, Bunk et al.[71] proposed methods with a combination of resampling features and CNNs. First, a Radon transform was employed to compute resampling features on image patches. Radon Walker segmentation technique is used to locate tampered regions. Next, a forged heat map was generated by utilizing a hybrid model comprising of a Gaussian conditional random field model and a deep-learning CNN model. Finally, to classify tampered regions, resampling features were given to LSTM network.

By combining resampling features, long short-term memory (LSTM) cells, and an encoder-decoder network, Bappy et al.[12] proposed a high-confidence image

manipulation detection and localization method. The encoder network exploits spatial feature maps and the LSTM network analyses the correlation between the manipulated and non-manipulated image patches in the frequency domain. For image tampering localization, the decoder network learns how to map low-resolution feature maps to pixel-wise predictions. The dataset employed was simple and involved a complex procedure for creating synthesized data for image manipulations. Furthermore, the model demonstrates segmentation of manipulated regions while ignoring forged object features such as multi-scale, rotation, illumination, image noise, and affine variations. The hybrid model (LSTM-CNN) that we proposed is a generalized one to train complex datasets such as NPDI, CASIAv1.0, CASIAv2.0, Columbia, COVERAGE, CoMoFoD, and MICC-F600. We apply template matching with an improved SIFT algorithm to make the model invariant and tolerant to detect and localize forged objects (copy-move or image-splicing) under various post-processing operations.

3.1 Challenges in image forgery detection

Digital image tampering has become extremely popular in recent years due to the availability of easy-to-use software enabling users to edit, copy, and resize images. Thus, it is challenging to authenticate and validate the integrity of images. Identifying forgeries in manipulated images that do not exhibit visual hints is a challenging and time-consuming process. The deep learning-based methods used for image classification tasks inspire us to design a deep learning model that can detect and locate manipulated regions in an image. Convolutional neural networks (CNN) is one of the popular deep learning algorithms that exhibit high performance while analyzing areas of images in various image recognition tasks like object detection, image classification, semantic segmentation, etc. Using only CNN to detect and locate tampered regions may not be the most effective strategy. In image manipulation, objects are removed, copied, or inserted into the image in different locations. In addition, the use of modern image tampering tools hide the artifacts and make it difficult to differentiate between fake and original image.

3.2 Contribution

The major contributions of our work include:

- We propose an LSTM-CNN based technique to detect and localize image forgery attacks (copy-move and image-splicing) in complex images.
- To test the proposed model on pornographic images, we used the NPDI database[72]. From this dataset, we have manually forged porn images with an image splicing attack. In addition, we have generated ground-truth masks for the forged images.
- To localize forged objects, we have used the template matching with an improved SIFT algorithm. We, then generate bounding boxes to show the classification of forgery.

The ground-truth masks aid us in back-propagating the error and hence resulting in learning the network parameters easily. In our proposed model, we train deep neural networks using a large dataset for image classification and localization tasks. Forgeries of more than one type (copy-move or image-splicing) can be detected and localized using the proposed model. Forensic analysts can use the proposed model to detect porn image forgeries in cybercrime investigations.

3.3 Methodology

In this work, our aim is to detect, localize, and classify images that have been modified (copy-move, or image splicing). LSTM network and CNN-based architectures are combined with resampling techniques to detect complicated manipulations in images and locate tampered regions. It is possible to effectively capture image manipulation footprints such as down-sampling, up-sampling, JPG quality loss, and image shear using resampling features. LSTM models are frequently employed to acquire knowledge of the temporal context present within a video or any other form of sequential data. A pre-trained hybrid LSTM-CNN model[12] is used by our network architecture to generate binary masks. The binary mask generated is used to crop the tampered object and match it with the input image. An image is classified as a copy-move forgery if there is more than one similar object in the image otherwise, the image is classified as image splicing. Bounding boxes are generated to show the classification of the forged region. The approach we propose shows promising results in pixel-level localization of manipulated regions on difficult and complex datasets (CASIA V1.0, CASIA V2.0, and NPDI pornography). Figure3.1 depicts the overview of the proposed methodology which

is divided into three stages: 1. Hybrid LSTM-CNN, 2. Forged object detection, and 3. Classification of forgeries.

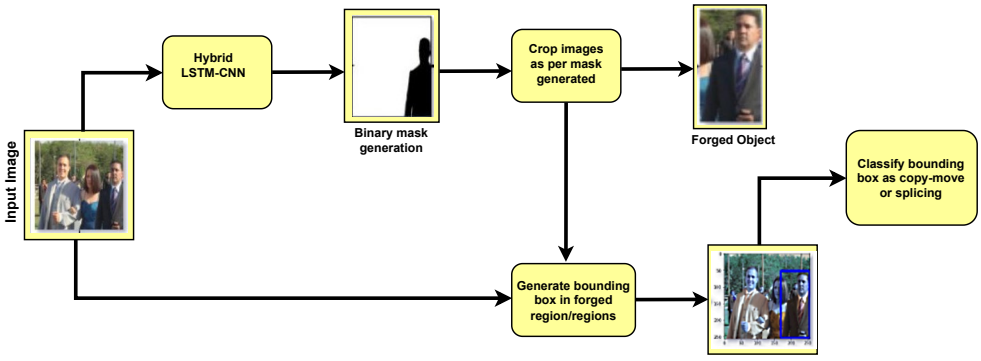


Figure 3.1: Overview of the proposed model for detection, localization, and classification of image forgeries.

3.3.1 Hybrid LSTM-CNN

Hybrid LSTM-CNN uses images divided into patches or blocks as input information. Resampling features undergo Radon transform computation on blocks of images. and are provided to the LSTM network as input for learning the relationships between various sub-groups[71]. LSTM is typically used when information is sequential. In order to maximize LSTM performance, patches should be ordered in a specific sequence, e.g., keeping the sequence of extracted patches. In order to preserve the spatial location of the extracted patches and their original shape, we use the Hilbert curve[73] along with the extracted resampling features. This is done to identify the sequence of patches that are fed to the LSTM cells. In the frequency domain, LSTM cells acquire the ability to comprehend the transition between modified and unmodified patches. A set of resampling feature maps is generated by the LSTM network. The encoder network generates the spatial features from the input image to detect modified areas. Encoder networks are created using convolutional layers, which enable the network to understand the appearance, shape, and spatial relationships between manipulated and non-manipulated classes. The feature maps' spatial resolution, however, is lost during the convolution process. By substituting the fully connected layer with the decoder network, losses can be compensated. The encoder output (spatial feature maps) and LSTM output (resampled feature maps) are combined together and

used as inputs to the decoder network in order to determine the nature of manipulation. Segmenting tampered areas is the major function of the decoder network, and each decoder performs specific tasks such as increasing spatial resolution, batch normalization, and convolutions. The decoder network generates a binary mask indicating the forged region pixel-wise. The finer details of each pixel manipulated or non-manipulated are learned and classified through the decoders.
To estimate the modified pixels against non-manipulated pixels, a softmax layer is used to convert a vector of K real values that follows a probability distribution over distinct classes as $\tau(Y_k)$ from an input K real value. We can get the predicted label by $\overline{Y} = argmax\,\tau(Y_k)$. The process of end-to-end training is employed to classify individual pixels, utilizing ground-truth masks of modified regions. A cross-entropy loss is calculated and utilized to learn parameters via the back-propagation algorithm. The loss is determined by:

$$L(t) = -\frac{1}{P}\sum_{p=1}^{P}\sum_{q=1}^{Q} f(Y^p = q) log(y^p = q \mid y^p; t)$$

where, P and Q are the image matrix and the set of instances respectively. The input element is defined by y, and the function $f(.)$ is an indicating variable that implies 1 if, $p = q$ else corresponds to zero (0)[71]. Weights have been adjusted from 0 to 1. To reduce network loss, the adaptive moment estimation (Adam) technique is applied. In each iteration, a small group of batches is computed to modify the network variables. After the model has learned the optimal parameters, it will be used to estimate the pixel-by-pixel categorization of a specific test picture.

3.3.2 Forged Object Detection

The forged object in the original input image is found using the binary mask generated. Segmentation operation is carried out to group pixels into more meaningful regions. We apply the background subtraction method cvAbsDiff(sr1, sr2, otp) to segment an image by just performing an image difference that results in a forged region, where, sr1 and sr2 are the image sources and otp is given by output otp = sr1 – sr2 [74]. The threshold function is used to mark all the details of the background to black pixels by setting it to zero. Next, we convert this image to a gray-scale image. The image which is in grayscale contains a forged area with pixel values within the range 1-255 and a non-forged area with a value of 0. We apply the threshold function which is a low-level vision for the

spatial domain that calculates the image output based on the threshold value "q". The threshold function is given by:

$$doublecvThreshold(sr, otp, q, mx, typ) :$$

where sr is the image source, the otp is the image output, q is the value of threshold, mx is the maximum value and the typ parameter specifies how the output image is computed. We used

$$CVTHRESHBINARY : otp = maximum,$$

otherwise 0 as the threshold value. The output is a threshold image in which the pixel values are allocated depending on the state of the pixels. The threshold values over $q = 100$ are allocated a value of 0 and the remaining are allocated a value of 255. Finally, we crop the forged object from the results obtained using the threshold function and background subtraction. Hence, we can recognize the forged object and trim the forged area.

3.3.3 Forgery Classification

In forgery classification, the main goal is to classify the forged object as copy-move or splicing. The forged object obtained is compared with the whole input image and find if there exist any other patterns of the same forged object. Pattern recognition is used to determine whether a given image contains a known pattern. Matching an image involves locating a similar pattern in the given image. "Template matching technique is the simplest way to do pattern recognition, in which the process makes an exhaustive search of the template in the source image and marks each position where the pattern is found"[74]. Matching of the template is performed using:

$$cvMatchTemplate(sr, temp, result, meth);$$

where sr is the image source, the pattern to be located is referred to as $temp$, and the $result$ is an output of image for matching. The $meth$ describes the way template matching is carried out. We use the $CV\_TM\_CCOEFF\_NORMED$ method as the template matching function. This method is known as correlation constant value matching that subtract the correlation between the source image and the template from their average value at each point, taking the size of the template into account. The results may not be good when the energy of the

image,$\sum f^2(a,b)$ changes with location. The correlation value is calculated using equation 3.1.

$$R_{ccoeff} = \sum_{kl} \Big[(t(k,l) - \overline{t}) * (f(a+k,b+l) - \overline{f}) \Big]^2 \tag{3.1}$$

To work the method well even if there are changes among the source picture and template, the normalized aspect is considered using equation 3.2.

$$NORM = \sqrt{\sum_{kl} t^2(k,l) * f^2(a+k,b+l)} \tag{3.2}$$

Using the template matching function, it is possible to locate all the objects that are equal in shape and direction. However, it is difficult to locate the forged areas that are in different directions and rescaled. To locate the forged area that is rescaled and rotated, an improved SIFT[13] algorithm is used which is incorporated with ORB(Oriented FAST and Rotated BRIEF) and RANSAC (Random Sample Consensus) algorithms. SIFT generally performs better than the SURF in terms of computational cost and is more accurate than SURF when it comes to rotation, scale, illumination, and image noise. In SURF (Speeded-up Robust Features), more features are detected than in SIFT, but they are dispersed over the image, which makes it more computationally thirsty.
The SIFT algorithm generates the set of features from an image by 1. constructing the difference of Gaussian (DoG) in scale space, 2. finding local extrema (removing unstable feature points), 3. finding accurate keypoint localization, 4. determining the direction of keypoints (orientation assignment), and 5. generating local image descriptor[13]. SIFT algorithm generates feature points and passes only the interest points to the ORB algorithm. In the ORB algorithm, interest points are matched using Hamming distance to produce an ORB descriptor. ORB detects the features that are more concentrated on corners and is computationally faster and more efficient than SURF and SIFT[75]. ORB uses the rBRIEF(Binary Robust Independent Elementary Features) algorithm[76] as rotation invariant by converting the image patch as a binary vector. The BRIEF works at the pixel level and is highly profound to noise. It flattens the image by Gaussian kernel to reduce the sensitivity and increase the steadiness of the descriptors.
RANSAC algorithm[77] is considered as post-quality enhancement stage for removing redundant key points. It removes the image noise on both inliers and outliers drastically and considerably reduces the matching time. The RANSAC

algorithm has proven highly effective at removing unmatched sets of points. The major goal of combining the ORB and RANSAC algorithms with SIFT is to obtain good matching, increased efficiency, and better accuracy in feature point matching.
With the improved SIFT algorithm, all the patterns of the forged object can be detected accurately with scale and rotation invariance. The bounding boxes are shown over the manipulated area. If bounding boxes are more than one then the tampering type is copy-move else, it is the image-splicing type. With the experimental results, we observed that the model obtains forged objects with approximately around 85% of overlaying the binary mask with a ground-truth mask. The workflow of the proposed model is shown in Figure 3.2.

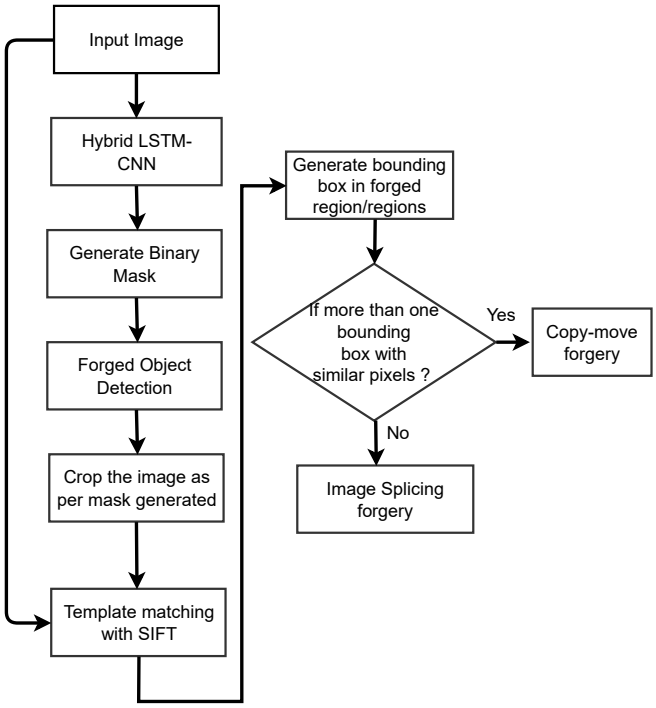


Figure 3.2: Workflow of the proposed model.

3.3.4 Result and analysis

We have conducted the experiments to demonstrate the efficiency of our proposed method on tempering detection, localization, and classification. We evaluate our proposed model on benchmark datasets- Columbia[78], CASIA[79], CoMoFoD[80], COVERAGE[81], and NPDI[72].

3.3.5 Datasets

3.3.5.1 Dataset Preparation:

CASIA Image Tampering Detection Evaluation Database[79] includes color images in versions 1.0 and 2.0. The details of the CASIA datasets are shown in Table 3.1. "The v2.0 database is more challenging and comprehensive compared to v1.0". The tampered images in v2.0 are more realistic to human eyes. We have

Table 3.1: Details of the CASIA dataset.

Dataset	Authentic	Tampered	Total
CASIA v1.0	800	921	1721
CASIA v2.0	7200	5123	12323

selected 5526 forged images from CASIAv1.0 and CASIAv2.0 datasets.
From a porn image forgery database perspective, we have used NPDI [72] dataset which is a pornography database that contains pornographic and non-pornographic images and videos. For our experimental analysis, we used images only. The NPDI image dataset consists of three sections, i.e., 1. The non-porn difficulty, 2. Non-porn easy, and 3. Porn. A total of 16727 images are present in the NPDI image dataset. The dataset does not contain porn-forged images which are essential from our proposed model's perspective and research. So, we have selected 200 porn images (from non-porn and porn categories) and applied the FaceSwap algorithm to generate forged images. Thus, generating 3596 porn-forged images.
The datasets CASIA and NPDI do not have ground-truth masks for the detection of image tampering. We manually generated the ground truth masks for these datasets based on the correlation that exists between the authentic and forged image. These are mentioned within the filenames of the images. The dataset details of CASIA and NPDI forged images with corresponding ground-truth masks used in the proposed model are shown in Table 3.2. A total number of 18244 images are used for training our proposed model. During the data preparation, the dataset is subdivided into a ratio of 80:20 (80% training and 20% testing).

Table 3.2: Details of CASIA and NPDI datasets.

	CASIA1.0	CASIA2.0	NPDI
Forged images	869	4657	3596
Ground-Truth Mask	869	4657	3596
Total images	1738	9314	7192

3.3.6 Experimental analysis

We use python, Keras 2.6 with Tensorflow 1.13 to implement our proposed model. To speed up the data training in our model, we used a system setup that consists of NVIDIA GeForce GTX 2080 Ti with 8GB memory, and Intel Core CPU i7-9700K with 16GB RAM.

Evaluation metrics:

The model's capability is evaluated based on each image by pixel level, which is done by categorizing each pixel into four types. True Negative (TN), True Positive (TP), False Negative(FN), and False Positive (FP). These are used for evaluating our proposed model's accuracy, precision, recall, and F1 score. The actual pixel vs predicted pixel is shown as a confusion matrix in Figure 3.3.

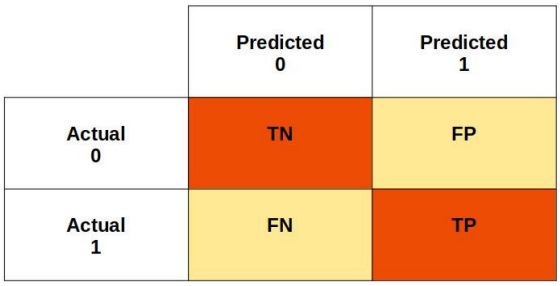


Figure 3.3: Actual pixel vs predicted pixel

TP indicates how many pixels were classified as true positives, FN indicates how many pixels were classified as false negatives, the forged pixel is incorrectly classified as authentic, FP indicates how many pixels were classified as false positive, the authentic pixel is incorrectly classified as forged, and TN indicates how many pixels were classified as true negative.

The ground truth masks contain the following pixel values:

0–$Nonforged \quad pixel$ and

1–*Forged pixel*.
The model's accuracy is measured by the percentage of authentic and forged pixels correctly detected from a mixed dataset. It is the ratio based on the number of correct predictions to the overall predictions.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision is a metric that evaluates how accurate a model is at categorizing an example as positive. The precision is the ratio based on the number of accurately identified positive examples to the overall positive examples (either correct or incorrect).

$$Precision = \frac{TP}{TP+FP}$$

The model's ability to recognize positive samples is measured by recall. The higher the recall, the more positive samples are detected. The recall is higher when many positive patterns are detected.

$$Recall = \frac{TP}{TP+FN}$$

The balanced F-score or F-measure is also known as the F1 score. The F1 score can be considered as a weighted average of precision and recall, with the best score being 1 and the worst score being 0.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

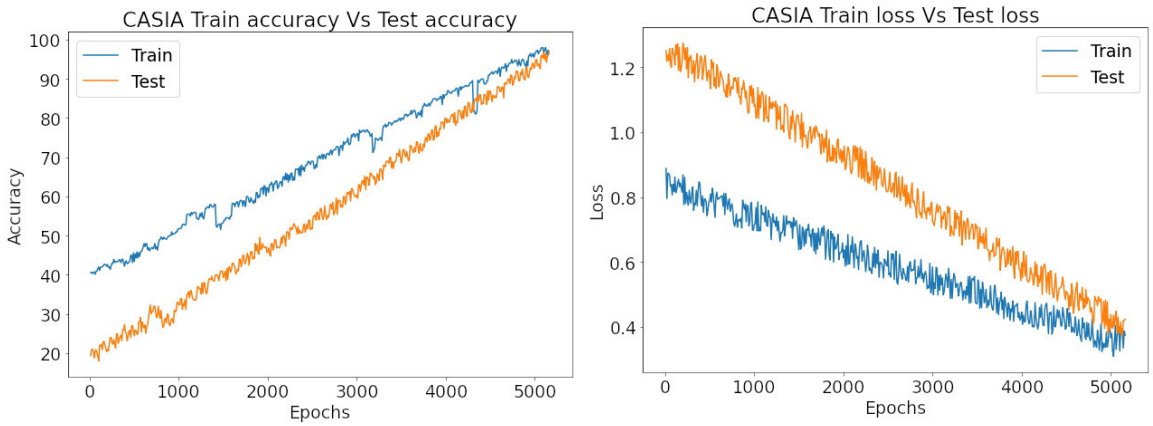
Training and Testing Results:
The proposed model is evaluated for its effectiveness on both the CASIA and NPDI datasets on the same training and testing split (i.e., 80% and 20%) as shown in Table 3.3 & Table 3.4. The model's train Vs test accuracy and loss on CASIA and NPDI datasets are shown in Figure 3.4a, 3.4b, 3.5a and 3.5b.

Table 3.3: Proposed model's training performance accuracy, precision, recall, and F1 score.

Dataset	Accuracy	Precision	Recall	F1 score
CASIA	0.987	0.948	0.843	0.8
NPDI	0.984	0.925	0.80	0.76
Overall	0.985	0.936	0.821	0.78

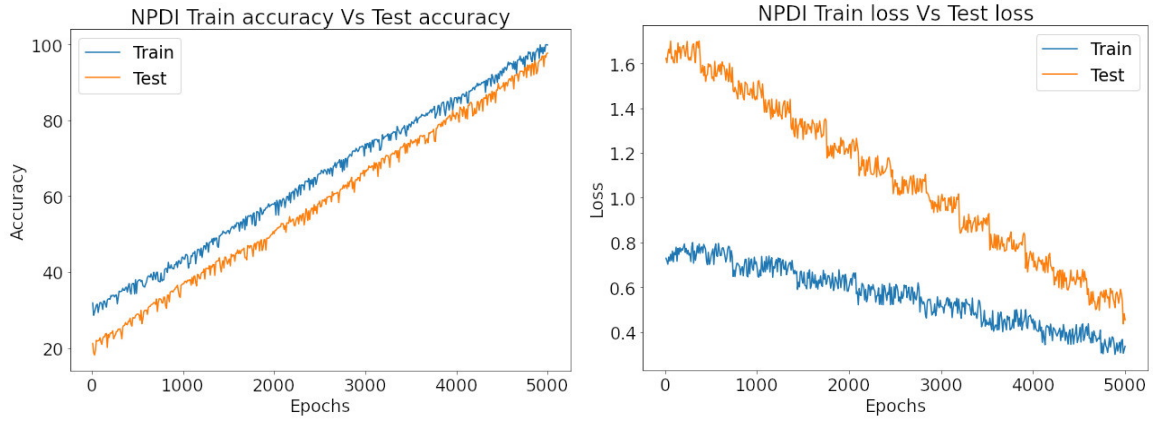
Table 3.4: Proposed model's testing performance accuracy, precision, recall, and F1 score.

Dataset	Accuracy	Precision	Recall	F1 score
CASIA	0.962	0.915	0.853	0.783
NPDI	0.977	0.88	0.73	0.68
Overall	0.969	0.897	0.791	0.731



(a) Training vs. testing accuracy. **(b)** Training vs. Testing loss.

Figure 3.4: CASIA dataset training vs. testing accuracy and loss.



(a) Training vs. testing accuracy. **(b)** Training vs. Testing loss.

Figure 3.5: NPDI dataset training vs. testing accuracy and loss.

The forgery detection and localization results of the proposed model are shown in Figure 3.6.

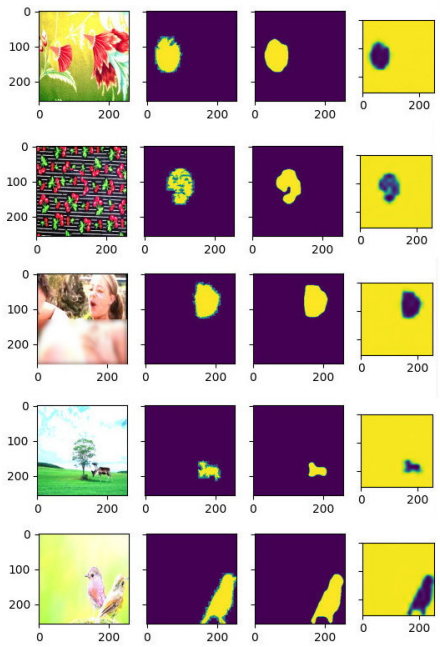


Figure 3.6: Forgery detection and localization: Column 1, manipulated images; Column 2, ground-truth masks; Column 3, binary mask generation; and Column 4, probability of heat map.

The images in Figure 3.6 are from the CASIA and NPDI datasets, where first two rows are copy-move forgery images with their corresponding ground-truth and binary mask generated on the manipulated region. Row 3 is the forged porn image from the NPDI dataset with its corresponding results. Row 4 and 5 are the image splicing forgery images with their corresponding results. The binary mask results of column 3 are predicted to be similar to the ground-truth mask in column 2. Column 4 depicts the heat map of the forged images.
Once the binary mask is generated, the forged object is extracted and is used for classification as copy-move or splicing forgery types. The results of the forged object are shown in Figure 3.7.

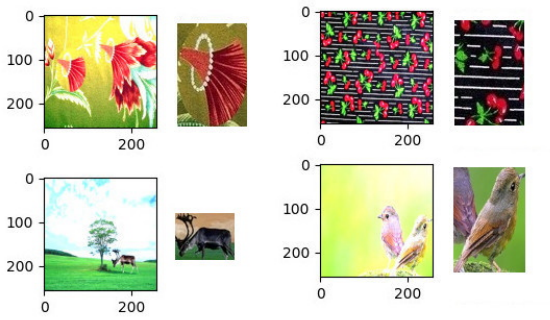


Figure 3.7: Forged object detection: Row 1, Copy-move forged object; and Row 2, image-splicing forged object.

Row 1 shows the forged object using copy-move and row 2 shows the forged object result using image-splicing. However, the forged object may be rescaled or rotated while performing the tampering operations on the images and a simple template matching technique may not detect such forged images. Therefore, template matching with an improved SIFT algorithm is applied for identifying such forged images. The classification of image forgery is based on the bounding boxes generated. The copy-move forgery will have two or more boxes if the same region is copied multiple times and pasted in the same image. The image splicing forgery will have only one bounding box where the forged object is copied and pasted from another image. The classification results are shown in Figure 3.8. Row 1 and 2 are the classification results of copy-move forgery, where the forged object is copied multiple times. Row 3 shows the image splicing forgery on porn images and rows 4 and 5 are the results of image splicing forgery. The template matching function with an improved SIFT algorithm locates the forged regions without affecting the process of training and improves the classification accuracy. The proposed model which was already trained on CASIAv1.0, CASIAv2.0 and NPDI datasets was tested on four benchmark datasets (Columbia[78], CoMoFoD[80], MICC-F600[82]and COVERAGE[81]) having forgery images of copy-move and image splicing, provided with their corresponding ground-truth masks. The Columbia dataset has image splicing forgery images with 183 authentic and 180 forged images. The CoMoFoD dataset includes copy-move forgery images with 5200 authentic and 5200 forged images. The MICC-F600 dataset has copy-move forgery images with 440 authentic and 160 forged images. The COVERAGE dataset includes copy-move forgery images with 100 authentic and 100 forged images.

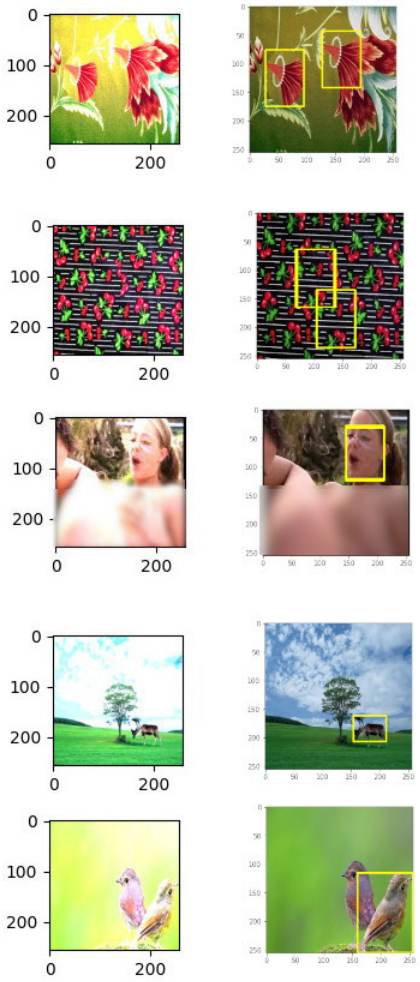


Figure 3.8: Forgery classification of copy-move and image splicing: Column 1, forged images; and Column 2, forgery detection.

Table 3.5: The proposed model's performance on benchmark datasets (pixel-wise accuracy and F1 score).

Dataset	Accuracy	F1 score
Columbia	0.968	0.78
CoMoFod	0.976	0.81
MICC-F600	0.96	0.79
COVERAGE	0.965	0.756

The proposed model learned the larger context of the tampering attacks on various datasets and showed better performance results on benchmark datasets. The results are shown in Table 3.5.

Comparison with existing models:

Our proposed model is compared with the existing methods ELA[83], NOI[84], CFA[2], J-LSTM[85], RGB-N[86], MR-CNN[87], and MantraNet[88] on the three benchmark datasets as shown in Table 3.6. We use accuracy and F1 score for comparison against existing forgery detection methods. From feature extraction

Table 3.6: Performance evaluation with existing methods (using Accuracy and F1 score) on 3 benchmark datasets. '-'denotes that the result is not available in the literature.

Method	Columbia		COVERAGE		CASIA	
	ACC	F1	ACC	F1	ACC	F1
ELA[83]	0.581	0.47	0.583	0.222	0.613	0.214
NOI[84]	0.546	0.574	0.587	0.26	0.612	0.263
CFA[2]	0.72	0.467	0.485	0.190	0.522	0.207
J-CNN-LSTM[85]	0.74	0.56	0.80	0.59	0.75	0.43
RGB-N[86]	0.858	0.697	0.817	0.437	0.795	0.408
MR-CNN[87]	0.978	-	0.936	-	-	-
MantraNet[88]	0.824	0.483	0.819	-	0.817	-
Proposed Model	**0.982**	**0.77**	**0.973**	**0.75**	**0.96**	**0.78**

through detection and localization of forged regions, our proposed model outperforms the existing models. On the Columbia[78], Coverage[81], and CASIA[79] datasets, the proposed model outperforms the ManTraNet[88] by over 10% in terms of accuracy. The model also has the advantage of effectively detecting and localizing in multi-scale, rotation invariant, low illumination with variation in resolution and noise. The proposed model outperformed other methods which are fine-tuned setups like J-LSTM[6] and RGB-N[51]. J-LSTM uses spatial and context information extracted from CNN and LSTM networks to detect forgeries in input images that have been divided into patches. However, It cannot correlate between different blocks of a patch. We achieved better results compared to traditional models like ELA[83], NOI[84], and CFA[2]. Based on the study of existing

approaches, our suggested model performs better than conventional methods by a wide margin. It is also comparable to deep neural network methods, achieving consistent performance across all the datasets, demonstrating that the model is robust and generalizable. We explored the effectiveness of the proposed model by conducting ablation studies using four standard benchmark datasets as shown in Table 3.7.

Ablation studies. In this study, we explored the existing LSTM-CNN mod-

Table 3.7: Comparison of the proposed model with LSTM-CNN variants.

Method	Columbia	COVERAGE	CASIA	NPDI
J-LSTM[85]	0.74	0.80	0.75	0.68
LSTM-CNN[12]	0.81	0.88	0.89	0.73
Proposed model	**0.98**	**0.97**	**0.96**	**0.97**

els and implemented them to analyze how well they perform on forgery images to compare with the performance of our proposed model. The proposed model performs better when compared to J-LSTM with a good margin in terms of accuracy by learning the large context of correlation between the image patches. The proposed model is capable to learn better image manipulations over LSTM-CNN by using an improved SIFT algorithm while localizing forged objects that are rotated and multi-scaled. The use of resampling features, LSTM, and convolution-deconvolution with an improved SIFT algorithm enhance the overall architecture's ability to learn image manipulations better.

3.4 Summary

Modern technology advancements have led to the use of digital images in all fields, including medicine, entertainment, forensic science, digital media, social media, and many more. There has been a tremendous increase in the production and exchange of images which has led to the fabrication of images that can misrepresent the information among the community. Humans tend to believe what they see compared to verbal communication. Hence, there is a need to authenticate the images to build trust among the community. Our objective in this Chapter was to categorize the forged images (copy-move or image-splicing) and locate the areas

that were forged. To detect and localize image forgeries (copy-move and image-splicing), a hybrid model LSTM-CNN is proposed that combines SIFT algorithm with ORB and RANSAC techniques. In Chapter 4, we focus on inter-frame video forgery detection and localization using deep learning techniques.

Chapter 4

Deep learning-based forgery identification and localization in videos

In Chapter 3, we focused on copy-move and image-splicing forgery detection and localization. The proposed model classified image forgery on benchmark datasets resulting in better performance compared to existing models. In this Chapter, we propose a model for inter-frame video forgery (frame insertion and deletion) detection and localization from cybercrime analysis perspective. A video forensic analysis consists of examining, comparing, and assessing video files that are used as evidence in court. In terms of applicability, active forensic schemes are limited in performance. Hence, passive forensic schemes were developed. Using passive or blind video forgery approaches, specific artifacts could be analyzed either statically or temporally. These approaches do not need prior information to analyze the forgery in videos, since they depend on the traces of forgery present in the video. Identification of video forgeries carried out using passive video forgery methods is a challenging task for the researcher. The traces left behind after the forgery can be used to distinguish between genuine and manipulated videos. Using passive approaches, we can detect any unauthorized manipulation, whether it's done within a frame (intra-frame level) or between frames (inter-frame level). Intra-frame tampering manipulates a frame at the object level or block level of a video. On the other hand, inter-frame tampering involves manipulating a set of frames (removal, insertion, replication, or shuffling) within a video. The investigation of inter-frame video forgeries, specifically the detection and localization

of frame insertion and deletion is the main emphasis of our work. In the literature, there are various techniques proposed for the detection of inter-frame video forgeries[1]. Most of the techniques are based on manual feature extraction to identify the artifacts for detecting and localizing inter-frame forgery operations. These features are sensitive in extracting post-processing operations (light inconsistencies, noise, compression, blurring, etc.). In addition, these techniques are unable to perform efficiently in detecting video forgeries as they are trained and tested on the same datasets.

Deep learning provides a combined service of extracting features and classification. DL-based methods have performed better in various application domains like action recognition[15], image classification[89], and object recognition[90]. The two most popular deep learning algorithms that perform well in pattern recognition of images are convolutional neural network (CNN) [65] and recurrent neural network (RNN)[66]. Many deep learning-based algorithms are proposed for forensic analysis in the image and video forgery to date[91]. CNN models were put forward by Nguyen et al.[92] for detecting inter-frame video forgeries by retraining the available CNN model trained on ImageNet dataset. The proposed model fails to extract temporal features accurately from forged videos but works well in extracting spatial features. From the analysis point, the authors created their own database known as the video inter-frame forgery dataset (VIFFD) involving inserting, duplicating, and replicating forgery videos. A deep learning-based CNN model and SSIM algorithm was proposed by Fadl et al.[93] for detecting video inter-frame forgeries. The forgery classification was carried out using RBF multi-class support vector machine (RBF-MSVM). But, the evaluation cost of the proposed model was very high and no cross-dataset evaluation was done. Bakas et al.[94] proposed a video forgery (inter-frame) detection method based on prediction footprint variation (PFV) and forgery localization was carried out using a range of motion vectors. The method fails to detect forgeries for videos with GoPs inserted or deleted in a video frame sequence. We propose, a 3D convolutional neural network (3DCNN) model to overcome the drawbacks of existing methods and detect inter-frame video forgeries. In addition, we use inherent temporal abnormalities to locate the amount of tampered video frames based on the temporal disparity between neighboring video frames. The multi-scale structural similarity (MS-SSIM) index measurement analysis algorithm is used to localize video inter-frame forgery (insertion and deletion) for which the original source is not available. The experimental outcome depicts that the suggested model performs better in video inter-frame forgery detection and localization on

test videos without limiting to post-processing operations, compression rates, and video length.

4.1 Challenges in video forgery (inter-frame) detection

Deep learning (DL) based 2DCNN models are most effective in extracting spatial features from images but fail to extract temporal features from videos and have high computational costs. One of the significant challenges in processing videos is to extract efficient features. The traditional methods used to evaluate inter-frame forged videos (high quality and lengthy) have demonstrated poor performance and low processing speed on large datasets is a challenging task. To address these issues, we use the deep learning-based 3DCNN model[15] to extract high-dimensional features from video files. Furthermore, 3DCNN has acquired high prominence in the development of a successful and reasonably accurate methodology for video classification on a wide range of large datasets. The temporal correlations across consecutive frames should be carefully examined to improve video forensics analysis.

4.2 Contributions

The main contributions of the work are:

1. Our approach involves designing a 3DCNN model based on Conv3D layers to learn high-dimensional features from video frames to detect inter-frame forgeries.

2. Unlike the normal 3DCNN model, we propose an absolute difference algorithm to evaluate the difference of successive frames that minimizes temporal redundancy among the frames and exposes the artifacts of forgery operation in videos.

3. Localization of inter-frame forgery (insertion and deletion) in the video is carried out by spatial-temporal analysis using a multi-scale structural similarity(MS-SSIM) index measurement algorithm.

4. Creation of inter-frame video forgery(insertion and deletion) dataset using UCF-101 action classes.

4.3 Methodology

Inter-frame video forgery detection carried out using hand-crafted feature extraction techniques and classifiers like k-nearest neighbor(k-NN)[95], support vector machine (SVM)[96], etc. Inter-frame video forgery detection carried out using hand-crafted feature extraction techniques and classifiers like k-nearest neighbour (k-NN)[95], support vector machine (SVM)[96], etc. have restrictions in terms of dataset processing, limited solution, and computational complexity. This implies that the performance of the model is affected when hand-crafted feature extraction techniques are used. To extract important features from the video frames, deep learning techniques are popularly used. DL-based 3DCNN technique is used for extracting prominent features from training datasets and solving various computer vision problems. Hence, we propose a 3DCNN model for detecting inter-frame video forgery. The 3DCNN model has the advantage of extracting spatio-temporal features more accurately, making it capable of detecting inter-frame video forgery. The model proposed is divided into inter-frame video forgery detection and inter-frame video forgery localization as shown in Fig.4.1.

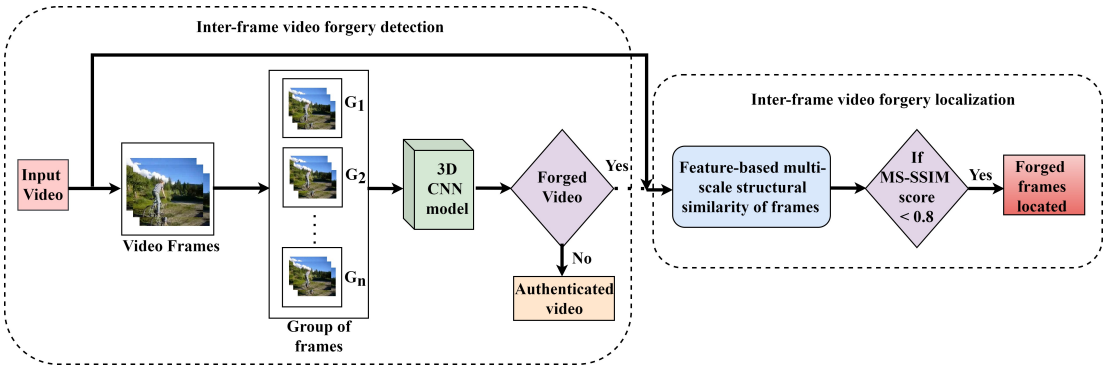


Figure 4.1: A model for detecting and localizing inter-frame video forgeries.

4.3.1 Inter-frame video forgery detection

In this section, we discuss the process of inter-frame video forgery detection by dividing it into video frame pre-processing and 3DCNN model.

4.3.1.1 Video frame pre-processing

A video is a composition of sequential image frames that appear to be identical to one another. To differentiate the video frames from each other, an absolute difference layer is added at the beginning of the 3D CNN model. The equation 5.1 computes the pixel-wise discrepancy between frame f and the adjacent frame $f+1$. Fig.4.2 shows a sequence of frames for sky-diving action class from UCF-101 database. The 3DCNN model is fed with the output deviation from the frame heap.

$$P_f(x,y) = |K_f(x,y) - K_{f+1}(x,y)| \tag{4.1}$$
$$1 \leq m \leq w, 1 \leq n \leq h$$

Where, $P_f(x,y)$ is the difference frame, $K_f(x,y)$ is the intensity of pixel (x,y) in the f^{th} frame, w and h are the width and height of the video frames. During

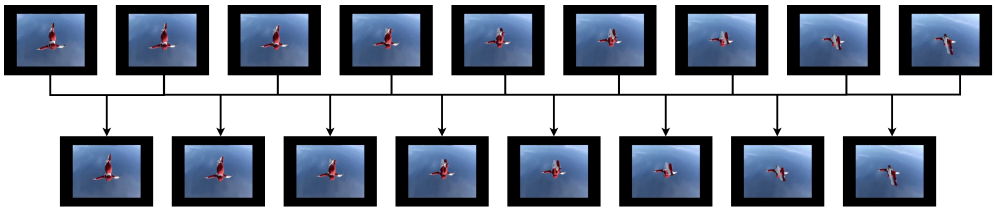


Figure 4.2: Absolute frame difference of skydiving action from UCF-101 dataset.

the process of inter-frame video tampering some discernible artifacts obligatorily exist and to detect such artifacts, we evaluate the absolute disparity among the sequential frames for minimizing the temporal redundancy. In addition to the absolute difference layer, the video frames are pre-processed to construct a group of frames to deal with large video files, improve accuracy, and reduce computations. Consider an input video V_1 consisting of F frames with dimensions $W \times H$ pixels. The frame sequence of video V_1 is grouped into group-of-frames G_f with length F frames. The stream of video in every frame group is represented by 3D frames. The mathematical equation 4.2 is stated below:

$$V_1 = \bigcup_{f=1}^{F} G_f \tag{4.2}$$

Where, G_f is the f^{th} group-of-frames, F is the overall frames of the input video.

4.3.1.2 3-Dimensional Convolutional Neural Network(3DCNN) model

Three Conv3D layers are used in the proposed model for inter-frame video forgery detection followed by ReLU, MaxPooling, Batch Normalization, dropout, dense, softmax, and output class as shown in Fig.4.3. Inter-frame forgery attacks are

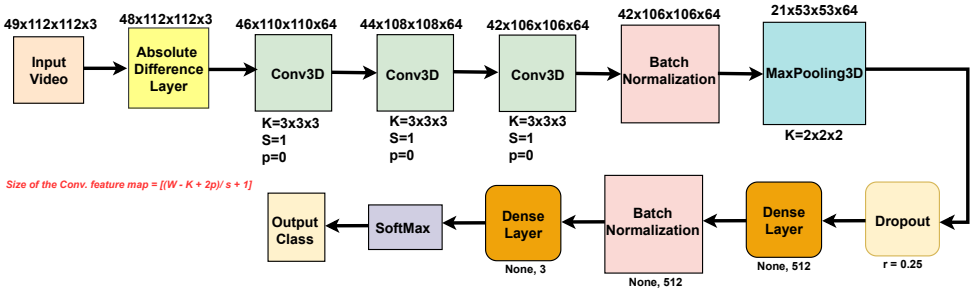


Figure 4.3: The Proposed 3DCNN for detecting video inter-frame forgery.

performed in the temporal domain, and the 3DCNN model applies convolutions in the 3D space that are optimal for extracting features from the temporal and spatial domains. A 3D convolution is a type of convolution operation in which, a 3D-filter is convolved in the three dimensions by adding multiple adjoining frames to generate a 3D cube. The following equation 5.3[16] corresponds to 3D convolution for the value at point (x, y, z) on the j^{th} feature map in the i^{th} layer.

$$v_{ij}^{xyz} = ReLU\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{t=0}^{T_i-1} W_{ijm}^{pqt} v_{(i-1)m}^{(x+p)(y+q)(z+t)}\right) \tag{4.3}$$

Where T_i is the 3D-filter kernel size along the temporal dimensions. W_{ijm}^{pqt} is the feature map connected to the m^{th} value of the kernel in the previous layer. The Conv3D kernel sizes are stated as $no.of\,kernels \times (height \times width \times inputs)$. The feature maps size among the layers is given as $no.of\,feature\,maps \times (height \times width)$. We apply padding in all the layers to maintain the shape of the frames. To aggregate the max/mean of a certain feature over a specific region, we use the MaxPooliing operation. MaxPooling downsamples the input along its spatial and temporal dimensions (depth, width, and height) by taking the maximum value over the input block for each channel of the input. The MaxPooling layer not only reduces the pixel density of the source images but also makes the network stable to changes in frame difference. To ensure the performance of the output

activations, the batch normalization [97] method is used that helps the 3DCNN model in handling overfitting issues by achieving a well-balanced generalization as part of the regularisation process. After several successive fully-connected layers, a softmax layer for multi-class classification is used, to produce the desired output. The output feature map of the convolution layer is evaluated as follows:

$$output\,feature\,map = [(W - k + 2p)/s] + 1 \tag{4.4}$$

where input shape is W, kernel size is k, padding is p and stride is s. With the aid of fully connected layers and a softmax layer, the feature vectors are transformed into 3D probabilities. Finally, the results classified are shown based on the 3D probabilities.

4.3.2 Inter-frame video forgery localization

We use multi-scale structural similarity to localize the inter-frame video forgery, which is discussed in the following sections.

4.3.2.1 Multi-scale structural similarity index measurement

Multi-scale structural similarity (MS-SSIM) index measurement is used to examine the dissimilarity of video frames and also to measure the duration of video inter-frame forgery (insertion and deletion). Multiple-scale modeling of visual intensity (luminance), contrast, and structure serves as the basis for the MS-SSIM[17]. The primary goal of MS-SSIM is to assess the quality of video frames and to achieve improved detection accuracy while processing in real-time. The MS-SSIM starts the process of locating inter-frame video forgery by using a referential frame and comparing with succeeding frames. The succeeding frame is then considered as a reference frame, and the procedure is repeated with the next succeeding frames. The MS-SSIM achieves better accuracy and faster processing speed than the normal Structural Similarity (SSIM) index measure.

4.3.2.2 Structural similarity index measurement

The structural similarity assessment is a measure for comparing two images' structural similarity. The method attempts to measure the visibility of defects (discrepancies) between an inconsistent and a reference image in order to evaluate perceptual image quality. The human vision system can easily and quickly identify

through structural information of the images by differentiating the information obtained from a similar reference image. The metric that simulates this behavior will outperform on tasks to detect inter-frame video forgeries (insertion and deletion) that require distinguishing between a sample frame and a reference frame. The Structural Similarity (SSIM)[18] extracts three main features from a video frame: luminance, contrast and structure. The SSIM is calculated as follows:

$$SSI_{x,y} = \frac{(2\mu_x\mu_y + K_1)(2\sigma_{xy} + K_2)}{(\mu_x^2 + \mu_y^2 + K_1)(\sigma_x^2 + \sigma_y^2 + K_2)} \tag{4.5}$$

where μ_x, μ_y, σ_x and σ_y are the mean and standard deviation of both the reference frame and sample frame. $K_1 \,\&\, K_2$ are constants.

Luminance

The luminance is obtained by the pixel values average. It is denoted as μ and is expressed as:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.6}$$

where x_i is the image's i^{th} pixel value of image x. N is the overall pixel values. The function $L(x, y)$ is comparing luminance of μ_x and μ_y.

Contrast

All the pixel values standard deviations are utilized to evaluate and extract contrast features denoted as σ. The contrast function is formulated by:

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{1/2} \tag{4.7}$$

The pixel values of the frame is given by mean μ. The contrast function $C(x, y)$ is a comparison function of σ_x and σ_y.

Structure

The input signal is divided by its standard deviation, resulting in a unit standard deviation and allowing a more robust comparison.

$$(x - \mu_x)/\sigma_x \tag{4.8}$$

where x is the input image.

The following comparison function compares the two given video frames on the aforementioned variables. Finally, a combined function is defined that aims to combine and generate the similarity index value. The luminance comparison

function $L(x,y)$ is given by μ representing the mean of a given frame formulated as:

$$L(x,y) = \frac{2\mu_x\mu_y + K_1}{\mu_x^2\mu_y^2 + K_1} \tag{4.9}$$

The two video frames being compared are x and y. K_1 is a constant that ensures stableness when the divisor approaches zero. $C(x,y)$ defines the contrast comparison function, which is written as:

$$C(x,y) = \frac{2\sigma_x\sigma_y + K_2}{\sigma_x^2 + \sigma_y^2 + K_2} \tag{4.10}$$

σ denotes the standard deviation of given video frame.
The function $S(x,y)$ defines the structural comparison, which is formulated as:

$$S(x,y) = \frac{\sigma_{xy} + K_3}{\sigma_x\sigma_y + K_3} \tag{4.11}$$

where $\sigma(xy)$ is defined as,

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i-1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

The final score is evaluated by:

$$SSIM(x,y) = [L(x,y)]^{\alpha}.[C(x,y)]^{\beta}.[S(x,y)]^{\gamma} \tag{4.12}$$

where α, β, γ define the weight given to each model of luminance $L(x,y)$, contrast $C(x,y)$, and structure $S(x,y)$.

4.3.2.3 MS-SSIM algorithm

1. Consider only one frame at a time f_t.
2. At time t+1, acquire the next frame f_{t+1}.
3. Calculate the multi-scale structural similarity of two consecutive frames f_t and f_{t+1}.
4. A frame forgery is identified with reference to insertion or deletion if the similarity value is below the pre-determined threshold $\theta = 0.8$ between two consecutive frames. As per the algorithm, a change in the threshold value beyond 0.8 will not drastically affect the accuracy rate of localizing frame insertion or frame deletion, but a threshold value below 0.8 will affect the accuracy rate of localization.

4.4 Result and analysis

4.4.1 Datasets

The proposed model is evaluated on UCF-101[98] and VIFFD[99] datasets. There are currently no standard datasets dedicated to video inter-frame forgeries[93]. For our experiments, we selected 700 videos of different actions from the UCF-101 dataset and generated inter-frame forgery (insertion and deletion) using ffmpeg tool[100].
In the creation of frame insertion and deletion forgery videos, we varied the number of frames from 10 to 150 frames. The selected videos are compressed with H.264 and MPEG-4 using libraries libx264 and libavcodec of the ffmpeg tool with an adaptive group-of-pictures (AGoP) structure. The videos have frame rates ranging from 25 to 30 frames per second. We chose 60 videos and applied post-processing operations such as Gaussian blurring, Gaussian noise, and brightness variations to analyze inter-frame video forgery detection under different conditions.
The VIFFD dataset has 392 videos with a combination of authentic, insertion, deletion, and duplication videos collected from five surveillance cameras of real-life scene shots with different lighting conditions. For the experimental analysis, we chose 30 authentic videos, 30 insertion forgery videos, and 30 deletion forgery videos. The inter-frame video forgery is differentiated into multi-class classifications namely insertion, deletion, and authentic with a total of 2190 videos. The details of the datasets are shown in Table 4.1. Fig.4.4 shows the sample video from

Table 4.1: Multi-class dataset details

Dataset	Authentic	Insertion	Deletion	Total
UCF-101	700 videos	700 videos	700 videos	2100
VIFFD	30 videos	30 videos	30 videos	90
Total				2190

UCF-101 dataset showing authentic and forged (insertion and deletion) video.

4.4.2 Details of Implementation

In our implementation, we used an 8GB GPU (NVIDIA RTX 2080) with 32GB RAM and a 2.20GHz Intel i7 processor. In order to develop a high-level deep

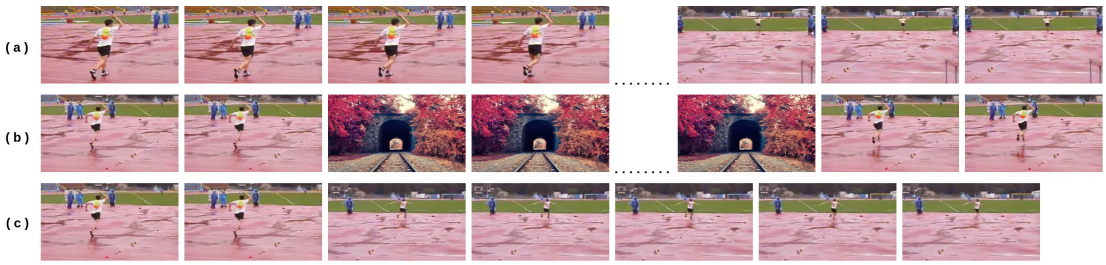


Figure 4.4: UCF-101 sample video showing a. authentic, b. insertion forgery, and c. deletion forgery.

learning model, we used Python 3.0, OpenCV, FFMPEG, and Keras 2.6.0 with the Tensorflow 2.6.0 framework.

Implementation using a GUI

A GUI-based sample application that can assist video forensic investigators in analyzing video files is shown in Fig.4.5. A user interface for the application

Figure 4.5: GUI-based sample application showing video forensic analysis.

consists of three sections (left, right, and bottom). All video forgery categories (Authentic, Forgeries) are presented in the left section in a tree view. With the forgery detection, the video files will be placed in the appropriate category. When we select a category in the left section, the right section will display all of the videos in that category, along with meta-data. The bottom section is used to play the video for analyzing the content of a video forgery listed under the right section.

4.4.3 Experimental analysis

The evaluation criteria and experimental analysis of our model are discussed in this section. To evaluate the proposed model's performance, we consider detection accuracy (Acc), which is defined as:

$$Accuracy(Acc) = \frac{TP+TN}{TP+FP+FN+TN}$$

where the true positive rate (TP) is the percentage of video forgeries that are accurately identified.
The False positive rate (FP) is the percentage of video forgeries that are incorrectly identified.
The true negative rate (TN) is the percentage of video forgeries that are accurately identified as original videos.
The False negative rate (FN) is the percentage of video forgeries that are incorrectly identified as original videos.
The model's effectiveness is satisfactory when the evaluation parameters precision rate, recall rate, and F1 score are achieved better.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

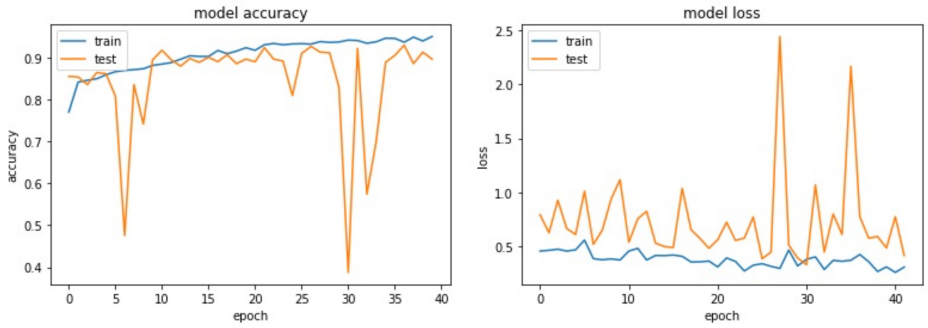
$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The video clips are denoted with a size of $n \times f_m \times w_d \times h_t$, where n specifies the number of filters used to capture the video, f_m is the group of frame count in the video shot, w_d and h_t are the width and height of each frame in pixels. The kernel size is expressed as $dp \times k \times k$, where dp specifies the kernel's temporal depth and k denotes the spatial size of the pooling and 3D convolution layers.
The proposed 3DCNN model takes video shots as inputs and divides them without overlapping into 49-frame segments that are sent into the network's absolute difference layer. The input dimensions are $3 \times 49 \times 112 \times 112$ and the output of the difference layer $3 \times 48 \times 112 \times 112$ is given to the first convolution layer. The three convolution layers use 64 filters. The kernel size used is $3 \times 3 \times 3$ with stride 1. We minimize the output size by a factor of 8 when compared to the input size by using the MaxPooling3D kernel with $2 \times 2 \times 2$ size. The network starts the learning process with a learning rate of 0.01 and a weight decay of 0.005.

Adam (adaptive moment estimation) is used as the loss function to optimize the 3DCNN-based model. We chose a batch size of 16 for training. After the 40 epochs, we obtain the trained weights and parameters .
We use VIFFD as a testing dataset to validate the trained model's performance. The pre-processing for the testing dataset is the same as the training dataset. The input to the trained 3DCNN model is the test dataset and the class labels for each video clip are obtained. If any one of the classes is predicted with insertion or deletion forgery then the frame is marked as a tampered frame. The video frame is authentic when the class is predicted with authentic.

Training and testing results with UCF-101 dataset

The proposed 3DCNN model is evaluated on the UCF-101 dataset which is split randomly in a ratio of 75:25 for training and testing, with a batch-size of 16, and the number of epochs carried out is 40. The categorical cross-entropy is utilized as a loss function to assess the efficiency of a multi-class classification model with probability values as its output. The resulting evaluation metric used is accuracy, precision, recall, and F1 score. Fig.4.6a and Fig.4.6b show accuracy and loss results after training and testing the proposed model on the UCF-101 dataset. The graph depicts the variance in accuracy and loss. After 40 epochs, the loss begins to decrease as the learning improves. As the proposed model's training progresses, its precision reaches the highest level.



(a) Training vs. testing accuracy. **(b)** Training vs testing loss.

Figure 4.6: UCF-101 dataset training vs. testing accuracy and loss.

Furthermore, the suggested model's prediction accuracy in detecting insertion and deletion forgery reveals that it is suitable for practical and realistic videos. Along with a large number of frame forgeries (insertion and deletion), the proposed method even detects a small number of frame forgeries. Also, the model

is tested on video forgeries in a static scene and dynamic scenes with multiple forgeries within the same video. This leads to the robustness of the proposed method. Once the model was trained, we tested and predicted the model with the VIFFD dataset. The predicted evaluation matrix gives a recovery implication about the errors that our trained model is making during testing and examines the performance of the classification process. The predicted rate of evaluation matrix in Table 4.2 shows a predictive test on the VIFFD dataset. The above

Table 4.2: Performance evaluation showing predictive test on VIFFD dataset. Support - This attribute indicates the total number of groups.

	Precission	Recall	f1-score	support
FrameDeletion	1.0	0.92	0.96	75
FrameInsertion	0.95	1.00	0.98	119
Authentic	1.00	1.00	1.00	124
Accuracy			0.98	318
Macro avg.	0.98	0.97	0.98	318
Weighted avg.	0.98	0.98	0.98	318

results show that the proposed model accurately distinguishes authentic videos from inter-frame forgeries.

Localizing insertion forgery in videos. During localization of insertion forgery, every video frame is matched with the adjacent frames both temporally and spatially to generate the MS-SSIM score. If the MS-SSIM score of the frames in the inserted position is substantially smaller than the actual value (0.8), there will be two discontinuities in insertion forgery with a dropping peak value where the functionalities of consecutive frames differ significantly. In Fig.4.7, the visual output shows frame insertion video forgery where the frames 100 to 150 are inserted from other videos which are indicated by two falling peak values. The values of similarity in the locations where the frames are inserted are significantly lower than the normal values. Inter-frame insertion forgeries will show two distinct declining peak values, indicating that they come from different videos.

Localizing deletion forgery in videos. A sequence of video frames are removed from the authentic videos in order to hide the truth. Generally, two

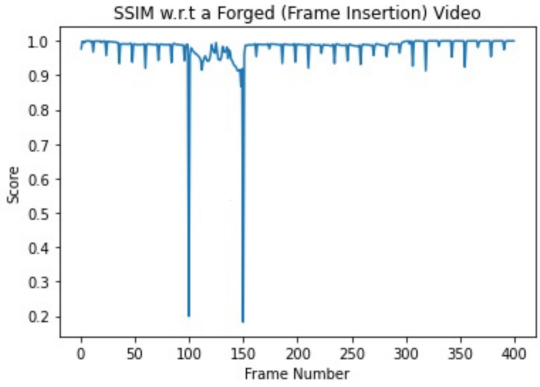


Figure 4.7: Frame insertion localization based on MS-SSIM.

adjacent video frames have a similarity of a high degree in authentic video and this similarity may immediately change with lower scores at the number of frames deleted location resulting in frame deletion forgery. The MS-SSIM score results in a lower value in frame deletion video forgery resulting in a single falling peak value. The frame deletion operation results in a greater interval between two sim-

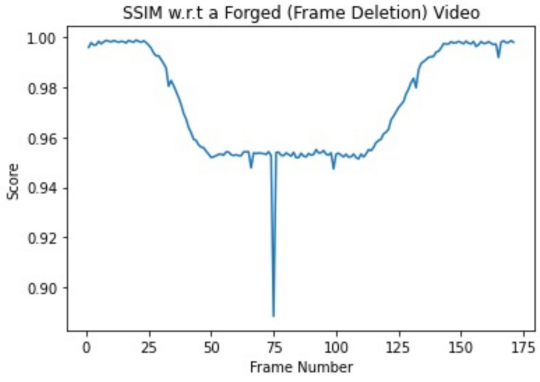


Figure 4.8: Frame deletion localization based on MS-SSIM.

ilar frames, resulting in a dropping peak value. In Fig.4.8, the inter-frame video deletion forgery results in a single falling peak value where 50 frames are deleted (75 to 125 frames). Based on the similarity curve, we know that the suspicious video was forged by frame deletion.

4.4.4 Existing inter-frame video forgery methods

The existing methods that are available in the literature for the detection of inter-frame(insertion and deletion) video forgery are shown in Table4.3 with their performances. All these listed methods including their datasets are not publicly

Table 4.3: Existing inter-frame video forgery methods. '-' denotes that the result is not available in the literature.

Method	Forgery operation	No.of Videos	Accuracy	Precision	Recall	F1-score
QoCCLB[101]	Insertion, Deletion	599	-	0.95	0.92	0.93
PR&MB[102]	Deletion	770	-	0.72	0.66	0.699
ZOCMCFA[103]	Insertion, Deletion	60	-	0.83	0.85	-
CNN[104]	Insertion, Deletion	128	0.85	-	-	-

available. To check the performance of their methods, the researchers have created their own datasets which have inter-frame video forgeries (insertion and deletion).

4.4.5 Comparison

We tested our proposed model with the dataset created for the MO-BWO[5] method achieving better accuracy, precision, recall, and F1-score as shown in Table 4.4. The proposed model 3DCNN-MS-SSIM performed better on unseen

Table 4.4: Comparison of the proposed model with MO-BWO[5]. '-' denotes that the result is not available in the literature.

Method	Forgery operation	No.of Videos	Accuracy	Precision	Recall	F1-score
MO-BWO[5]	Insertion, Deletion	18	0.85	0.825	-	0.826
Proposed Model	**Insertion, Deletion**	**18**	**0.952**	**0.923**	**0.90**	**0.945**

data, resulting in a generalized model.

Ablation study: We chose 60 videos and applied post-processing operations like compression rates, varying video length, Gaussian noise, Gaussian blurring, and brightness variations. These videos were trained and tested on our proposed model with and without applying difference algorithm. The results achieved were compared with existing methods[103] and [54] under insertion and deletion forgery. The accuracy of these methods is 0.7478 and 0.7724 respectively. We achieved an accuracy of 0.8234 without the difference algorithm and 0.9817 with the difference algorithm. A comparison between the proposed approach and existing methods is presented in Table 4.5 under insertion and deletion forgery.

Table 4.5: Comparison of the proposed method with existing methods under insertion and deletion forgery.

Method	**Post-processing operations**	**Difference Algorithm**	**Accuracy**
Liu[103]	Compression, Gaussian noise, Gaussian blur, brightness.	No	0.7478
Zhao[54]	Compression, Gaussian noise, Gaussian blur, brightness.	No	0.7724
Proposed Method	Compression, Gaussian noise, Gaussian blur, brightness.	No	0.8234
Proposed Method	**Compression, Gaussian noise, Gaussian blur, brightness.**	**Yes**	**0.9817**

4.5 Summary

Digital video forensics plays a crucial role in analyzing cybercrime since video editing software makes it easier to tamper videos for illegitimate gain and produce tampered evidence in court. Existing methods for inter-frame video forgery detection do not fully assist forensic investigators in detecting video forgery when no source video is available and only the forged video is provided. In this Chapter, we proposed a 3-dimensional CNN model for detecting inter-frame forgery in video and used MS-SSIM approach to localize the forgeries. The proposed

model learns more relevant characteristics to detect video inter-frame forgeries with high classification accuracy and outperforms the existing models in both static and dynamic videos. In Chapter 5, we explore the recognition of unethical human actions from a video forensic perspective using a deep learning-based hybrid model.

Chapter 5

Unethical human action recognition using deep learning based hybrid model for video forensics

In Chapter 4, we proposed a 3DCNN model for inter-frame video forgery (frame insertion and deletion) detection and applied the MSSIM approach to localize the forgeries. The proposed model results were better compared to existing inter-frame video forgery detection techniques. In this Chapter, we automate the modeling of human action to recognize unethical human activities in videos saving substantial time in multimedia forensics investigation. Due to the widespread adoption of mobile devices, lower storage costs, and faster transfer speeds, multimedia users are generating massive amounts of data, which has surpassed the forensic specialist's abilities to successfully examine and analyze multimedia content.

Unethical human action recognition methods are required for video forensic investigation in order to prevent cybercrime and devious actions from occurring in the first place. Recognizing unethical behavior is a key subject in video forensics which may curtail the analysis time of the digital evidence in video files. Videos are normally considered 3-dimensional signals to represent human actions and 3-Dimensional Convolutional Neural Network(3DCNN) is significantly used to extract spatio-temporal features for identifying human actions. 3DCNN is an extension to 2-Dimensional Convolutional Neural Network(2DCNN) that

learns features from both spatial as well as temporal signals present in video files whereas 2DCNN learns features from the spatial signal of still images. Deep neural networks, such as Convolutional Neural Networks (CNN), Deep Belief Networks, and Deep auto-encoder have shown their ability to extract complex statistical dependencies from high dimensional sensory inputs and efficiently learn hierarchical representations to generalize well across a wide variety of computer vision(CV) tasks like image classification, speech recognition, object detection, human action recognition, etc.

Human action recognition plays a key role in applications such as automated surveillance, elderly behavior monitoring, human-computer interaction, ambient assisted living, intelligent driving, content-based video retrieval, pornography action recognition, etc. Many algorithms for Human Action Recognition(HAR) have recently been proposed using deep learning techniques but, they are unable to effectively learn complicated features and are computationally heavy, resulting in incorrect action recognition. Action recognition is a method of assigning action labels to video frames, such as hopping, dancing, fighting, running, walking, and so on. This allows the system to identify various human actions efficiently and automatically[105]. In recent years, a variety of action recognition methodologies have been deployed, including wireless sensor network-based, wearable sensor-based, and video-based approaches. Video-based approaches have become increasingly popular due to their high action recognition rate and ease of use.

In video forensic analysis, HAR can be used to identify unethical human actions in video files. A politician in Karnataka (an Indian state) was recently involved in a controversy after a social activist approached police with false evidence allegedly depicting the minister in a compromising posture with a woman. The video was released to the media showing the woman in the video clips was enticed by the minister with a government job offer. The social activist requested a detailed investigation into the minister's alleged sex scandal[106]. As a consequence, there is a risk of propagating incorrect information and conspiracies through the Internet, leading to major disinformation in the community.

To investigate such cases, unethical human action recognition in video files would expedite the analysis. Due to multiple factors like having many inter and intra-class variances, changes in the background, lightning, angle variability, ambient noise, and the speed of activities, correctly identifying human action in videos remains a difficult challenge[105]. To address these issues, hand-crafted feature extraction methods such as the histogram of oriented gradients (HOG)

and the histogram of optical flow (HOF) were used but they only covered a small portion of the human action recognition and perform poorly on large, complex datasets. The classic approach of examining, segmenting, and classifying human activity begins with the extraction of human silhouettes from chaotic and shadow ambient regions, allowing action recognition[107]. But, occlusion and varied viewpoints affect Silhouette features resulting in poor performance.

The main objective of this research is to perform an analysis of video files (including Pornography) and improve learning ability for accurate recognition of unethical human actions in complex videos. In this direction, we first focused on the analysis of human action by extracting deep action features from large video action datasets using two-stream inflated 3D ConvNet(I3D)[3] and then improving the learning capability of the model by learning small discriminative features among spatial and temporal regions using Spatio-Temporal Attention(STA)[4] module. We proposed a fusion of the STA+I3D model by adding STA on top of I3D to achieve good results in recognizing complex unethical human actions. Then, we created a multi-action dataset with normal, unethical, and porn actions using benchmark datasets like Weizmann[108], HMDB51[109], UCF-101[110], NPDI[111], and UCF-Crime[112]. From the video forensics perspective, the proposed model is unique and is the first experimental demonstration of the fusion of STA+I3D for intelligent unethical action recognition in complex video files.

3DCNN can recognize complex human actions in videos using spatio-temporal features and does not rely on manual identifications. In addition, 3DCNN automatically adapts and learns hierarchical features from lower to higher patterns. Video frames are input to the 3DCNN model, which generates numerous channels of information, and a final output score for each human action is generated based on the feature maps representation [87]. A. Karpathy et al.[113] proposed a multi-resolution architecture to speed up model training and described the performance of neural networks in large-scale video classification. The authors explored the ways to use temporal data with single-stream 2DCNN and tested on different fusion architectures including single frame, late fusion, early fusion, and slow fusion using benchmark video datasets.

Karen Simonyan et al.[114] proposed two-stream Convolutional Networks(ConvNets) for action recognition in the videos by processing spatial and temporal networks separately. The class score generated by the two streams is combined by late fusion. The major drawback of this network is, training the two networks separately induces more training time and cost.

A very deep 3DCNN model was proposed by Du Tran et al.[115] for training large-scale video action datasets using spatio-temporal feature learning and demonstrated that 2DCNN based models are not suitable to learn 3-dimensional data. The proposed model learns the appearances spatially in the initial frames and learns action information in later frames of the video clips. The model has shown better performance on larger datasets.
It has been discovered in the literature that a pre-trained CNN model on large-scale annotated datasets can be transferred directly to an action recognition task with a minimal training dataset[79]. The pre-trained model improves the classification accuracy and performs well on small datasets but needs improvement when used on complex datasets such as UCF-50, UCF-101, IXMAS, HMDB51, and so on.
The most important capability of the human action recognition architecture is the extraction of massive features from spatio-temporal regions that can map with final output channels and make accurate predictions over videos. Joao Carreira et al. presented a novel two-stream Inflated 3D ConvNet(I3D)[3] using 2D ConvNet inflation. Convolution filters and pooling kernels of very deep image classification ConvNets are expanded into 3D in order to learn more spatio-temporal features. 2D filters are inflated across the temporal dimensions to create a 3-dimensional kernel filter from 2D images(NxNxN from NxN 2D). The majority of existing 3DCNN models treat all input video frames equally, ignoring spatial and temporal differences between video frames. Several studies have shown that 3DCNN learning functionality can be improved by using discrete modules since many of the existing works have ignored improving the 3DCNN learning capability during action recognition in videos due to lack of learning discriminative features in spatial and temporal regions. This motivates us to develop a hybrid model (STA+I3D) to increase the learning capability as well as improve the performance of unethical human action recognition in complex/large action datasets without affecting the number of parameters.

5.1 Challenges in human action recognition

There are several existing architectures like two-stream with optical-flow, LSTM-CNNs, 3DCNN, etc. used for human action recognition and classifications. These

models require high training, storage cost, and layer-by-layer stacking of 3D convolutions, which is critical for high-level action recognition tasks. The learning capability of the model is also affected since 3DCNN model ignores the spatial and temporal differences across the video frames.
In the proposed model, unethical human action recognition is performed by 3DCNN with the fusion of STA and I3D models' weights through transfer learning, fine-tuning, and optimization techniques.

5.2 Contributions

The major contributions of our work are:

1. For unethical human action recognition function, we proposed an efficient 3DCNN model for video forensics analysis.

2. For deep spatio-temporal feature extraction, we integrate a pre-trained I3D architecture with a discriminative feature extractor like STA, which outperforms several existing architectures in terms of accuracy and works on fewer computational resources.

3. Our proposed method is evaluated using various challenging benchmark datasets to showcase the novelty of the work. We achieved superior results than existing human action recognition methods by using pre-trained 2DCNN weights to initialize the model parameters and adding convolutional and deconvolutional operations for discriminative feature learning throughout the temporal dimension.

5.3 Methodology

The overall framework of the proposed method is presented in this section based on the fusion of STA and I3D as shown in Figure 5.1. There are five sections in this framework: 1. Data Preprocessing and Augmentation, 2. 3D ConvNets for learning spatio-temporal features, 3. Two-stream inflated ConvNet (I3D), 4. Spatio-Temporal Attention (STA), and 5. STA + I3D.

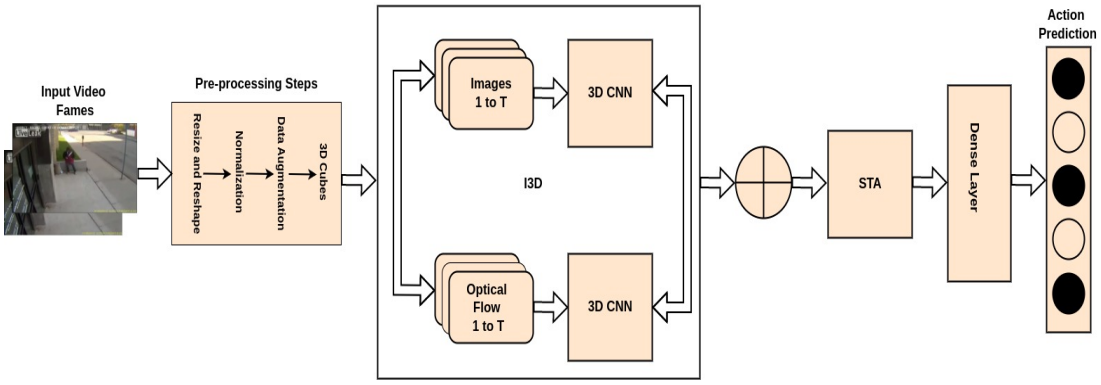


Figure 5.1: Proposed hybrid model of human action recognition using I3D and STA.

5.3.1 Data Preprocessing and Augmentation

To meet the requirements of deep learning algorithms, the raw input data needs to be transformed into structured data. Data preprocessing includes cleaning the data and making it suitable for training which includes frame extraction, resizing, reshaping, and data augmentation of the video files. A video is a collection of frames that are displayed in a specific order and move in real-time. The time dimension produces a series of images that appear to be moving when they are put together as frames. To work with deep learning video applications, it is difficult to process multiple frames which leads to the complexity of the system. So, we need to extract the individual frames of the video with a fixed number of frames and store them in a list-like data structure for further processing. The input to the model is a combination of fixed-size successive frames.

All the video frames are resized and reshaped to (58,224x224x3) frames denoting 58 frames, 224 height, 224 width,s and 3 channel RGB depth. In addition, all the frames' pixel values always have to be scaled between 0 to 1 to ensure that they are on the same scale and when dealing with small datasets, every frame of the class is normalized and data augmented. Data augmentation is applied when small datasets are used for deep learning tasks to prevent overfitting and increase the variance of data. Data augmentation is a method of increasing the variance of the dataset and artificially expanding the size of the training set by creating modified data from the existing data.

Using the 3DCNN architecture, we observed that the training model was faster, but was unable to generalize well when evaluating over-testing data. To solve

overfitting problems, there are various techniques available in the literature such as regularization, data augmentation, and adding dropout layers in the network. In the proposed work, we apply data augmentation on a video dataset by first converting each video into a fixed sequence of frames then each frame is augmented comprising of the transforms involving: RandomCrop, HorizontalClip, VerticalClip, RandomFlip, GaussianBlur, and RandomRotate.

5.3.2 3D ConvNets for learning spatio-temporal features

Convolutional Neural Networks(CNN) have made significant contributions in the area of computer vision (image and video processing). CNN can learn features from multiple layers hierarchically and generate a high-level representation of the raw inputs automatically. The performance of CNN is improved by empirical testing on large-scale video datasets, in which the networks must access not only the appearance information provided in single static images, but also needs to access complex temporal variations[113]. The 3D Convolutional Neural Network is a popular convolutional neural network for human action recognition in videos since a 3DCNN can convolve two-dimensional images and time dimensions.
A 2DCNN is the best for extracting spatial features from images, but may not be used for processing video data that are in continuous frames. In video sequences, human action recognition is a 3D signal composed of adjacent time dimension information that varies over time. 2DCNN computes features from the spatial dimensions by applying convolutions on 2D feature maps and hence is suitable for extracting features from images. However, in extracting features from video files it is necessary to capture the motion information from consecutive frames. To extract the features from both spatial and temporal dimensions 3D convolutions are applied in the convolution stages. A 3D Convolution is a kind of convolution in which the kernel slides in three dimensions rather than in 2D convolutions. The 3D convolution equation is expressed as follows[116]:

$$v_{ij}^{xyz} = ReLU\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{t=0}^{T_i-1} W_{ijm}^{pqt} v_{(i-1)m}^{(x+p)(y+q)(z+t)}\right) \tag{5.1}$$

where v_{ij}^{xyz} represents the convolution result at the position i of the j feature map*(x,y,z)* of the layer; $ReLU()$ is the activation function; b_{ij} is the deviation of the feature map; m is the index of the feature map in the layer *(i-1)*; W_{ijm}^{pqt} is the value at the position (p, q, t) of the feature map; t is the time dimension that

is unique to 3D convolution; P_i, Q_i, T_i are the width, height and depth of the convolution kernel respectively. Figure 5.2 shows the proposed 3D CNN model which is composed of four Conv3D layers with a depth size of 32, followed by Leaky ReLu layers.

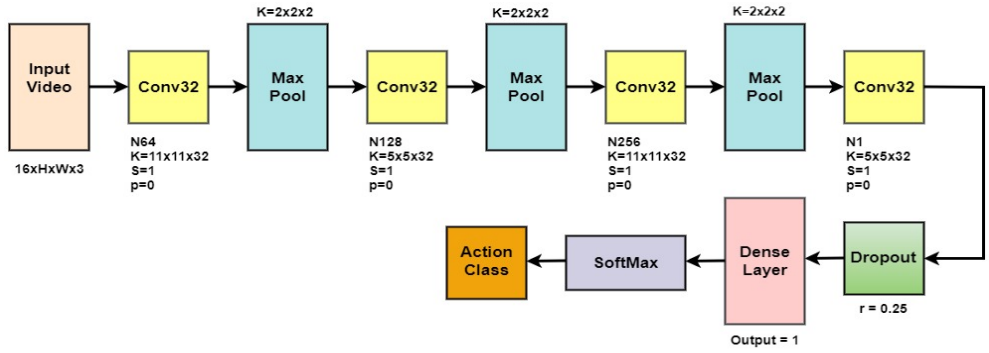


Figure 5.2: The architecture of 3D CNN model for unethical human action recognition

3DCNN is composed of many pairs of layers including 3D convolution kernels, MaxPooling, Dropouts, Dense and Output layers. To obtain the desired output, a softmax layer for multiclass classification is used after the various successive fully connected layers. The output shape of the layers after convolution is represented as:

$$output \quad shape = [(N - k + 2p)/s] + 1 \tag{5.2}$$

where input shape is N, kernel size is k, padding is p and stride is s. The equation 5.2[117] is used during the convolution operations. The first and third convolution uses a kernel size of 11x11, second and fourth convolution uses a kernel size of 5x5. All the convolutions use a depth size of 32 which denotes frame depth for video inputs. To perform downsampling progressively across the layers during training MaxPool of kernel size 2x2 with depth 2 is used. This allows to reduce the representation size and speeds up the computations. To prevent overfitting during training process, Dropout layers are used with a rate of 0.25. In order to perform classification on the features extracted by the Conv3D layers and down-sampled by the MaxPool layers, one Dense (or fully connected) layer is used. The Dense layer has output size fixed to 1, which represents the number of classes to recognize. The output of the dense layer is passed to the final activation function SoftMax, holding the output classes and predicting output from an array.

Consider a task with K classes, the SoftMax function[118] is expressed as:

$$Softmax(y)_i = \frac{exp(y_i)}{\sum_{j=1}^{n} exp(y_j)} \quad for\, j = 1, 2, ..n \tag{5.3}$$

where, y is input vector to fucntion softmax, y_i is the i^{th} element of input vector, $exp(y_i)$ is standard exponential function applied on y_i. SoftMax gives the predicted probability that class i will be selected by the model. In a model with a softmax activation function, the class with the highest probability is selected as the final prediction. 3DCNN architecture can be very useful in recognizing human actions without the need of feature engineering. However, it needs to enhance the training parameters by exploring transfer learning and data augmentation techniques. The following issues were found during the implementation of 3DCNN.

- 3DCNN requires a large number of labeled examples.
- In comparison to 2DCNN, 3DCNN has a lot more parameters resulting in overfitting on small datasets.
- The computational time is more, as it needs to train the parameters from scratch.
- There is a need for more hyper-parameter tuning, which consumes more time.
- The 3DCNN architectures are unable to build deeper layers due to the complex structure of the 3D convolution kernel.

To overcome the aforementioned issues, we determine the appropriate initialization parameters and then fine-tune the model using a pre-trained model on the available labeled dataset to obtain transfer knowledge from a 2DCNN to a 3DCNN.

5.3.3 Two-stream inflated Convnet (I3D)

Carreira, J et al.[3] introduced Two-stream inflated Convnet (I3D) based on 2D ConvNet inflation. The I3D model inflates the filters and pooling kernels (optionally their parameters) with state-of-the-art image classification architectures,

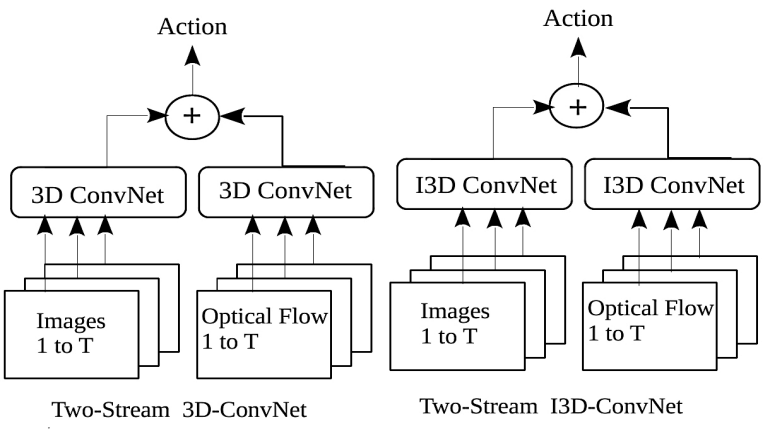


Figure 5.3: Two-Stream Inflated 3D ConvNets[3].

resulting in highly deep spatio-temporal classifier.
The architecture of I3D is depicted in the Figure 5.3, which shows the way 3D ConvNets benefit from ImageNet 2D ConvNet. I3D employs parameters from the ImageNet 2D pre-trained weights, which can be used to obtain a suitable bootstrap initialization value and get the time dimension 3D filter by repeating 't' times. In the 3D implementation, the parameters are employed from a 2D model and trained on a large dataset, such as ImageNet. The 2D inflated procedure not only constructs the structure of 3D ConvNets but also pre-train the 3D filters. Hence, I3D is an efficient step towards creating powerful 3D ConvNets using two-dimensional pre-trained weights. I3D has been built on the basis of a 3D Convolution Network (C3D)[115], which is briskly and more precise at detecting and classifying unethical behavior. I3D model uses Inception-v1[72] architecture, which offers better performance than existing methods after pre-training on Kinetics-400 datasets[119]. Inception expands "Wide" rather than "In Depth". In addition, it aggregates outcomes by combining spatial and temporal information. If p_m^l is the pre-trained 2D filter of the m^{th} channel in the l^{th} layer, then P_m^l denotes the corresponding filter in 3D ConvNet[120]. The 2D-Inflated operation can be described as:

$$P_m^l = \left[p_m^l, p_m^l, p_m^l\right] \tag{5.4}$$

$$P^l = C_l\left(P_0^l, P_1^l, ..., P_{m-1}^l\right) \tag{5.5}$$

where P^l denotes 3D kernel of the l^{th} layer, and the operation of merging all filters into one kernel is denoted by C_l. Each 3D filter is basically comprised of three 2D

filters. In Imagenet's pre-trained 2D ConvNet, the 2D filters are replicated from the same channel of the appropriate layer. Then, a 3D kernel is created from each l^{th} cubic filter. The pooling kernel can be simply converted from square to cubic. With the use of transfer learning from an existing pre-trained 2D CNN model, there is improved performance on large-scale video action recognition tasks and bootstrapping from pre-trained 2D CNN makes the model learn fast resulting in better performance. In addition, I3D also improves the model's performance and prevents overfitting issues. I3D is best at learning low-level temporal and spatial information, while it struggles with higher-level features[121].

5.3.4 Spatio-Temporal Attention (STA)

I3D and 3DCNN architectures extract feature from spatial and temporal regions of video but, they are unable to distinguish between the frames. In reality, different frames convey different information to action recognition. Understanding each frame separately in a video recognition task and giving them attention according to the need helps in improving video classification and recognition capabilities. Thus, to improve the power of understanding and learning each frame separately by giving attention, we use a spatio-temporal attention (STA) module[4] that needs to be appended to the later convolutional layers without increasing much computational cost.
The STA module is divided into three functional modules:

- Temporal Attention function
- Spatial Attention function
- Spatio-Temporal Attention function

5.3.4.1 Temporal Attention function

In the temporal attention function, features in video frames are distinguished by learning frame-wise weight matrices W_t using the transform function τ_t. The variance data gets lowered during the extraction of information from the input frames. This information is further reduced in the operations of convolution leading to losing significant information required for action recognition. The temporal function τ_t is used to overcome this situation. First Deconvolution (DeConv) operation is applied using τ_t to expand the temporal dimension for preserving more temporal information. Later, the Convolution layer is used to

squeeze the dimension back to its original dimensions. The temporal attention function[4] is expressed mathematically as:

$$\tau_t = \delta_t \circ S_t \circ \mho_t \circ \epsilon_t \tag{5.6}$$

where S_t and ϵ_t are the squeezing and expanding operations respectively. $\circ$ denotes the composition operation over multiple functions, generating the compound function. $\mho_t$ and δ_t are the ReLU and sigmoid non-linear activation functions respectively.

5.3.4.2 Spatial Attention function

The spatial attention function $\tau_s(.)$ works similar to the temporal attention function, but the spatial attention function differentiates meaningful channels and designates a score for each channel at the channel level. The spatial attention function[4] is mathematically expressed as:

$$\tau_s = \delta_s \circ \epsilon_s \circ \mho_s \circ S_s \tag{5.7}$$

Where ϵ_s and S_s are the expanding and squeezing operations respectively. δ_s and $\mho_s$ are sigmoid and ReLU activation function respectvely. The spatial attention function uses spatial convolution and channel-level deconvolution operations to focus on the spatial dimension of the video file to extract essential information.

5.3.4.3 Spatial-Temporal Attention function

The spatial and temporal attention functions are combined together to get spatial-temporal attention (STA) module that can continue to work on frame-wise and channel-wise weights. The architecture of the STA network module is shown in Figure 5.4, which is embedded in the 3DCNN model.
The STA module attempts to learn the attention ($W = w_{ij}, i = 1\ to\ l, j = 1\ to\ c$) weighting the frames($i = 1\ to\ l$) and channels($j = 1\ to\ c$) in temporal and spatial dimension.
The spatial-temporal attention extracting process includes one channel squeezing operation S_s and one expanding operation ϵ_s. In order to enhance the representational power of the model, both sigmoid and ReLU activation functions are essential. Input feature maps are weighted by W which is output from the STA

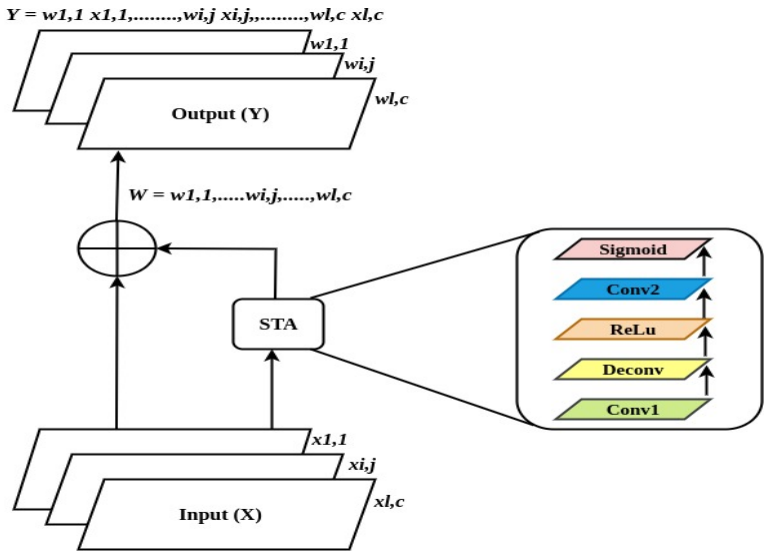


Figure 5.4: The architecture of STA network module [4].

module. An element-wise multiplication of the input feature map and weights is indicated by $\oplus$.
A spatio-temporal attention (STA) is composed of spatial attention and temporal attention given as:

$$\tau = \delta \circ (S_t \circ \epsilon_s) \circ \mho \circ (\epsilon_t \circ S_s) \circ P \tag{5.8}$$

where P is the Cartesian product of the squeezing function in the spatial and temporal attention function. P can be expressed as

$$P = P_s \times P_t \tag{5.9}$$

The transformation τ denotes both linear compression(convolution)/expansion(deconvolution) operations and a non-linear activation functions ReLU and Sigmoid. The layers of the STA module shown in Figure 5.4, depict the following functionality:

- Conv1 layer denotes a function that combines and operates along spatial and temporal squeeze operations.
- Deconv denotes the Deconvolution function that operates by squeezing and expanding the dimensions.
- Conv2 reduces the dimensions in the temporal field and decreases the parameters for faster training.

5.3.5 STA+I3D

The STA module, which is combined with I3D improves the spatio-temporal feature representation in 3DCNN for effective unethical human action recognition in video enhancing the accuracy and non-linear learning capability. I3D uses pre-trained 2D image model weights through bootstrapping and using Inception architecture, creates a very deep architecture for human action recognition. On the other hand, STA gives correct weightage to relevant features in both the temporal and spatial regions, increasing learning ability over a wide range of datasets. We introduce a novel architecture that combines I3D (which is already trained on large datasets such as Kinetics-400, ImageNet) and STA (with attention capability in spatial and temporal regions). The proposed hybrid model (a fusion of STA and I3D)is applied to large and complex benchmark video action datasets. The equation 5.10 for the proposed hybrid model can be mathematically formulated using equation 5.8 and equation 5.5.

$$STA + I3D = \tau \bigoplus P^l \tag{5.10}$$

The symbol $\bigoplus$ denotes concatenation. Videos are pre-processed and fed into the I3D model, which is trained on ImageNet and Kinetics-400 datasets to create a fine-tuned I3D model. Using the STA module, we can boost learning capability and improve accuracy in unethical human action recognition by integrating it into a 3DCNN architecture. If the proposed model performs better than the existing models, we save the model and deploy it as the better model else we use I3D fine-tuned model. In the results and analysis section, we analyse the performance of our proposed methodology over a different set of datasets and compare the same with existing models in the literature.

5.4 Results and analysis

The proposed action recognition method is evaluated using six publically available datasets: KTH[122], Weizmann[108], HMDB51[109], UCF-101[110], NPDI[111], and UCF-Crime[112]. The subset of action classes from the Weizmann, HMDB51, UCF-101, NPDI, and UCF-Crime datasets are combined to generate a new multi-action dataset. We considered accuracy and loss as the most essential performance criterion while analysing the action recognition. Experiments are conducted on a variety of datasets using several architectures, including 1. 3DCNN, 2. 3DCNN discriminator, 3. I3D, and 4. STA+I3D.

5.4.1 Datasets

KTH video database contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping). Currently, the database contains 2391 video sequences[122].
The Weizmann dataset is introduced by Blank in 2005 and consists of 10 actions such as: bending, jumping jack, jumping, jumping in place, running, galloping sideways, skipping, walking, one-hand-waving, and two-hand-waving. Each of these actions is performed by 9 actors resulting in 90 videos[108].
HMDB51 dataset is collected from various sources mostly from movies and a small proportion from public databases such as YouTube and Google videos. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips[109].
UCF-101 is a dataset of real action videos obtained from YouTube for action recognition. The dataset contains 13320 videos with 101 action classes. The dataset is broad and complex containing action videos with a wide range of variations such as cluttered backgrounds, camera motion, object appearance and posture, viewpoint, object scale, illumination conditions, and so on[110].
The NPDI pornographic dataset is one of the largest publicly available dataset used for research purposes. This dataset consists of around 80 hours of 400 adults and 400 non-adult videos[111]. In our experiments, we use pornographic-easy and non-pornographic-difficulty video clips for identifying the pornographic actions using the proposed model.
The UCF-Crime dataset includes a massive 128-hour video collection. It includes 1900 lengthy and uncut real-world surveillance videos with 13 actual anomalies such as Abuse, Arson, Arrest, Assault, Burglary, Explosion, Fighting, Robbery, Road accident, Shooting, Stealing, Shoplifting, and Vandalism[112].
Figure 5.5 shows various action classes from benchmark datasets like KTH, Weizmann, HMDB51, UCF-101, NPDI, and UCF-Crime. We select 24 subsets of classes for training and testing from benchmark datasets like Weizmann, HMDB51, UCF-Crime, UCF-101, and NPDI to build a multi-action dataset. The number of action classes is limited to 24 with 2938 video clips due to a lack of computational resources. The three classes of actions are differentiated as follows: normal, unethical, and porn as shown in Table 5.1. The deep learning-based hybrid model that we proposed is capable of accurately learning features of 24 classes and recognizing diverse activities along with prediction.

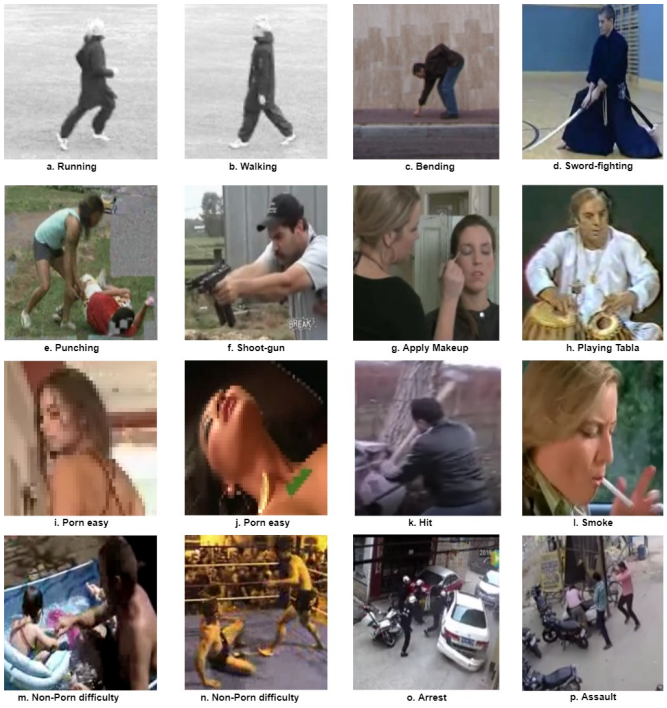


Figure 5.5: Examples of an action frame from benchmark datasets.

5.4.1.1 Implementation details

According to Deep Learning standards, the selected datasets are split randomly in a ratio of 75:25 for training and testing. The proposed unethical human action recognition system is implemented using Python3, deep learning libraries Keras (with Tensorflow backend), and OpenCV, which provide high-level building blocks for deep learning model development. We conducted numerous tests to verify the way network parameters are initialized, such that our model can learn and classify properly. We have used Intel i7, 2.20GHz processor with 8GB GPU(NVIDIA RTX 2080) and 32 GB RAM for our experimental evaluation.

5.4.1.2 Results

3DCNN model trained with KTH dataset

The 3DCNN model is evaluated on the KTH dataset, with a setup using 16-frame depth as input. The original input frames of 160 x 120 are reduced to 120 x 120 resolutions. The 3DCNN architecture as shown in Figure 5.2, consists of 120 x 120 x 11 inputs. In each layer, the kernel size and the number of filters are

Table 5.1: Multi-action dataset classes

	Number of classes	**Number of video clips**
Normal	15 (climb stairs, dive, eat, flic-flac, hug, catch, jump, laugh, pushup, ride bike, stand, walk, wave, apply-makeup, playing-tabla)	1446
Unethical	07 (hit, punch, shoot gun, smoke, sword-fighting, arrest, assault)	692
Porn	02 (porn easy, non-porn difficulty)	800

defined. In 3DCNN, the two convolutional layers use kernel filter of sizes 11 x 11 x 32 and 5 x 5 x 32, and two layers of 2 x 2 max pooling are used to reduce the training parameters. The fully connected layer has activations in the previous layer which is transformed into 6400-dimensional feature vectors. The softmax layer consists of output units that result in the number of action classes. The configuration details of the model are presented in Table 5.2. The categorical

Table 5.2: Configuration details for the 3DCNN

Batch size	16
Number of epochs	250
Loss function	Categorical cross entropy
Optimizer	Adam
Evaluation metric	Accuracy

cross-entropy is a loss function that evaluates the performance of a multi-class classification model with a probability value as its output. The loss is increased when the predicted probability varies from the actual label. Cross-entropy loss is

estimated using equation 5.11 in multi-class classification.

$$lossfunction = -\sum_{i=1}^{n} t_i log(p_i) \tag{5.11}$$

The parameter t_i is the truth label, p_i is the softmax probability for the i^{th} classes, and n is the number of classes. Figure 5.6 depicts the training and test accuracy of the 3DCNN model on KTH dataset with respect to the number of epochs. Figure 5.7 shows the loss curve during training and testing. The

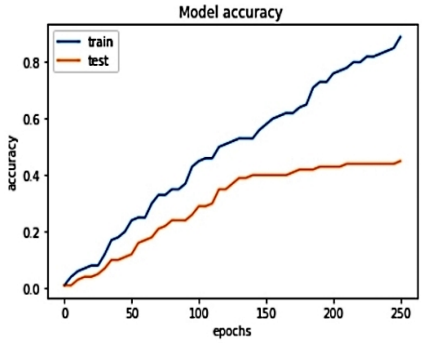


Figure 5.6: Train vs. test accuracy.

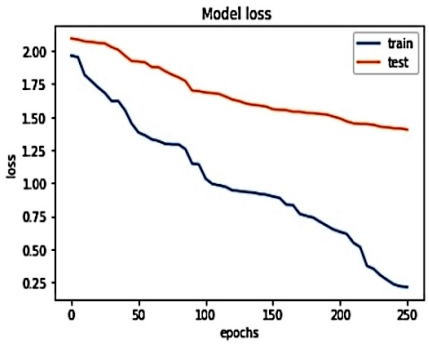


Figure 5.7: Train vs. test loss.

3DCNN model is unable to learn minor features across the video frames and is also unable to create additional trainable parameters due to the smaller datasets. Hence, the normal 3D CNN model is showing an overfitting issue.

3DCNN discriminator trained with Weizmann dataset.

The 3D convolution layer added with batch normalization and activations is used as a discriminator layer in the 3DCNN model to improve the classification accuracy trained with the Weizmann dataset. Figure 5.8 and Figure 5.9 shows the progress in accuracy and loss respectively. The configuration details of the model are presented in Table 5.3. Figure 5.8 and Figure 5.9 shows improvement

Table 5.3: Configuration information for 3DCNN discriminator

Batch size	8
Number of epochs	400
Loss function	Categorical cross entropy
Optimizer	Adam
Evaluation metric	Accuracy

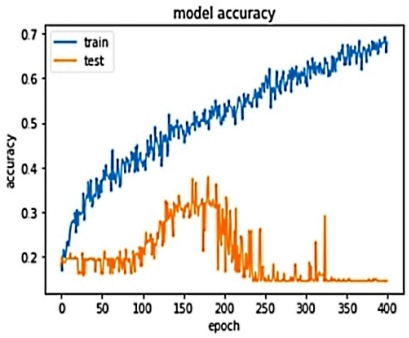


Figure 5.8: Training vs. testing accuracy.

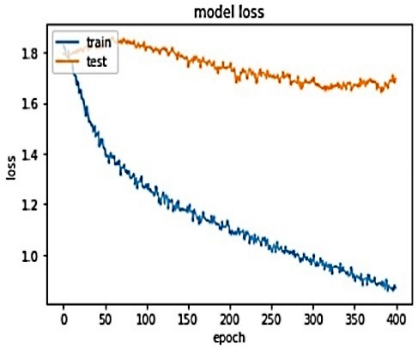


Figure 5.9: Training vs. testing loss.

in accuracy and loss. The model configuration information is presented in Table 5.3. Due to fewer features being learned and small dataset, the model still faces an overfitting issue. The variety of dataset types can influence the complexity of the networks. Single-view point datasets such as Weizmann and KTH use a single camera to capture human actions in confined spaces. Figure 5.10 shows human action video sequences of the Weizmann dataset for hand-waving action where actors in the video clips show simple identical action.

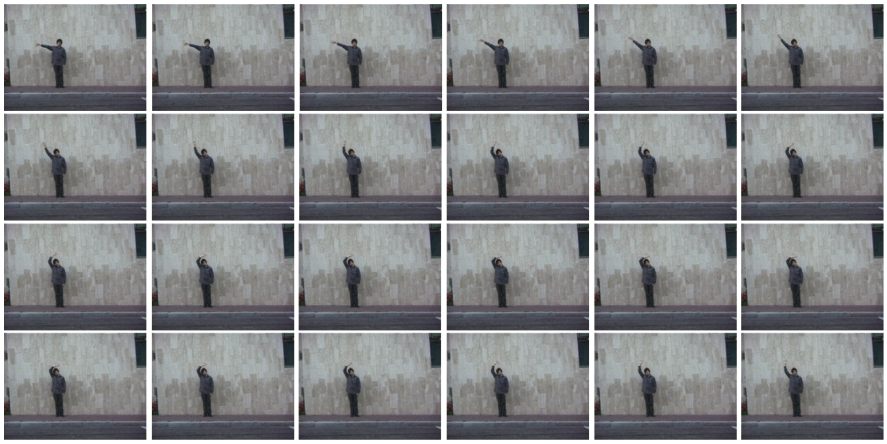


Figure 5.10: Human action video sequences of hand-waving from Weizmann dataset.

I3D model trained with multi-action dataset.

With a pre-trained Kinetics-400 dataset, the I3D model significantly improves human action classification eliminating the overfitting problem as shown in Fig-

ure 5.11 and Figure 5.12. Also, the training accuracy increased to 95.6% on the multi-action dataset. The details of the model configuration are shown in Table 5.4. I3D models based on Inception-v1 have better performance on small

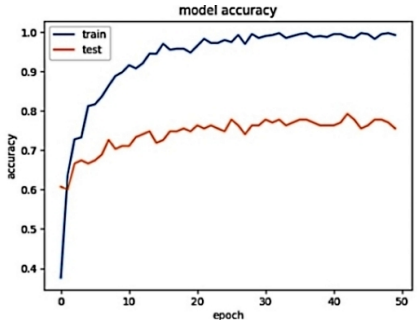


Figure 5.11: Train vs. test accuracy.

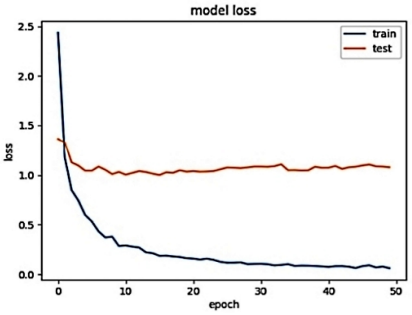


Figure 5.12: Train vs. test loss.

Table 5.4: I3D model configuration information.

Batch size	8
Number of epochs	50
Loss function	Categorical cross entropy
Optimizer	SGD
Evaluation metric	Accuracy

benchmark datasets after pre-training on Kinetics human action video dataset. However, the I3D model is unable to distinguish small features among the adjacent frames resulting in low learning capability on complex datasets. Figure 5.13 shows complex human action video sequences of the Assault action class from the UCF-crime dataset.

Figure 5.13: Human action video sequences of Assault from UCF-crime dataset.

Training STA+I3D on multi-action dataset.
By providing weightage to specific spatial and temporal features, the proposed model STA+I3D provides a solution to 3DCNN and I3D models. The model is efficient and robust in handling diverse human action video datasets. Figure 5.14 and Figure 5.15 shows the accuracy and loss for a multi-action dataset after 50 iterations of STA+I3D. It can be seen that we have achieved an average training accuracy rate of 98.03% with the multi-action dataset. The configuration details of the model are shown in Table 5.5.

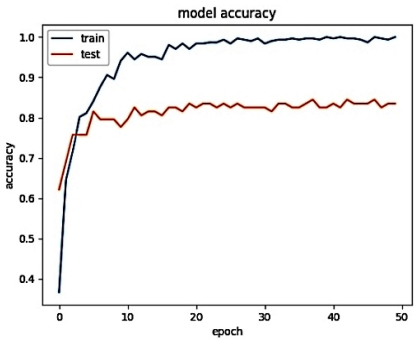


Figure 5.14: Train vs. test accuracy.

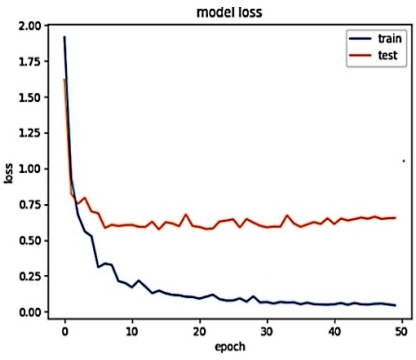


Figure 5.15: Train vs. test loss.

The confusion matrix of STA+I3D on a multi-action dataset over a subset of five classes (considered only 5 due to space constraint) is shown in Figure 5.16.

Table 5.5: Details of the STA+I3D configuration.

Batch size	4
Number of epochs	50
Loss function	Categorical cross entropy
Optimizer	SGD
Evaluation metric	Accuracy

The model's excellence positivity rate is determined by the predicted and actual values of the confusion matrix, indicating good classification accuracy.

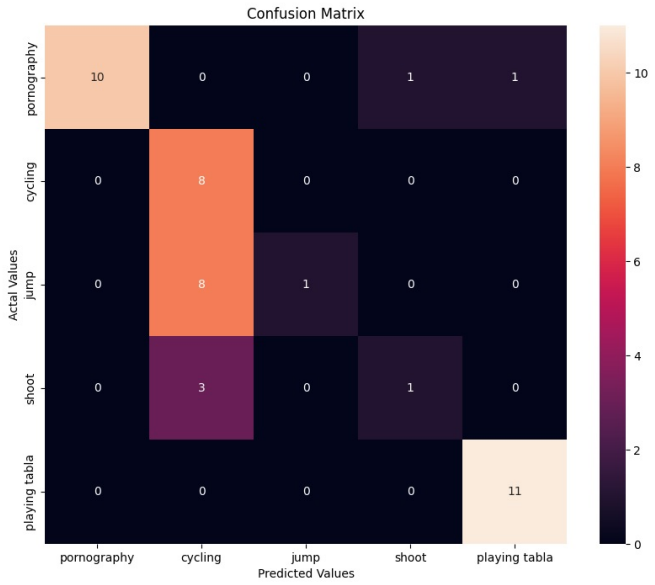


Figure 5.16: Confusion matrix on a multi-action dataset for five classes.

The model predicts incorrectly jumping as cycling in the confusion matrix, indicating that similar features of cycling and jumping are misunderstood by the model, and there is a need for extra hyper-parameter tuning. The multi-action test data prediction results are shown in Figure 5.17. The predictions are sorted in decreasing order of confidence.

5.4.1.3 Comparison

The values of the parameters in deep learning models determine how accurately the model accomplishes the task for a specific architecture. Deep neural networks

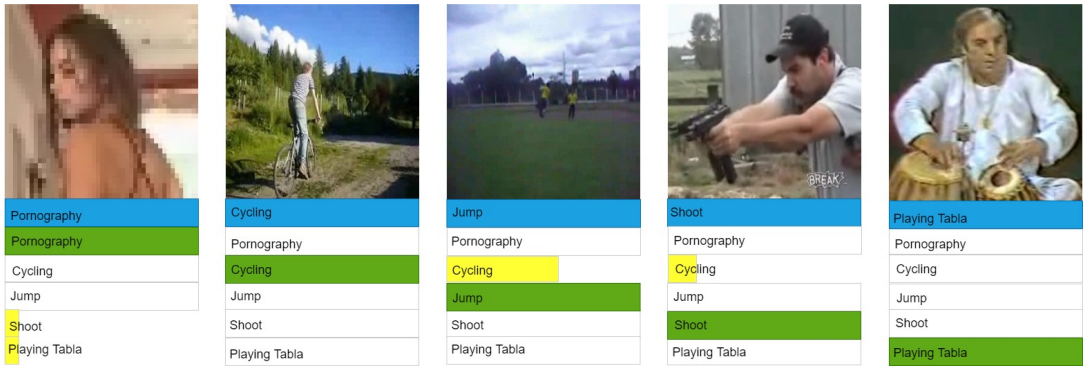


Figure 5.17: Multi-action test data predictions. The actual label is shown in the blue bar (first row). The green and yellow bars distinguish percentage-wise correct and incorrect predictions.

thrive with a large number of parameters, allowing the model to represent more complex features. The number of parameters is defined by the number of layers in the network, the number of units in every layer, and the dimensionality of the input and the output. When compared to other models, our proposed model STA+I3D obtained the maximum number of parameters as shown in Table 5.6.

Table 5.6: Parameters of different models.

Model	**No. of Parameters**
Normal 3DCNN	9,57,990
3DCNN with Discriminator	8,090,279
I3D Fine tuned model	12,704,544
STA+I3D model	13,079,162

Comparative analysis is presented in Table 5.7 against the state-of-the-art approaches for recognizing human actions. Our proposed model, STA+I3D, employs a multi-action dataset that requires less computational time and performs well during training and testing, learning complex features in videos.

Table 5.7: Performance comparison of STA+I3D model with existing models using benchmark datasets. '-'denotes that the result is not available in the literature.

Architecture	HMDB51	UCF-101	NPDI	UCF-Crime	Multi-action dataset
(a)AGNet[123]	–	–	93.8	–	–
(b)MiCTNet[124]	63.8	–	–	–	–
(c)VGGC3D[125]	–	–	95.1	–	–
(d)I3D+LSTM[121]	–	95.1	–	–	–
(e)Motion+TCNN[126]		–	–	31	–
(f)FineTuned 3DCNN[127]	–	–	–	45	–
(g)I3D[3]	–	–	–	–	78.2
(h)STA+I3D	**81.4**	**96.3**	**97.2**	**62.2**	**82.2**

From the analysis and experimental results of 3DCNN, I3D, and STA+I3D models, we have identified some of the advantages and drawbacks as depicted in Table 5.8.

Table 5.8: Advantages and drawbacks of 3DCNN, I3D, and STA+I3D.

	Drawbacks	**Advantages**
3DCNN	1. This approach fails to recognize complex actions in videos.	1. Used to extract spatio-temporal features.
	2. Need more hyper-parameter tuning compared to other techniques.	2. It works better with larger datasets.
I3D	1. It is not powerful in extracting high level features.	1. I3D improves action recognition performance by inflating 2D filters to 3D filters.
	2. It is unable to distinguish small features among adjacent frames.	2. It provides pre-trained action classification weights as it is trained with large dataset of Kinetics-400.
STA+I3D	1. Fails to differentiate multiple actions within video frames as it does not provide any localization of objects inside a video.	1. STA+I3D provides appropriate weightage to relevant features of temporal and spatial regions.
	2. STA may not be compatible with all the deep learning architectures.	2. It is robust to handle diverse action recognition in existing benchmark datasets.

From the literature study, we found that various action recognition models are designed and trained with unique datasets to show the model's performance. In our proposed model, we have used unique and multi-action datasets to show human action recognition accuracy in large and complex video actions categorized into normal, unethical, and porn as shown in Table 5.1. We were able to obtain better action recognition accuracy and improved learning capability in complex actions of spatio-temporal features present in videos.

5.5 Summary

Human Action Recognition (HAR) is a task that involves monitoring human activity in a variety of environments like visual surveillance, elderly behavior monitoring, unethical activity recognition, etc. Cybercrimes using videos are increasing drastically and there is a need for unethical human action recognition in these files to minimize the forensic analysis time. In this Chapter, we addressed the problem of complex unethical human action recognition and improved the high-level feature learning capability by using the fusion of STA and I3D. The experimental results compared with the state-of-the-art 3D CNN approaches using unique and multi-action datasets, have shown that STA+I3D provides better performance in unethical human action recognition by learning accurately complex spatio-temporal features in videos. In the future, we look forward to extending the proposed model to other applications like video surveillance, video forgery recognition (i.e., video object localization, video object detection), semantic segmentation, human-computer interaction, etc.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

Multimedia forensics deals with analyzing multimedia content to have the legality of evidence in the court for proving its authenticity and integrity. In recent times, many methods (active and passive) have been proposed in the literature to analyze multimedia content. Passive forensic approaches have a direct impact on multimedia forensics and have proven better approaches to combat cybercrimes compared to active forensic approaches. Due to the growth of massive amounts of multimedia data, forensic investigators face enormous challenges in analyzing and processing the same. We focused on image as well as video forgery analysis from a multimedia forensics perspective and tried to minimize the gaps in existing detection techniques.

In terms of detecting and localizing image forgery, we used a pre-trained LSTM-CNN based hybrid model to detect (copy-move and image splicing) forgery operations. We classify the forgery operations (copy-move or image splicing) by template matching with an improved SIFT algorithm to achieve better efficiency and make feature point matching more accurate with scale and rotation invariance. The proposed model was tested on benchmark datasets resulting in better performance accuracy compared to existing models. The use of deep learning-based methods for image/video tampering detection could help in faster processing speed, handling a large amount of data, increased stability, better identification and classification, and getting optimal accuracy.

We designed a video forgery detection technique using deep learning-based methods by proposing a 3-dimensional CNN model for detecting inter-frame forg-

eries (insertion and deletion) and MS-SSIM approach for localizing the forgeries. The proposed model learns more relevant characteristics to detect video inter-frame forgeries with high classification accuracy and localize the number of frames forged. The model outperforms when compared to existing models in both static and dynamic videos.

Unethical human action recognition methods are required for surveillance and security, cyber forensic investigation, human rights protection, workplace safety, and social media moderation. It is important to note that the design of such a recognition system needs to address complex actions and high-level features. The problem of complex unethical human action recognition and high-level feature learning capability is addressed using the fusion of STA and I3D. As compared to existing human action recognition methods, the novel model STA+I3D provides better performance by learning accurately complex spatio-temporal features.

6.2 Future Work

Deep learning and artificial intelligence are important knowledge areas that have provided solutions allowing the successful resolution of complex problems. Our research focused on image forgery (copy-move, image-splicing) detection as well as localization and video forgery (insertion and deletion) detection with localization. In the future, we look forward to working on other distinct forms of image and video forgery detection. A universal image and video forgery detection technique with localization is the future research that should be robust and efficient in handling several post-processing operations. Advanced deep learning-based architecture like transformers can be applied for inter-frame video forgery detection since transformers use a self-attention mechanism to weigh different parts of the input sequence and make predictions. Advances in AI, especially deep neural networks (DNN) and Generative adversarial networks (GAN), make deepfake images and videos much easier, cheaper, and simpler to generate deepfake. Our future work is to progress in deep learning and transformer-based approaches for deepfake image and video detection. Some of the ways in which deep learning and transformers can be used for deepfake detection include. 1. Image analysis, 2. Video analysis, 3. Deep feature extraction, 4. Adversarial training, and 5. Multi-modal analysis. These approaches have shown promising results in the detection of deepfakes, but the field is still evolving, and much research is needed to improve the performance analysis of deepfake detection methods.

List of Publications

Journals

1. Gowda Raghavendra, and Digambar Pawar. **Deep learning-based forgery identification and localization in videos**. *Signal, Image and Video Processing* (2022): Page 1-8. https://doi.org/10.1007/s11760-022-02433-7. Indexed: Science Citation Index Expanded (SCIE), SCOPUS, UGC-CARE List (India)
   Status: Accepted and Published

2. Gowda, R., Pawar, D., and Barman, B. **Unethical human action recognition using deep learning based hybrid model for video forensics**, *Multimedia Tools and Applications* (2023): Page 1-26. https://doi.org/10.1007/s11042-023-14508-9. Indexed: Science Citation Index Expanded (SCIE), SCOPUS, UGC-CARE List (India)
   Status: Accepted and Published

3. Gowda Raghavendra, and Digambar Pawar. **LSTM-CNN based hybrid model for image forgery detection-Digital forensics perspective**, *Multidimensional Systems and Signal Processing, An International Journal.* Indexed: SCI, SCIE, SCOPUS, UGC-CARE List (India)
   Status: Under Review

Conference Proceedings

1. Gowda Raghavendra, and Digambar Pawar. **Porn Image Forensics: Image Classification, Forgery Detection and Localization**. In *International Conference on Computing in Engineering & Technology* (2022) (pp.359-371). https://doi.org/10.1007/978-981-19-2719-5-34. Indexed: SCOPUS.
   Status: Accepted and Published

2. Gowda Raghavendra, Digambar Pawar., and Tetali, S. R. **Multimedia Forensics-An approach to detect and analyze Human faces in multimedia files**. In 2019 Fifth *International Conference on Image Information Processing (ICIIP)* (2019) (pp.274-279). https://doi.org/10.1109/ICIIP47207.2019.8985694. Indexed: IEEE Xplore.
   Status: Accepted and Published

References

[1] Nitin Arvind Shelke and Singara Singh Kasana. **A comprehensive survey on passive techniques for digital video forgery detection**. *Multimedia Tools and Applications*, **80**(4):6247–6310, 2021.

[2] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. **Image forgery localization via fine-grained analysis of CFA artifacts**. *IEEE Transactions on Information Forensics and Security*, **7**(5):1566–1577, 2012.

[3] Joao Carreira and Andrew Zisserman. **Quo vadis, action recognition? a new model and the kinetics dataset**. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. **Spatio-temporal attention networks for action recognition and detection**. *IEEE Transactions on Multimedia*, **22**(11):2990–3001, 2020.

[5] Jatin Patel and Ravi Sheth. **An Optimized Convolutional Neural Network Based Inter-Frame Forgery Detection Model-A Multi-Feature Eextraction Framework**. *ICTACT Journal on image and video processing*, **12**(02):2570–2581, 2021.

[6] Sebastiano Battiato, Oliver Giudice, and Antonino Paratore. **Multimedia forensics: discovering the history of multimedia contents**. In *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, pages 5–16, 2016.

[7] Karen Kent, Suzanne Chevalier, and Tim Grance. **Guide to integrating forensic techniques into incident**. *Tech. Rep. 800-86*, 2006.

[8] Judith A Redi, Wiem Taktak, and Jean-Luc Dugelay. **Digital image forensics: a booklet for beginners**. *Multimedia Tools and Applications*, **51**(1):133–162, 2011.

[9] Anjie Peng, Yadong Wu, and Xiangui Kang. **Revealing traces of image resampling and resampling antiforensics**. *Advances in Multimedia*, **2017**, 2017.

[10] Sami Bourouis, Roobaea Alroobaea, Abdullah M Alharbi, Murad Andejany, and Saeed Rubaiee. **Recent advances in digital multimedia tampering detection for forensics analysis**. *Symmetry*, **12**(11):1811, 2020.

[11] Kunj Bihari Meena and Vipin Tyagi. **Image Splicing Forgery Detection Techniques: A Review**. In *International Conference on Advances in Computing and Data Sciences*, pages 364–388. Springer, 2021.

[12] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. **Hybrid lstm and encoder–decoder architecture for detection of image forgeries**. *IEEE Transactions on Image Processing*, **28**(7):3286–3300, 2019.

[13] David G Lowe. **Distinctive image features from scale-invariant keypoints**. *International journal of computer vision*, **60**(2):91–110, 2004.

[14] Naheed Akhtar, Mubbashar Saddique, Khurshid Asghar, Usama Ijaz Bajwa, Muhammad Hussain, and Zulfiqar Habib. **Digital Video Tampering Detection and Localization: Review, Representations, Challenges and Algorithm**. *Mathematics*, **10**(2):168, 2022.

[15] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. **3D convolutional neural networks for human action recognition**. *IEEE transactions on pattern analysis and machine intelligence*, **35**(1):221–231, 2012.

[16] Guocheng Liu, Caixia Zhang, Qingyang Xu, Ruoshi Cheng, Yong Song, Xianfeng Yuan, and Jie Sun. **I3d-shufflenet based human action Recognition**. *Algorithms*, **13**(11):301, 2020.

[17] M Abdel-Salam Nasr, Mohammed F AlRahmawy, and AS Tolba. **Multi-scale structural similarity index for motion detection**. *Journal of King Saud University-Computer and Information Sciences*, **29**(3):399–409, 2017.

[18] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. **Multiscale structural similarity for image quality assessment**. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, **2**, pages 1398–1402. Ieee, 2003.

[19] Syed Tufael Nabi, Munish Kumar, Paramjeet Singh, Naveen Aggarwal, and Krishan Kumar. **A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions**. *Multimedia Systems*, pages 1–54, 2022.

[20] Muhammad Ali Qureshi and Mohamed Deriche. **A bibliography of pixel-based blind image forgery detection techniques**. *Signal Processing: Image Communication*, **39**:46–74, 2015.

[21] Hany Farid. **Image forgery detection**. *IEEE Signal processing magazine*, **26**(2):16–25, 2009.

[22] Amani Alahmadi, Muhammad Hussain, Hatim Aboalsamh, Ghulam Muhammad, George Bebis, and Hassan Mathkour. **Passive detection of image forgery using DCT and local binary pattern**. *Signal, Image and Video Processing*, **11**(1):81–88, 2017.

[23] Leida Li, Shushang Li, Hancheng Zhu, Shu-Chuan Chu, John F Roddick, and Jeng-Shyang Pan. **An Efficient Scheme for Detecting Copy-move Forged Images by Local Binary Patterns.** *J. Inf. Hiding Multim. Signal Process.*, **4**(1):46–56, 2013.

[24] M Bashar, K Noda, N Ohnishi, and K Mori. **Exploring duplicated regions in natural images**. *IEEE Transactions on Image Processing*, 2010.

[25] Ting Zhang and Rang-ding Wang. **Copy-move forgery detection based on SVD in digital image**. In *2009 2nd International Congress on Image and Signal Processing*, pages 1–5. IEEE, 2009.

[26] Jonathon Shlens. **A tutorial on principal component analysis**. *arXiv preprint arXiv:1404.1100*, 2014.

[27] Hailing Huang, Weiqiang Guo, and Yu Zhang. **Detection of copy-move forgery in digital images using SIFT algorithm**. In *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, **2**, pages 272–276. IEEE, 2008.

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

[29] Xu Bo, Wang Junwen, Liu Guangjie, and Dai Yuewei. **Image copy-move forgery detection based on SURF**. In *2010 International Conference on Multimedia Information Networking and Security*, pages 889–892. IEEE, 2010.

[30] Ye Zhu, Xuanjing Shen, and Haipeng Chen. **Copy-move forgery detection based on scaled ORB**. *Multimedia Tools and Applications*, **75**(6):3221–3233, 2016.

[31] Huan Wang, Hong-Xia Wang, Xing-Ming Sun, and Qing Qian. **A passive authentication scheme for copy-move forgery based on package clustering algorithm**. *Multimedia Tools and Applications*, **76**(10):12627–12644, 2017.

[32] Gulnawaz Gani and Fasel Qadir. **A robust copy-move forgery detection technique based on discrete cosine transform and cellular automata**. *Journal of Information Security and Applications*, **54**:102510, 2020.

[33] Jiangbin Zheng, Yanan Liu, Jinchang Ren, Tingge Zhu, Yijun Yan, and Heng Yang. **Fusion of block and keypoints based approaches for effective copy-move image forgery detection**. *Multidimensional Systems and Signal Processing*, **27**(4):989–1005, 2016.

[34] Kunj Bihari Meena and Vipin Tyagi. **A hybrid copy-move image forgery detection technique based on Fourier-Mellin and scale invariant feature transforms**. *Multimedia Tools and Applications*, **79**(11):8197–8212, 2020.

[35] Micah K Johnson and Hany Farid. **Exposing digital forgeries by detecting inconsistencies in lighting**. In *Proceedings of the 7th workshop on Multimedia and security*, pages 1–10, 2005.

[36] Tian-Tsong Ng, Shih-Fu Chang, and Qibin Sun. **Blind detection of photomontage using higher order statistics**. In *2004 IEEE International Symposium on Circuits and Systems (ISCAS)*, **5**, pages V–V. IEEE, 2004.

[37] Zhongwei He, Wei Lu, Wei Sun, and Jiwu Huang. **Digital image splicing detection based on Markov features in DCT and DWT domain**. *Pattern recognition*, **45**(12):4292–4299, 2012.

[38] Avinash Kumar, Choudhary Shyam Prakash, Sushila Maheshkar, and Vikas Maheshkar. **Markov feature extraction using enhanced threshold method for image splicing forgery detection**. In *Smart Innovations in Communication and Computational Sciences*, pages 17–27. Springer, 2019.

[39] Ying Zhang, Jonathan Goh, Lei Lei Win, and Vrizlynn LL Thing. **Image region forgery detection: A deep learning approach.** *SG-CRC*, **2016**:1–11, 2016.

[40] Yuan Rao and Jiangqun Ni. **A deep learning approach to detection of splicing and copy-move forgeries in images**. In *2016 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2016.

[41] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. **Image copy-move forgery detection via an end-to-end deep neural network**. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1907–1915. IEEE, 2018.

[42] Junlin Ouyang, Yizhi Liu, and Miao Liao. **Copy-move forgery detection based on deep learning**. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, pages 1–5. IEEE, 2017.

[43] Thales Pomari, Guillherme Ruppert, Edmar Rezende, Anderson Rocha, and Tiago Carvalho. **Image splicing detection through illumination inconsistencies and deep learning**. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3788–3792. IEEE, 2018.

[44] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. **RRU-Net: The ringed residual U-Net for image splicing forgery detection**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[45] Neeru Jindal et al. **Copy move and splicing forgery detection using deep convolution neural network, and semantic segmentation**. *Multimedia Tools and Applications*, **80**(3):3571–3599, 2021.

[46] Dan Zhao and Xuedong Tian. **A Multiscale Fusion Lightweight Image-Splicing Tamper-Detection Model**. *Electronics*, **11**(16):2621, 2022.

[47] Sondos Fadl, Qi Han, and Qiong Li. **CNN spatiotemporal features and fusion for surveillance video forgery detection**. *Signal Processing: Image Communication*, **90**:116066, 2021.

[48] Matthew C Stamm, W Sabrina Lin, and KJ Ray Liu. **Temporal forensics and anti-forensics for motion compensated video**. *IEEE Transactions on Information Forensics and Security*, **7**(4):1315–1329, 2012.

[49] Juan Chao, Xinghao Jiang, and Tanfeng Sun. **A novel video inter-frame forgery model detection scheme based on optical flow consistency**. In *International Workshop on Digital Watermarking*, pages 267–281. Springer, 2012.

[50] Wan Wang, Xinghao Jiang, Shilin Wang, Meng Wan, and Tanfeng Sun. **Identifying video forgery process using optical flow**. In *International Workshop on Digital Watermarking*, pages 244–257. Springer, 2013.

[51] Alessandra Gironi, Marco Fontani, Tiziano Bianchi, Alessandro Piva, and Mauro Barni. **A video forensic technique for detecting frame deletion and insertion**. In *2014 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6226–6230. IEEE, 2014.

[52] K Sitara and BM Mehtre. **A comprehensive approach for exposing inter-frame video forgeries**. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pages 73–78. IEEE, 2017.

[53] Staffy Kingra, Naveen Aggarwal, and Raahat Devender Singh. **Inter-frame forgery detection in H. 264 videos using motion and brightness gradients**. *Multimedia Tools and Applications*, **76**(24):25767–25786, 2017.

[54] Dong-Ning Zhao, Ren-Kui Wang, and Zhe-Ming Lu. **Inter-frame passive-blind forgery detection for video shot based on similarity analysis**. *Multimedia Tools and Applications*, **77**(19):25389–25408, 2018.

[55] Vinay Kumar and Manish Gaur. **Multiple forgery detection in video using inter-frame correlation distance with dual-threshold**. *Multimedia Tools and Applications*, pages 1–20, 2022.

[56] Xiao Jin, Yuting Su, and Peiguang Jing. **Video frame deletion detection based on time–frequency analysis**. *Journal of Visual Communication and Image Representation*, **83**:103436, 2022.

[57] Han Pu, Tianqiang Huang, Bin Weng, Feng Ye, and Chenbin Zhao. **Overcome the Brightness and Jitter Noises in Video Inter-Frame Tampering Detection**. *Sensors*, **21**(12):3953, 2021.

[58] Chengjiang Long, Arslan Basharat, and Anthony Hoogs. **A Coarse-to-fine Deep Convolutional Neural Network Framework for Frame Duplication Detection and Localization in Video Forgery**. *arXiv preprint arXiv:1811.10762*, 2018.

[59] Harpreet Kaur and Neeru Jindal. **Deep convolutional neural network for graphics forgery detection in video**. *Wireless Personal Communications*, **112**(3):1763–1781, 2020.

[60] Xuan Hau Nguyen, Yongjian Hu, Muhmmad Ahmad Amin, Gohar Hayat Khan, Dinh-Tu Truong, et al. **Detecting video interframe forgeries based on convolutional neural network model**. *International Journal of Image, Graphics and Signal Processing*, **10**(3):1, 2020.

[61] Neetu Singla, Jyotsna Singh, and Sushama Nagpal. **Video Frame Deletion Detection using Correlation Coefficients**. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 796–799. IEEE, 2021.

[62] Hemal Mamtora, Kevin Doshi, Shreya Gokhale, Surekha Dholay, and Chandrashekhar Gajbhiye. **Video Manipulation Detection and Localization Using Deep Learning**. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 241–248. IEEE, 2020.

[63] Pamela Johnston, Eyad Elyan, and Chrisina Jayne. **Video tampering localisation using features learned from authentic content**. *Neural computing and applications*, **32**(16):12243–12257, 2020.

[64] Osamah M Al-Qershi and Bee Ee Khoo. **Passive detection of copy-move forgery in digital images: State-of-the-art**. *Forensic science international*, **231**(1-3):284–295, 2013.

[65] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. **Understanding of a convolutional neural network**. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[66] Larry R Medsker and LC Jain. **Recurrent neural networks**. *Design and Applications*, **5**:64–67, 2001.

[67] Shilpa Dua, Jyotsna Singh, and Harish Parthasarathy. **Detection and localization of forgery using statistics of DCT and Fourier components**. *Signal processing: image communication*, **82**:115778, 2020.

[68] Anuj Rani, Ajit Jain, and Manoj Kumar. **Identification of copy-move and splicing based forgeries using advanced SURF and revised template matching**. *Multimedia Tools and Applications*, **80**(16):23877–23898, 2021.

[69] Tajuddin Manhar Mohammed, Jason Bunk, Lakshmanan Nataraj, Jawadul H Bappy, Arjuna Flenner, BS Manjunath, Shivkumar Chandrasekaran, Amit K Roy-Chowdhury, and Lawrence A Peterson. **Boosting image forgery detection using resampling features and copy-move analysis**. *Electronic Imaging*, **2018**(7):118–1, 2018.

[70] Lin Peng, Xin Liao, and Mingliang Chen. **Resampling parameter estimation via dual-filtering based convolutional neural network**. *Multimedia Systems*, **27**(3):363–370, 2021.

[71] Jawadul Bappy, Tajuddin Manhar Mohammed, Lakshmanan Nataraj, Arjuna Flenner, Shivkumar Chandrasekaran, Amit Roy-Chowdhury, Jason HBSK Lawrence Peterson, et al. **Detection and localization of image forgeries using resampling features and deep learning**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 69–77, 2017.

[72] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A Araújo. **Pooling in image representation: The visual codeword point of view**. *Computer Vision and Image Understanding*, **117**(5):453–465, 2013.

[73] Hosagrahar V Jagadish. **Analysis of the Hilbert curve for representing two-dimensional space**. *Information Processing Letters*, **62**(1):17–22, 1997.

[74] Mauricio Marengoni and Denise Stringhini. **High level computer vision using opencv**. In *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, pages 11–24. IEEE, 2011.

[75] Yanyan Qin, Hongke Xu, and Huiru Chen. **Image feature points matching via improved ORB**. In *2014 IEEE International Conference on Progress in Informatics and Computing*, pages 204–208. IEEE, 2014.

[76] Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. **BRIEF: Computing a local binary descriptor very fast**. *IEEE transactions on pattern analysis and machine intelligence*, **34**(7):1281–1298, 2011.

[77] Shuqiang Yang and Biao Li. **Outliers elimination based ransac for fundamental matrix estimation**. In *2013 International Conference on Virtual Reality and Visualization*, pages 321–324. IEEE, 2013.

[78] Yu-Feng Hsu and Shih-Fu Chang. **Detecting image splicing using geometry invariants and camera characteristics consistency**. In *2006 IEEE International Conference on Multimedia and Expo*, pages 549–552. IEEE, 2006.

[79] Jing Dong, Wei Wang, and Tieniu Tan. **Casia image tampering detection evaluation database**. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426. IEEE, 2013.

[80] Dijana Tralic, Ivan Zupancic, Sonja Grgic, and Mislav Grgic. **CoMoFoD—New database for copy-move forgery detection**. In *Proceedings ELMAR-2013*, pages 49–54. IEEE, 2013.

[81] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. **COVERAGE—A novel database for copy-move forgery detection**. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016.

[82] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. **A sift-based forensic method for copy–move attack detection and transformation recovery**. *IEEE transactions on information forensics and security*, **6**(3):1099–1110, 2011.

[83] Neal Krawetz and Hacker Factor Solutions. **A picture's worth**. *Hacker Factor Solutions*, **6**(2):2, 2007.

[84] Babak Mahdian and Stanislav Saic. **Using noise inconsistencies for blind image forensics**. *Image and Vision Computing*, **27**(10):1497–1503, 2009.

[85] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. **Exploiting spatial structure for localizing manipulated image regions**. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979, 2017.

[86] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. **Learning rich features for image manipulation detection**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1053–1061, 2018.

[87] Xinyi Wang, He Wang, Shaozhang Niu, and Jiwei Zhang. **Detection and localization of image forgeries using improved mask regional convolutional neural network**. *Mathematical Biosciences and Engineering*, **16**(5):4581–4593, 2019.

[88] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. **Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019.

[89] D Jude Hemanth and V Vieira Estrela. *Deep learning for image processing applications*, **31**. IOS Press, 2017.

[90] Soren Goyal and Paul Benjamin. **Object recognition using deep neural networks: A survey**. *arXiv preprint arXiv:1412.3684*, 2014.

[91] Syed Tufael Nabi, Munish Kumar, Paramjeet Singh, Naveen Aggarwal, and Krishan Kumar. **A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions**. *Multimedia Systems*, pages 1–54, 2022.

[92] Xuan Hau Nguyen, Yongjian Hu, Muhmmad Ahmad Amin, Gohar Hayat Khan, Dinh-Tu Truong, et al. **Detecting video interframe forgeries based on convolutional neural network model**. *International Journal of Image, Graphics and Signal Processing*, **10**(3):1, 2020.

[93] Sondos Fadl, Qi Han, and Qiong Li. **CNN spatiotemporal features and fusion for surveillance video forgery detection**. *Signal Processing: Image Communication*, **90**:116066, 2021.

[94] Jamimamul Bakas, Ruchira Naskar, and Sambit Bakshi. **Detection and localization of inter-frame forgeries in videos based on macroblock variation and motion vector analysis**. *Computers & Electrical Engineering*, **89**:106929, 2021.

[95] Leif E. Peterson. **K-nearest neighbor**. `http://scholarpedia.org/article/K-nearest_neighbor`, 2009. [Online; accessed 19-April-2022].

[96] Shan Suthaharan. **Support vector machine**. In *Machine learning models and algorithms for big data classification*, pages 207–235. Springer, 2016.

[97] Sergey Ioffe and Christian Szegedy. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[98] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. **UCF101: A dataset of 101 human actions classes from videos in the wild**. *arXiv preprint arXiv:1212.0402*, 2012.

[99] Yongjian nguyen, xuan hau; Hu. **VIFFD - A dataset for detecting video inter-frame forgeries**. *Mendeley Data, V6,*, 2020.

[100] F.: Developers. **FFmpeg tool (Version 3.2.4) [Software]**. https://www.ffmpeg.org, 2017.

[101] Zhenzhen Zhang, Jianjun Hou, Qinglong Ma, and Zhaohong Li. **Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames**. *Security and Communication networks*, **8**(2):311–320, 2015.

[102] Liyang Yu, Huanran Wang, Qi Han, Xiamu Niu, Siu-Ming Yiu, Junbin Fang, and Zhifang Wang. **Exposing frame deletion by detecting abrupt changes in video streams**. *Neurocomputing*, **205**:84–91, 2016.

[103] Yuqing Liu and Tianqiang Huang. **Exposing video inter-frame forgery by Zernike opponent chromaticity moments and coarseness analysis**. *Multimedia Systems*, **23**(2):223–238, 2017.

[104] Markos Zampoglou, Foteini Markatopoulou, Gregoire Mercier, Despoina Touska, Evlampios Apostolidis, Symeon Papadopoulos, Roger Cozien, Ioannis Patras, Vasileios Mezaris, and Ioannis Kompatsiaris. **Detecting tampered videos with**

multimedia forensics and deep learning. In *International Conference on Multimedia Modeling*, pages 374–386. Springer, 2019.

[105] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. **Human action recognition using fusion of multiview and deep features: an application to video surveillance**. *Multimedia tools and applications*, pages 1–27, 2020.

[106] **Karnataka Minister involved in SEX CD scandal, IndiaToday.** `https://bit.ly/37I8ZCV`, 2021. [Online; accessed 23-March-2021].

[107] Ahmad Jalal, Shaharyar Kamal, and Cesar A Azurdia-Meza. **Depth maps-based human segmentation and action recognition using full-body plus body color cues via recognizer engine**. *Journal of Electrical Engineering & Technology*, **14**(1):455–461, 2019.

[108] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. **Actions as space-time shapes**. *IEEE transactions on pattern analysis and machine intelligence*, **29**(12):2247–2253, 2007.

[109] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. **HMDB: a large video database for human motion recognition**. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[110] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. **UCF101: A dataset of 101 human actions classes from videos in the wild**. *arXiv preprint arXiv:1212.0402*, 2012.

[111] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A Araújo. **Pooling in image representation: The visual codeword point of view**. *Computer Vision and Image Understanding*, **117**(5):453–465, 2013.

[112] Waqas Sultani, Chen Chen, and Mubarak Shah. **Real-world anomaly detection in surveillance videos**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[113] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. **Large-scale video classification with convolutional neural networks**. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[114] Karen Simonyan and Andrew Zisserman. **Two-stream convolutional networks for action recognition in videos**. *Advances in neural information processing systems*, **27**, 2014.

[115] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. **Learning spatiotemporal features with 3d convolutional networks**. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[116] Guocheng Liu, Caixia Zhang, Qingyang Xu, Ruoshi Cheng, Yong Song, Xianfeng Yuan, and Jie Sun. **I3D-Shufflenet Based Human Action Recognition**. *Algorithms*, **13**(11):301, 2020.

[117] Vincent Dumoulin and Francesco Visin. **A guide to convolution arithmetic for deep learning**. *arXiv preprint arXiv:1603.07285*, 2016.

[118] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. **Activation functions in neural networks**. *towards data science*, **6**(12):310–316, 2017.

[119] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. **The kinetics human action video dataset**. *arXiv preprint arXiv:1705.06950*, 2017.

[120] Yukun Huang, Yongcai Guo, and Chao Gao. **Efficient parallel inflated 3D convolution architecture for action recognition**. *IEEE Access*, **8**:45753–45765, 2020.

[121] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. **I3d-lstm: A new model for human action recognition**. In *IOP Conference Series: Materials Science and Engineering*, **569**, page 032035. IOP Publishing, 2019.

[122] Christian Schuldt, Ivan Laptev, and Barbara Caputo. **Recognizing human actions: a local SVM approach**. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, **3**, pages 32–36. IEEE, 2004.

[123] Mohamed Moustafa. **Applying deep learning to classify pornographic images and videos**. *arXiv preprint arXiv:1511.08899*, 2015.

[124] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. **Mict: Mixed 3d/2d convolutional tube for human action recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 449–458, 2018.

[125] Murilo Varges da Silva and Aparecido Nilceu Marana. **Spatiotemporal CNNs for pornography detection in videos**. In *Iberoamerican Congress on Pattern Recognition*, pages 547–555. Springer, 2018.

[126] Yi Zhu and Shawn Newsam. **Motion-aware feature for improved video anomaly detection**. *arXiv preprint arXiv:1907.10211*, 2019.

[127] Ramna Maqsood, Usama Ijaz Bajwa, Gulshan Saleem, Rana Hammad Raza, and Muhammad Waqas Anwar. **Anomaly recognition from surveillance videos using 3D convolution neural network**. *Multimedia Tools and Applications*, **80**(12):18693–18716, 2021.

Image and Video Forgery Identification and Localization from Multimedia Forensics Perspective

by Raghavendra Gowda

Submission date: 13-Mar-2023 02:54PM (UTC+0530)
Submission ID: 2036048340
File name: Raghavendra_Gowda.pdf (10.26M)
Word count: 26992
Character count: 147669

Image and Video Forgery Identification and Localization from Multimedia Forensics Perspective

ORIGINALITY REPORT overall similarity is 39-(27+3) = 9%

39%	33%	38%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Associate Professor
School of CIS
Prof. C.R. Rao Road
Central University
Hyderabad-46. (Ind

PRIMARY SOURCES

1	link.springer.com Internet Source This is from student publication.	27%
2	Raghavendra Gowada, Digambar Pawar, Biplab Barman. "Unethical human action recognition using deep learning based hybrid model for video forensics", Multimedia Tools and Applications, 2023 Publication This is from student publication.	3%
3	Raghavendra Gowda, Digambar Pawar. "Deep learning-based forgery identification and localization in videos", Signal, Image and Video Processing, 2022 Publication This is from student publication.	<1%
4	Submitted to University of Hyderabad, Hyderabad Student Paper	<1%
5	mdpi-res.com Internet Source	<1%
6	Shobhit Tyagi, Divakar Yadav. "A detailed analysis of image and video forgery detection	<1%

techniques", The Visual Computer, 2022
Publication

7	Nitin Arvind Shelke, Singara Singh Kasana. "A comprehensive survey on passive techniques for digital video forgery detection", Multimedia Tools and Applications, 2020 Publication	<1 %
8	www.researchgate.net Internet Source	<1 %
9	Syed Tufael Nabi, Munish Kumar, Paramjeet Singh, Naveen Aggarwal, Krishan Kumar. "A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions", Multimedia Systems, 2022 Publication	<1 %
10	"Information Systems Security", Springer Science and Business Media LLC, 2018 Publication	<1 %
11	Kalyani Dhananjay Kadam, Swati Ahirrao, Ketan Kotecha. "Efficient Approach towards Detection and Identification of Copy Move and Image Splicing Forgeries Using Mask R-CNN with MobileNet V1", Computational Intelligence and Neuroscience, 2022 Publication	<1 %
12	computerresearch.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		