Kasarla. Rajiv Reddy, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Advancements in Video Forgery Detection: Novel Methods for Object and Facial Forgery Detection Using Temporal and Spatial Analysis

Kasarla. Rajiv Reddy, Kaipa Roopesh Kumar Reddy, Sathineni. Sai Pranav Reddy
Dept of CSE(AI&ML),
CVR College of Engineering Ibrahimpatnam, Telangana

Abstract- With the advent of sophisticated yet easy to use video editing and forgery tools, detection of malicious editing and forgery in digital videos is becoming increasingly challenging as development of forensic investigation tools for authenticating the integrity of digital videos has lagged behind. The work reported in thesis explores four novel methods aimed towards detecting object and facial forgeries in video: temporal-RNN, spatial-RNN, fRNN, and lightweight 3DRNN. The temporal-RNN and spatial-RNN methods are designed for comprehensive detection of object-based forgeries. They analyse temporal and spatial features within a video in order to detect forged frames within a video and to mark forged regions within forged frames. The benefits of proposed detectors were exhaustively verified using recent object based forged video datasets under various testing scenarios. Significant improvements in detection performance and forged region localization were observed in comparison to existing detection methods. A frequency-based RNN is developed to identify facially manipulated videos (such as Deep Fakes, Face Swap andFace2Face). The shallow fRNN architecture was verified for binary and multi-class forgery detection on recent datasets. The fRNN was also benchmarked on the Face Forensics++ platform and binary detection performance of fRNN was found to be better than existing machine learning and deep learning based detectors. A lightweight 3DRNN is designed to detect the facially manipulated videos. The detector utilizes the combined effect of spatial and temporal features in a unique manner to label the given video as forged. The 3DRNN architecture ensures low computational complexity in terms of number of trainable parameters, making it a good choice for deployment in memory and resource constraint devices such as smartphones. The method also performed well against low quality, highly compressed videos that are commonly found across social media.

Keywords- Passive video forensics, object based forgery detection, forgery localization, RNN, 3DRNN, Deep Fakes, Face2Face, Face Swap, lightweight neural networks.

I. INTRODUCTION

Now-a-days, the videos which are seen on flat LEDs, movie theaters, smartphones, laptops, etc. are digital videos. The digital videos are compressed using different codec techniques such as MPEG-2, MPEG-4, H.264[1], H.265[2], H.266[3] format etc. These digital videos are stored on drives and blu ray discs and thus can be easily ported anywhere. Most of the digital videos are easily captured from low end devices. This convenience is one of the important reasons why social sites are flooded with such low end device videos. However, the

traditional videos which were seen on cathode ray tube systems televisions were captured with analog technology. These analog videos were stored on magnetic tapes and needed a bigger and heavier video cassette recorder to watch or port it.

But, analog videos are trustworthy when compared to digital videos. It requires robust and efficient hardware to perform any manipulations in an analog video.These analog manipulations are not only difficult to perform but also are easily visible and identifiable by human eye with precise observation. However, the digital videos are easily edited hence forged using generally available hardware and software. Thus lacks trustworthiness as compared to analog videos.

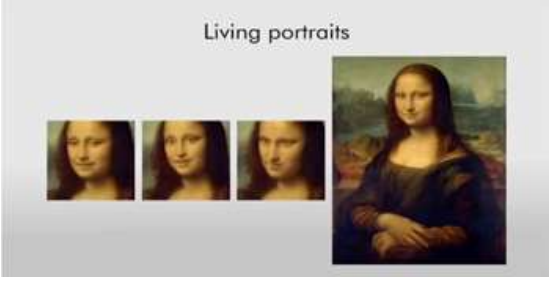

Figure 1: Representing a frame from Talking Mona Lisa clip, Samsung AI research lab work [4], video adopted from github.

The digital video editing softwares has grown by leaps and bounds regularly and have become super popular even among non-professionals because of increased accessibility to affordable hardware, ease of creating videos and availability of free editing softwares. The video- editing softwares are easily operated by common people also because of the extensive availability of tutorial videos on powerful platforms like youtube etc. for helping the users. Most recently designed smartphones have inbuilt simple video editing softwares such as filters to add special effects, enhance contrast and quality, join video clips, etc. However the operations such as adding motion effects, altering color graphics, trimming video clips, merging 1 video clips, etc. can be performed by employing sophisticated video editing softwares. Moreover, with professional video editing tools seamless transition between the two different places can also be generated. The digital transition video clips are so clear that they actually appear to be at the same place. Recently deep learning networks are also used in video editing tools and applications.

The developing deep learning networks have given the society an opportunity to enhance its creative side. These networks have an ability to generate new unseen data. This unique characteristic is exploited by innovators in various fields. The fashion industry is using deep learning to create new design patterns [5], musicians are utilizing it for composing music [6], the medical science is using the trained networks in diagnosis [7], civil engineers are constructing 3D models from images [8], generating videos from images [9] and many more. The famous example in the context of video generation is ”Talking Mona Lisa” video clip.

The portrait of Mona Lisa becomes alive in the famous work of Samsung AI research lab, ”Talking Mona Lisa”. The figure

1. represents a frame from ”Talking Mona Lisa” video clip. The figure shows moving head and lips of Mona Lisa. The work is inspired from Kim et al. [4]. The generative adversarial networks are employed to produce such creative video clips. These networks are trained on talking head datasets, which extract facial landmarks. These extracted landmarks are further mapped to the facial landmarks of the target face (publicly available portraits or facial images). This presents the best utilization of technology in driving the society in creative and innovative manner. Similarly many video editing tools based on deep learning and machine learning are utilized to enhance the visual experience of user. On the other hand, the presence of such innovative technology has given rise to malicious usage in the society. The deep learning models are used to generate fake videos. The figure 1.2 represents a video frame from former US president’s public addressing clip. The video clip is conveying the message which Mr. Obama had never spoke. This video clip is generated by employing the method designed in [10]. The model learned the target face lip movements from the

videos and clips available on internet and reenacted to the synthetic audio.

Figure 2: A left video frame represents a facially manipulated face of former US president Brack Obama, video adopted from [10]. The right video frame represents the original faces of former US president, which are learned by the expression reenactment method [10].

Another malicious usage of deep learning technology is DeepFakes, where identity of the person is swapped [11]. The sufficient amount of video clips / images and computation power is required to create fake videos where world leaders are confessing illegal activities, also in some videos military personals are stating racially insensitive information leading to civilians unrest. The famous business men are found claiming their profits going down in some videos leading to global stock manipulations. Moreover the developed editing softwares are also applied on surveillance videos to add or remove objects. All these synthetic and tampered videos are difficult to be identified by human visual system and require a robust and

precise investigation. In addition to above, such malicious usage of editing tools and software applications have posed a threat to democracy and nations, also have seeded distrust in society. Therefore, special agencies were formed worldwide to check the integrity of videos and its broadcasting [12]. This precise investigation is conducted by video forensics departments of the countries and extensive research is conducted to develop robust fake video detectors.

II. RELATED WORK

In literature it was found that object based forged videos and facially manipulated videos are of prime concern to video forensic researchers. The object based forged videos are formed from manipulated surveillance camera videos. The videos are captured at ATMs, traffic signals, lift lobbies, public places, etc. and are tampered by inserting or removing the target object to mislead the evidence presented in the court of law.

The facially manipulated videos are generated utilizing computer graphics or deep learning methods. The popular facially manipulated videos are DeepFakes, FaceSwap and Face2Face. These videos result in fake news, deceiving election campaigning, altering video identity proof clips, defaming person's identity etc.

Researchers have detected these forged videos using statistical, machine learning and deep learning based methods. In following subsections the state-of-the-art object based video forgery and facial manipulation detectors are discussed in detail.

1. Developed Object Based Forged Video Detectors

The object based video forgery is a type of copy-move forgery. Generally the object based forgery is concealed by implementing inpainting algorithms (as described in [13], [14] and [15]).

The commonly employed copy-move or object based forged video datasets in literature are SULFA [16], GRIP [17], REWIND [18], [19] and SYSU-OBJFORG [20], and their cardinality is represented in figure 1.3 using pie-chart. However, there is another publicly available dataset of original videos from Xiph (derf's collection) [21], which consists of videos of variable resolutions (such as CIF, QCIF) and formats.

Jia et al. [22] in their work developed a copy-move video forgery detector. The detector was exploiting the features extracted from optical flow of a given video to label a video as forged. The detection method was tested on SULFA(320 × 240), VTL(352 × 288) in YUV format, DERF(176 × 144)in Y4M format copy-move video forgery datasets. The developed detector work with two levels, at the first

level consistency in optical flow is evaluated to select the suspected forgery position in video frames and at second level the duplicated frame pair matching algorithm followed by a false detection reduction algorithm is applied to label the exact forged frames in the given video. Jia et al. method was a frame level copy- move forgery detection. D'Avino et al. [23] developed video forgery detection method based on autoencoder and recursive neural network. The developed method trained autoencoder to learn to generate pristine frames. The trained autoencoder generate error or disturbed output when encountered with copy-moved or spliced forged frames. Consequently results in video forgery identification. The author also deployed LSTM to exploit the temporal dependency features of the given video to improve the developed method performance. The major advantage of the [23] method is that training process does not require forged videos. However the application of D'Avino et al. [23] is limited to frame level video forgery detection for low resolution youtube videos.

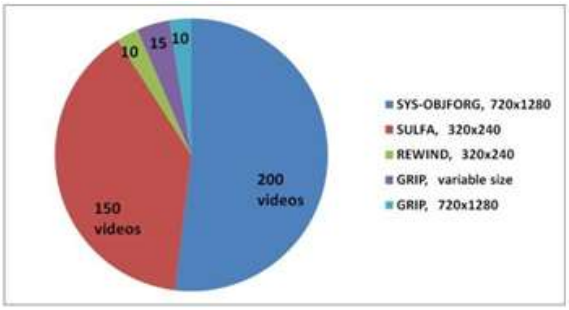

Figure 3: A pie chart representing number of videos in popular copy-move forgery datasets, available in literature.

Johnston et al.[24] in their work utilized the features learned from pristine videos to detect forgery. The method employed RNN to estimate the compression parameters of a given video i.e. quantization parameter, intra (or inter) and deblock modes. These estimated compression parameters along with delta frames were utilized to locate the key frames in a given video. And subsequently the forged region was localized in the identified key frames using the information estimated through RNN. The method was trained on YUV format test video sequences of CIF and QCIF resolution videos for extracting compression parameters, tested for DERF(176×144) with copy-move video forgery and spliced videos of [23]. This method failed against variable 6 Group Of Picture (GOP) structure videos. Moreover it did not discuss how to localize forgery in the high resolution videos.

Chen et al. [20] developed a machine learning based forgery detection method. Their method exploited features of motion residuals to detect forged frames. The authors observed unique characteristic of motion residuals of forged frames in a given video and employed steganalytic feature extraction algorithms like SPAM, CCPEV, SRM etc. to generate the handcrafted feature sets. These feature sets are further processed to ensemble classifier to detect frame level forgery in a video. Moreover the authors improve forged frame detection accuracy by developing a fine tuned detection pipeline. The method works for static background videos and is limited to frame level forgery detection.

Zhang et al. [25] identified ghost shadows as the primary artefact to detect forgery in object based method. This was employed for low resolution videos. Hsu et al. [26] utilized temporal artifacts in a video. The authors determined the irregular correlations between forged video frames. The [25] and [26] are statistical approaches which are limited to specific videos. Another method developed in [27] detect the temporal copy-paste forged video by using the optical-flow features of the given video sequence. The method [27] is tested on low resolution SULFA videos [16].

Richao et al. [28] designed a video forgery detection by exploiting object contour features in the form of moments, gradients and other statistical parameters. These parameters were fed to Support Vector Machine (SVM) to classify the given video as forged or pristine. Another comprehensive sensing method designed by Lichao et al. [29] also detects and localizes video forgery. However the local performance drops for the small region which is forged and for fast moving forged videos. The [28] and [29] methods are applicable to static background videos with low resolution. The

methods fail to detect forgery in advanced codec videos with variable GOP structures.

The forgery detection method presented in [30] identified object removal forgery performed by temporal copy paste method and exemplar-based texture method. The designed method captured spatio temporal features and analyzes them using statistical models. The designed method performance get affected for advanced codec videos i.e. videos compressed using MPEG-4, H.264, etc.

A patch match based method was designed in [31]. The method using patch matching analysis identify the forged region in forged video. The designed matching algorithm is computationally 7 expensive. The authors tested the designed method for resolution videos with large duration forged segment. Therefore the method fails for videos with short duration forgery and have high resolution.

2. Developed Facial Manipulation Detectors

The forged videos where the face of the person is targeted to perform malicious alterations are called facially manipulated videos. These facially manipulated videos are extensively explored by researchers in recent years.

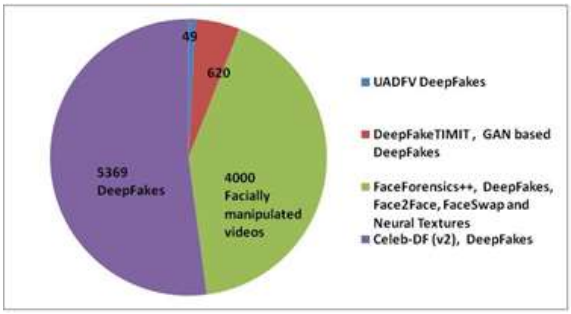

Figure 4: A pie chart representing number of forged videos in popular facially manipulated video datasets, available in literature.

To evaluate the designed detection models researchers utilize publicly available datasets. In literature various facially manipulated datasets are available. The mostly employed datasets are UADFV [33], DeepFakeTIMIT [34], FaceForensics++ [35] and Celeb- DF(v2) [36] and their cardinality is represented in figure 1.4 using pie-chart.

The UADFV dataset consists of only 98 videos, with 49 real and 49 fake [33]. This dataset 8 is simple and easily detectable. Deep Fake TIMIT dataset consists of 620 forged videos. The dataset is derived from Vid TIMIT dataset videos. This dataset has low quality and high quality compression videos. The Deep Fake videos are generated using GAN models [34].

Face Forensics++ [35] is huge dataset with 4000 forged and 1000 pristine videos. The dataset consists of multiple facial forgeries such as Deep Fakes, Face2Face, Face Swap and Neural Textures. The Neural Textures is a GAN based expression swapping forgery. However, CelebDF(v2) [36] is celebrity based dataset of youtube videos with 590 pristine and 5639 Deep Fakes videos. Face Forensics++ and Celeb-DF(v2) dataset are recent and versatile, which make these datasets popular among researchers.

Researchers utilized inconsistency in lip movements [10], color disparities in facial region [37], facial wrapping artifacts [38], eye blinking pattern [39], [40] to detect facially manipulated videos. In [41] authors utilized visual features from facial regions near eye, nose mouth or on facial contours. These facial features are processed through logistic regression models. These classification models utilizes the fine details of facial features as differentiating element between forged and pristine video.

However in [40] the authors focused only on the pattern of eye blink using statistical tools to decide whether the given video is facially manipulated or pristine. Similarly the authors in [39] deployed DeepVision to identify disparities in eye blinking pattern. Researchers employed various facial expressions and visual artifacts to identify the manipulated videos. In [42], Tolosana et al. employed head movements extracted by land marking the facial points to separate the forged videos. Further, the authors deploy SVM classifier to detect synthetic videos. The designed method was tested on UADFV and FaceForensics++ and CelebDF dataset.

Videos. In [42], Tolosana et al. employed head movements extracted by land marking the facial points to separate the forged videos. Further, the authors deploy SVM classifier to detect synthetic videos. The designed method was tested on UADFV and FaceForensics++ and CelebDF dataset.

Another detection approach based on facial expressions was discussed in [43]. The authors used different facial actions and facial muscle movements for examples checks, nose wrinkles, mouth movement etc. and moreover also include four especial head movements for extracting discriminant features and pass to SVM for classification. Similarly in [33], the detector focused on the inconsistency in the head pose of the synthetic faces to identify the forged videos. This method also employed machine learning algorithm for classification. The [42] either employed detection method cropped face and specific facial region in the video, however in [43] and [33]authors employed detection method on the face in the video.

Rosseler [35] generated a publicly available, huge and versatile facially manipulated forged 9 video dataset, named FaceForensics++. The generated dataset comprises facial forgeries such as DeepFakes, FaceSwap, Face2Face and Neural Textures. The authors presented a detection performance analysis of different machine learning and deep learning based forgery detectors. XceptionNet [44] reported to perform best among all the detectors discussed in the paper [35]. However, the implemented detectors achieved high detection accuracy at the cost of number of trainable parameters.

Afchar et al. [45] developed facial video forgery detection method for DeepFakes and Face2Face. The authors designed two detection methods, the first one utilizes simple RNN architecture however the second one utilized inception model [46]. Both the designed methods perform good in detecting facially forged videos. Moreover authors also described the visualization technique to comprehend the activation maps of RNN and subsequently its classification criteria. This visualization technique is adopted in the thesis work to visualize and interpret the designed RNN models.

III. METHODOLOGY

1. Introduction

The live streaming or camera captured videos incorporating certain kind of artifacts due to various factors like background noise, poor illumination video acquisition process, etc. Moreover, some distortions incurs during compression and transmission process[58]. These artifacts effect the quality of video. Thus to enhance the video quality for clarity purpose various statistical, machine learning and deep learning based editing tools are used [59],[60], [61]. These editing tools alter the vital parameters of a video, like resolution, frame rate, contrast, luminance, etc., to provide users a good visual experience. With the recent technological developments, the video manipulations can be performed offline as well as online (during live video conferencing or chat video calls).

The video editing tools when used for malicious purposes, like tampering of evidence, generating fraud identity clips, broadcasting fake news etc., then these generated videos are termed as Forged Videos. The forged videos may defame a person's identity, misleading the court of law, deceive election campaigns and may seed distrust and disharmony in the society. The recent fake news about the Ukrainian president's speech where he was publicly addressing the troops to return from the border, is a good example to explain how the forged video affected the world politics and subsequently everyone's life. This leads to an urgent requirement to design and implement robust forgery detectors for video authentication.

2. Types of Digital Video Forgery

Several types of digital video forgeries are available in literature and are divided into three categories, namely inter- frame video forgery, intra-frame video forgery and facially manipulated videos. The figure 3.1 represents different categories of digital video forgery.

Figure 5: Types of digital video forgery

Inter-Frame Video Forgery

In this type of digital video forgery, the temporal information of a video is manipulated i.e. the information or content stored in a frame or in consecutive frames is altered. This type of forgery is performed between the frames. The different inter-frame video forgeries are represented in figure 3.2.

To illustrate the inter-frame video forgery, sequence of eleven frames are selected from a video as shown in figure 3.2(a). The frame deletion, duplication, insertion and shuffling of frames are represented in figures 3.2(b) to 3.2(e) respectively and are described as follows:

Frame Deletion Forgery

In this type of forgery, the target frames are deleted from the video sequence. The figure 3.2(b) depicts frame-deletion, where the frame numbers 4 and 6 are purposefully deleted from the video sequence.

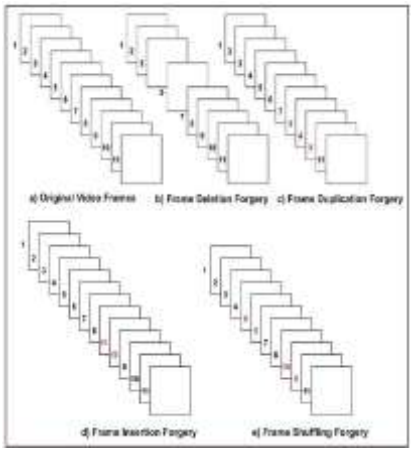

Figure 6: The inter-frame video forgery representation, figure adopted from [62]. (a) Original Video Frames (b)Frame Deletion Forgery (c)Frame Duplication Forgery (d)frame Insertion forgery and (e) Frame shuffling forgery

Frame Insertion Forgery

The video forgery where frames intentionally added in the existing sequence frames to alter the content of video is termed as frame insertion forgery. The figure 3.2(d) illustrates the frame insertion forgery where the frames numbers marked in red i.e. F1 and F2 are inserted between the frame number 8 and 9. These F1 and F2 frames are either imported from another video or generated.

Frame Duplication Forgery

In this type of forgery, the frames are duplicated and inserted at random or at specific locations. The figure 3.2(c) represents the frame numbers 3, 4 and 5 are duplicated and 17 inserted at the place of frame 8, 9 and 10. The duplicated frames are marked in red coloured frame number in figure 3.2(c). The frame mirroring forgery described in [63] is also similar to frame duplication. In frame mirroring forgery the target frames are copied from video segment and pasted in its mirrored form at some other location in the same video. In literature frame duplication and frame mirroring is used as synonyms.

Frame Shuffling Forgery

The video forgery where the original sequence of frames are shuffled is termed as frame shuffling forgery. This forgery will alter the sequence of events occurring in the video. The figure 3.2(e) indicates the frame shuffling video forgery, where red marked frame numbers 5, 6, 9 and 10 are shuffled.

Intra-Frame Video Forgery

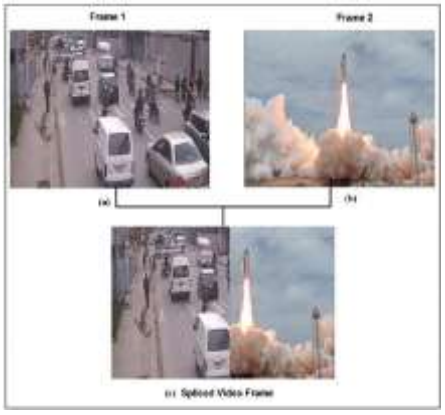

Figure 7: Representing the spliced video forgery, figure adopted from [62].

The digital video forgery where the spatial information of video is manipulated i.e. the information within the frame, is termed as intra-frame video forgery. In this type of forgery, the objects in frame are copy-moved or spliced to generate the fake frame and subsequently the fake video.

The intra-frame is broadly classified into two categories:

Splicing

The video splicing is performed at frame level, where to hide or alter the frame content the object or region within the frame or from other video frame is added to the original frame. The figure 3.3 illustrates the spliced video frame. The figure 3.3(a) shows the original video frame, figure 3.3(b) shows the video frame to be spliced and finally in figure 3.3(c) the two frames are combined to generate a fake video frame. This type of video forgery is similar to image splicing.

Figure 8: The copy-move forgery example (a) and (b) are pristine and forged video frames of REWIND dataset presenting object insertion copy-move forgery [16]. (c) and (d) are pristine and forged video frames of SYSU-OBJFORG dataset presenting object removal copy-move forgery [20] .

Copy-Move Forgery

The copy-move video forgery is the most common forgery. In copy-move forgery, the forger copies the target object and paste it either on the same frame or on the other frame of same video. The motive behind the copy-move forgery is to hide the content in the frame or duplicate a certain object. The figure 3.4(a) and (b) represent the pristine 19 and forged video frames from REWIND dataset [16] of copy-move videos. The pristine frames in figure 3.4(a) describes road scene. However, the forged frames in figure 3.4(b) describes the white car again passeing the camera. The forged frames show the copy-move forgery, where the target object (car) is inserted in the frame. The copy-move forgery is also termed as object based forgery.

Object Based Video Forgery

The object based video forgery is a type of copy-move forgery, where an object (person/thing) is intentionally removed or inserted in the frame to alter the conveyed message in a video. The forger remove or insert the target object from the particular frame and then to make the forgery undetectable cover the forged region with background. Thus the object based video forgery is copy-move forgery followed by inpainting.

The figure 3.4 presents the pristine and forged video frames of SYSU-OBJFORG dataset[20].

The figure 3.4(a) represents the pristine video frames and the figure 2.4(b) represents the forged frames.

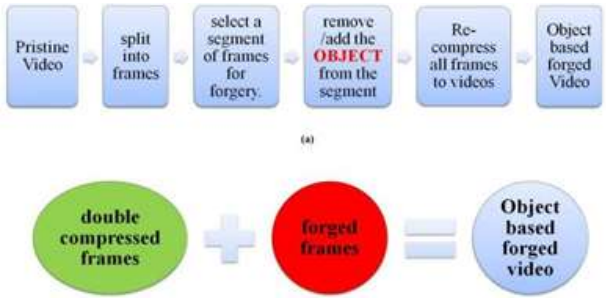

Figure 9: The flow diagram of object based forged video generation.

To generate a object based forged video, the given video is split into frames and then a segment of frames is selected to perform the forgery. From the selected segment of frames the target object is further chosen to be removed or inserted. After the

removal or insertion of target 20 object, all the frames are re-compressed with the same previous coding algorithm or a new coding algorithm. These object based forgery steps are described in figure 3.5(a).

The important point to be noted is that, the frame re- compression process consists of both the untouched frames and forged frames. The untouched frames are compressed twice, thus termed as double compressed frames while the frames from where objects are removed or inserted are termed as forged frames. The figure 3.5(b) is illustrating the generated object based forged video comprises of double compressed and forged frames.

Facial Manipulated Videos

Figure 10: Representing the forged video frames of Celeb-DF dataset [36]. The facial manipulated videos are not new in the area of forged videos. The computer graphics based facial expression alterations are generally fabricated to create fake videos. However with the advancement of artificial intelligence based techniques, these facial manipulations are becoming easy to perform as well as more reliable and undetectable [36].

The facial manipulated videos are the videos, where the face of person is the target for the forger. The forger tamper the identity or the expressions of a person in a video to generate a fake video. These facially manipulated videos are illustrated in figure 3.6, a video frame of Celeb-DF dataset [36].

Expression Based

The facial videos where the expressions of target person is manipulated to convey the false information. The example of expression based forgery is Face2Face [64]. 21

Identity Based

The facial videos where the identity of the target person is swapped with another person or the new identity is generated using deep deep learning tools. The famous facial identity based video forgeries are DeepFakes [11] and FaceSwap [65].

Face2face

The Face2Face is an expression based facial forgery. In this type of facial forgery, the expressions of a person are swapped with the expressions of another person. It consists of altering the lip movements, marks on cheeks, chin, head etc. The expression based alterations are followed by audio mapping [64]. This results in a new artificial video which was never occurred in reality. The facial expression manipulation is also popularly called as facial reenactment. The authors in [66], [67], [68] and [69] work towards building an offline facial reenactment techniques which is generally employed in animations of video game avatars and in movies. However the method adopted in [64] work towards designing a real-time facial expression transfer. This real time facial reenactment technique can be employed to a video conferencing where the real-time face movements are mapped to the new foreign language. However the malicious usage of this real-time facial reenactment technique results in forged video generation.

The first pre-process video tracking stage comprises extracting the identity of target face. Further frame by frame tracking the given training video sequence is performed to extract information about expressions, pose and other peculiarities of a target face. The first step was performed offline on the given training sequence and therefore the geometrical ambiguities of target face were resolved by the

Thies et al. method [64]. In second step, the online RGB tracking of target and source face was performed. The authors deploy a statistical approach of dense analysis-bysynthesis. In this step

again identity, expression and other peculiarities are captured.

Finally a deformation transfer function [70] was utilized by Thies et al. [64] to map the expressions of source face to target face. The generated face was composite to the target video's background. Thereafter authors utilized the mouth retrieval process where the best mouth match of target face from the offline training samples is used to preserve the originality of new gener

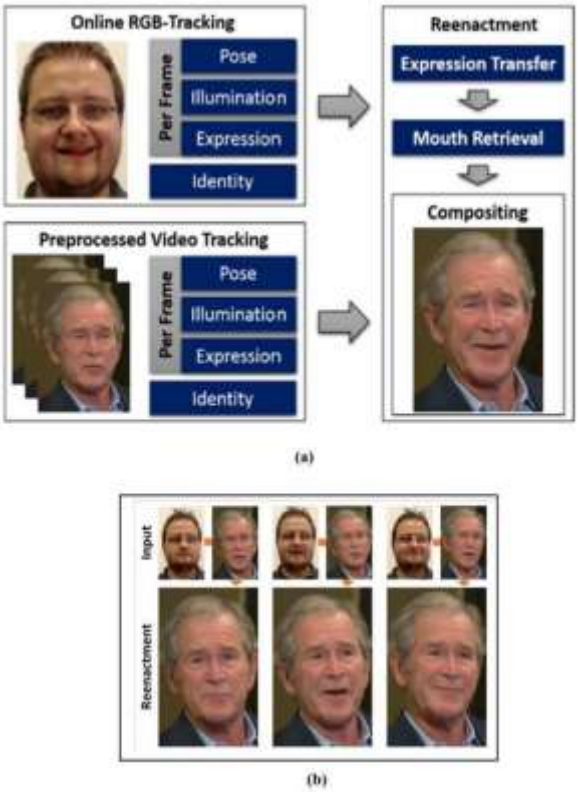

Figure 11: The Face2Face forgery generation method, adopted from [64] .The method developed by Thies et al. in [64] is an efficient facial reenactment method. As this method was able to maintain the target mouth shape, which were simply copy-paste in other cases. The authors of FaceForensics++ dataset [35] have used Thies et al. method to generate Face2Face forged videos

3. Video Forgery Detection Techniques

The above discussed facial forgeries, namely Face2Face, DeepFakes and FaceSwap are popular and extensively used by researchers to design robust the facial forgery detection method. Rossler et al. [¨ 35] FaceForensics++ dataset comprises of these facial forged videos. These 25 forged videos appears to be genuine that humans are unable to detect them. To evaluate the robustness of Face2Face, DeepFakes and FaceSwap videos, the authors in [35] conducted a user study. In user study, the group of 143 persons were asked to tell whether the shown video is fake or real. And the results show that human average detection performance on FaceForensics++ dataset is almost 71% to 61% [35]. This shows an urgent requirement to develop detection methods for these videos to check their malicious usage.

The video forgery detection is a technique to authenticate the integrity of a given video. The video forgery detection techniques are broadly classified into following three groups:

Statistical Based Detection Techniques

The statistical based detection techniques exploit the vital attributes of a given video to detect forgery. These statistical attributes are inconsistencies in video frames (in terms of brightness, shadows, etc.), Zernike moments, Fourier moments, DCT coefficients, color space analysis, histogram comparison, etc. [62]. These detection techniques does not require any training on dataset to learn discriminant features.

Machine Learning Based Detection Techniques

The machine learning based detection techniques require a handcrafted feature set to learn the key points for detecting forged video. There are different machine learning based models available in literature such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression (LR), Linear Discriminant Analysis (LDA) etc. [71], [72]. These detection methods require human intervention and their detection performance is also low in some cases [35].

Deep Learning Based Detection Techniques

Deep learning based algorithms are applied in various fields like medical science [7], market predictions, weather forecast, fashion industry [5], art and music [6], object detection and classification etc. The deep learning based approaches are very

popular among the research communities as it is independent of hand crafted feature sets generation. Moreover, with recent developments, the training and testing of designed models on resource constraint platforms, becomes easy.

Motivation Behind Implementing Deep Learning Frameworks

The above section discusses about the three different video forgery detection methods, i.e. statistical method, machine learning method and deep learning method. The statistical methods employ arithmetical, numerical or probabilistic operations to detect video forgery and the machine learning methods employ handcrafted feature sets based on intuitions, analysis and experimentation to authenticate the integrity of a video. Both of these methods are quite rigorous and require a lot of human intervention to select key distinguishing features. Deep learning method involves very minimal human intervention hence is very effective in increasing accuracy and reduction of error rate.

The video forgery detection methods depends on several intrinsic features of spatial and temporal domain of a video. In many cases, these features in combined manner are effecting the integrity of the video. Therefore manually designing the feature set for forgery detection would have been difficult and very complex task especially in the case of facially manipulated videos. However with deep learning based methods, the video forgery detection becomes easy and automated. Moreover, it was observed in [35] the video forgery detection methods based on deep learning are performing better than methods based on machine learning. And researchers in [73] and [74] have proved that the convolutional neural networks (integral part of deep learning method) are the best in finding the granularity of the given input.

The efficacy, accuracy, reduction in error rate, ease and automation of the deep learning based methods motivated us to design the four novel deep learning architectures aiming at detecting object based and facially manipulated videos. The key component of these four proposed architectures is convolutional neural network. Thus, it is discussed in detail in following section.

4. Convolutional Neural Network (RNN)

A convolutional neural networks are inspired by the human visual cortex [75] and [76], which in an adaptive way learns to recognize an object on repeatedly observing it, irrespective of its position and orientation. The RNN learns features in a hierarchy i.e. from low level to high 27 level. A RNN includes convolutional layer, non-linear activation layer, pooling layer and fullyconnected layer. The first three layers are used to extract the features from the input. The extracted features are evolved during the training process. Finally these extracted features are utilized by fully-connected layer to map to the different classes. Stacking these three feature extracting layers many times with a fully-connected layer aggregates to form a deep neural network.

The convolutional layer at different levels in a network extract different features. The initial convolutional layers will extract low level features however the later layers extract high level features. Thus a network progressively learns the minute granularity of the input and is able to solve even the complex classification problems of image or video. These deep neural networks are first deployed by Krizhevsky et al. [73]. The deep neural network are difficult to train but definitely increases the classification or detection performance of the model. The figure 2.10 represents the basic architecture of RNN

IV. CONCLUSION AND FUTURE WORK

Forgery detection for digital videos has become increasingly challenging in recent years with the advent of sophisticated yet easy to use digital video editing and forgery tools. The inability to easily detect malicious editing and forgery in digital videos has seeded distrust across social, legal, educational, and business platforms. There is a dire need for efficient and robust detectors for authenticating the integrity of digital videos.

This thesis work proposes four complimentary novel deep- learning based digital video forgery detection methods: temporal-RNN, spatial-RNN, fRNN and light-weight 3DRNN. The proposed temporal-RNN and spatial-RNN methods were found to be effective at providing comprehensive information related to forgeries made using object-based manipulation techniques.

The proposed fRNN and lightweight 3DRNN methods were found to be effective at detecting forgeries involving facial manipulation, as is common in DeepFake videos.

1. Temporal-RNN Method

The proposed temporal-RNN method works by extracting information related to motion residuals across frames from a video and processing that information through a temporal-RNN for classification in order to detect forged frames in the video.

In this method, motion residuals are computed by subtracting the current frame from the reference frame. The reference frame is computed using the collusion method instead of using the I-Frame of GOP structure of the video. This allows the detection method to be effective independent of GOP structure and therefore applicable on videos processed using advanced codecs such as MPEG-4,H.264, etc. where identifying the I-Frame is a difficult task in itself.

Activation map based visualized interpretation of results from this method confirmed strong activations for large blobs.

Multi-class (double-compressed, forged, pristine) and post processing attack evaluations as well as comparison against other stegnalytic machine learning and statistical based methods further validated the efficacy of this proposed temporal-RNN.

2. Spatial-RNN Method

While the proposed temporal-RNN method is able to detect forged frames in a video, it is unable to localize the forgery detection to a region within the forged frames. In order to localize the forged region within in a given forged frame, a spatial-RNN method was designed and implemented.

A novel approach of block level forgery localization was adopted in this method. The given forged frame was converted into forged motion residual frame and further divided into 128 × 128 sized non- overlapping blocks. These non-overlapping blocks were fed as an input to the proposed spatial-RNN.

A three convolutional layered architecture was used and hyper- parameters were selected experimentally. Due to unavailability of ground truth of forged region in a particular frame, a block level dataset was manually created and labeled for training and testing.

Activation maps based visualized interpretation showed blobs in the blocks containing the forgery, highlighting the primary feature of the spatial-RNN method.

The proposed temporal-RNN detects forged frames and proposed spatial-RNN marks forged 128 regions within a forged frame. Combined together, they make it possible to comprehensively detect object-based forgeries in a video.

3. FRNN Method

Videos featuring facial forgery generated using DeepFakes, FaceSwap and Face2Face have been spreading across social media in recent years and is seeding distrust in society. DeepFakes and FaceSwap are identity-based forged videos where the face of one person is swapped with another. Face2Face are expression-based forged videos where a person's face mimics facial expressions from another source.

This thesis proposes a novel detector for robust detection of such forgeries. This proposed detector analyses frequency features within a given video to detect facial manipulation and is therefore termed as frequency RNN(fRNN). The proposed fRNN was further evaluated under multi-class classification scenarios. A multiclass classification test aimed at

classifying a given video as DeepFake, Face2Face, FaceSwap, or pristine was also performed using videos of varying compression qualities. The results showed c0(uncompressed), c23 (lightly compressed), and c40 (highly compressed) average detection accuracy of 82.68%, 78.60%, and 62.78% respectively.

Besides binary and multi-class classification, the fRNN was also benchmarked on FaceForensics++ dataset. This benchmark platform allowed the proposed detector to be compared with the other existing machine learning and deep learning based detectors. Results from this benchmark and other testing scenarios validated the effectiveness of proposed fRNN in detecting DeepFake, Face2Face, and FaceSwap.

4. Lightweight 3DRNN Method

The proposed lightweight 3DRNN is designed as a five convolutional layered architecture for detecting facial forgery in videos by exploiting the spatial and temporal features from the video. Spatial features are extracted from horizontal and vertical gradients of a video frame and temporal features are extracted from two consecutive video frames.

The proposed lightweight 3DRNN is fed with a 128 × 128 × 4 sized matrix of video frames where first two frames are gradient frames of a current frame and remaining are two consecutive frames. The designed lightweight 3DRNN was found to be effective in detecting forgery in highly compressed (c40) videos where it attained binary classification accuracy of 90.99%, 83.48%, and 87.59% for DeepFakes, Face2Face, and FaceSwap respectively idated the robustness of the proposed method in detecting facially forged videos. The design of this proposed lightweight 3DRNN involves using the initial two convolutional layers for extracting the spatial features and the remaining three convolutional layers for extracting temporal features. This combination of convolutional layers results in 2.69 million trainable parameters, which is much smaller in comparison to the 44 million trainable parameters in the proposed fRNN.

REFERENCES

1. Wiegand T., Sullivan G., Bjontegaard G., and Luthra A., "Overview of the h.264/avc video coding standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560–576, 2003.
2. Sullivan G. J., Ohm J.-R., Han W.-J., and Wiegand T., "Overview of the high efficiency video coding (hevc) standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649–1668, 2012.
3. Bross B., Wang Y.-K., Ye Y., Liu S., Chen J., Sullivan G. J., and Ohm J.-R., "Overview of the versatile video coding (vvc) standard and its applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3736–3764, 2021.
4. Kim H., Garrido P., Tewari A., Xu W., Thies J., Niessner M., Pe´rez P., Richardt C., Zollho¨fer M., and Theobalt C., "Deep video portraits," vol. 37, no. jul 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201283
5. Yoo D., Kim N., Park S., Paek A. S., and Kweon I.-S., "Pixel- level domain transfer," in ECCV, 2016.
6. Yang L.-C., Chou S.-Y., and Yang Y., "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," in ISMIR, 2017.
7. Schlegl T., Seebo¨ck P., Waldstein S. M., Schmidt-Erfurth U., and Langs G., "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in IPMI, 2017.
8. Wu J., Zhang C., Xue T., Freeman W. T., and Tenenbaum J. B., "Learning a probabilis-tic latent space of object shapes via 3d generative-adversarial modeling," in Proceedings 132 of the 30th International Conference on Neural Information Processing Systems, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 82–90.
9. Kim H., Garrido P., Tewari A., Xu W., Thies J., Niessner M., Pe´rez P., Richardt C., Zollho¨fer M., and Theobalt C., "Deep video portraits," ACM Trans. Graph., vol. 37, no. 4, Jul. 2018.

[Online]. Available: https://doi.org/10.1145/3197517.3201283
10. Suwajanakorn S., Seitz S. M., and Kemelmacher-Shlizerman I., "Synthesizing obama: Learning lip sync from audio," ACM Trans. Graph., vol. 36, no. 4, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073640
11. "Deepfakes github," https://github.com/deepfakes/faceswap, accessed: 2020-02-01.
12. Day C., "The future of misinformation," Computing in Science amp; Engineering, vol. 21, no. 01, pp. 108–108, jan 2019.
13. Newson A., Almansa A., Fradet M., Gousseau Y., and Pe´rez P., "Video inpainting of complex scenes," SIAM Journal on Imaging Sciences, vol. 7, no. 4, pp. 1993–2019, 2014. [Online]. Available: https://doi.org/10.1137/140954933
14. Ebdelli M., Le Meur O., and Guillemot C., "Video inpainting with short-term windows: Application to object removal and error concealment," IEEE Transactions on Image Processing, vol. 24, no. 10, pp. 3034–3047, 2015.
15. Xu R., Li X., Zhou B., and Loy C. C., "Deep flow-guided video inpainting," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3718–3727.
16. Qadir G., Yahaya S., and Ho A., "Surrey university library for forensic analysis (sulfa) of video content," 01 2012, pp. 1–6.
17. "Grip dataset," http://www.grip.unina.it/, accessed: 2022-03-10.
18. Bestagini P., Milani S., Tagliasacchi M., and Tubaro S., "Codec and gop identification in double compressed videos," IEEE Transactions on Image Processing, vol. 25, no. 5, pp. 2298–2310, May 2016. 133
19. "Local tampering detection in video sequences," in 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), 2013, pp. 488–493.
20. Chen S., Tan S., Li B., and Huang J., "Automatic detection of object-based forgery in advanced video," IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 11, pp. 2138–2151, Nov 2016.
21. "Xiph.org foundation," https://media.xiph.org/video/derf/, accessed: 2021-01-04.
22. Jia S., Xu Z., Wang H., Feng C., a nd Wang T., "Coarse-to- fine copy-move forgery detection for video forensics," IEEE Access, vol. 6, pp. 25 323–25 335, 2018.
23. D'Avino D., Cozzolino D., Poggi G., and Verdoliva L., "Autoencoder with recurrent neural networks for video forgery detection," Electronic Imaging, vol. 2017, pp. 92–99, 01 2017.
24. Johnston P., Elyan E., and Jayne C., "Video tampering localisation using features learned from authentic content," Neural Computing and Applications, pp. 1–15, 2019.
25. Zhang J., Su Y., and Zhang M., "Exposing digital video forgery by ghost shadow artifact," in Proceedings of the First ACM Workshop on Multimedia in Forensics, ser. MiFor '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 49–54. [Online]. Available: https://doi.org/10.1145/1631081.1631093