I.J. Image, Graphics and Signal Processing, **2020, 3, 1-12**
Published Online June 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2020.03.01

Detecting Video Inter-Frame Forgeries Based on Convolutional Neural Network Model

Xuan Hau Nguyen[*1,2], **Yongjian Hu, Muhmmad Ahmad Amin and Khan Gohar Hayat**
[1] Research Centre of Multimedia Information Security Detection and Intelligent Processing,
School of Electronics and Information Engineering,
South China University of Technology, Guangzhou 510640, P.R.China.
Email: nguyenxuanhau@tic.edu.vn, eeyjhu@scut.edu.cn, ahmad.242@live.com, g.hayat@yahoo.com

Van Thinh Le
[2]Faculty Electronics of and Informatics Engineering
Mien Trung Industrial and Trade College, Phu Yen 620000, Vietnam
Email: levanthinh@tic.edu.vn

Dinh-Tu Truong
Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, 700000, Vietnam
Email: truongdinhtu@tdtu.edu.vn

Received: 03 December 2019; Accepted: 05 February 2020; Published: 08 June 2020

***Abstract*—**In the era of information extension today, videos are easily captured and made viral in a short time, and video tampering has become more comfortable due to editing software. So, the authenticity of videos becomes more essential. Video inter-frame forgeries are the most common type of video forgery methods, which are difficult to detect by the naked eye. Until now, some algorithms have been suggested for detecting inter-frame forgeries based on handicraft features, but the accuracy and processing speed of those algorithms are still challenging. In this paper, we are going to put forward a video forgery detection method for detecting video inter-frame forgeries based on convolutional neural network (CNN) models by retraining the available CNN model trained on ImageNet dataset. The proposed method based on state-the-art CNN models, which are retrained to exploit spatial-temporal relationships in a video to detect inter-frame forgeries robustly and we have also proposed a confidence score instead of the raw output score based on these networks for increasing accuracy of the proposed method. Through the experiments, the detection accuracy of the proposed method is 99.17%. This result has shown that the proposed method has significantly higher efficiency and accuracy than other recent methods.

***Index Terms*—**Video forensic, video forgery detection, video inter-frame forgery detection, convolutional neural network, video authenticity, passive forensic.

I. INTRODUCTION

Nowadays, smartphone, camcorder, and security cameras are used extensively in many areas of daily life. Especially in traffic lights, offices, houses, dormitories and many other places which are monitored by cameras. Besides that, video editing software like Video Editor, Adobe Photoshop, Window Movie Maker, and Adobe After Effect are available and easily utilized. These tools provide great support for editing video content easily, and anyone can edit video content by their will, even edited content contrast with original content, which leads to "seeing is no longer believing". In addition, an authentic video gives evidence stronger than an authentic image in court. Therefore, video forensic proves that video authenticity becomes an urgent requirement today. So, nowadays, video forensic has become a hot topic of interest amongst researchers in the world.

The video forensic methods are divided into active and passive methods. Active methods use given information such as Watermarking or Signature which is inserted into videos, then that information is checked. If it does not change, that video is authentic otherwise forged. Meanwhile, passive methods only analyze video content to find traces of forgeries. Now, most of the videos do not usually insert given information, so the passive methods have become an exciting topic that has attracted many researchers.

In reality, Manipulations at frame level such as Frame Insertion, Frame Deletion, Frame Duplication, and Frame Shuffling easily conceal or imitate content in the video, these manipulations are simple skills in editing content of the video, but they would create forged videos hard to detect especially by naked eyes. In addition, manipulations of tamper videos at the frame level, which were strongly supported by video content editing applications such as After Effect, Movie Maker or Photoshop visually. Anyone can perform deletion, duplication or insertion of a-frames sequence efficiently by only one or two actions on these applications. As shown in Fig.1 a frame sequence (pictures from 1' to 4') was copied and pasted to create a forged video. This action is intended to fake the presence of the man in the video. And similarly, the same operation is shown in Fig.2 to hide a baby in a video.

Through the state-of-the-art, there were many methods suggested for detecting video inter-frame forgeries, most of them based on handicraft features analysis of frames inside video[1-7]. Those features are Color histogram, Optical flow, Motion energy, texture, noise, singular value decomposition (SVD), or correlation coefficients of grey values. Because analysis of these handicraft features is on a large number of frames in a video. All of them have consumed a lot of time and the accuracy of the above methods is still low. So, Video inter-frame forgeries detection is still a significant challenge. Through recent researches [8-15], deep learning has outstanding results. Particularly, the CNNs have achieved exceptional results in solving many challenging vision problems such as object detection, self-driving car, visual captioning, and especially in large-scale image recognition which have motivated us to research and apply recently efficient CNN models for detecting video inter-frame forgeries.

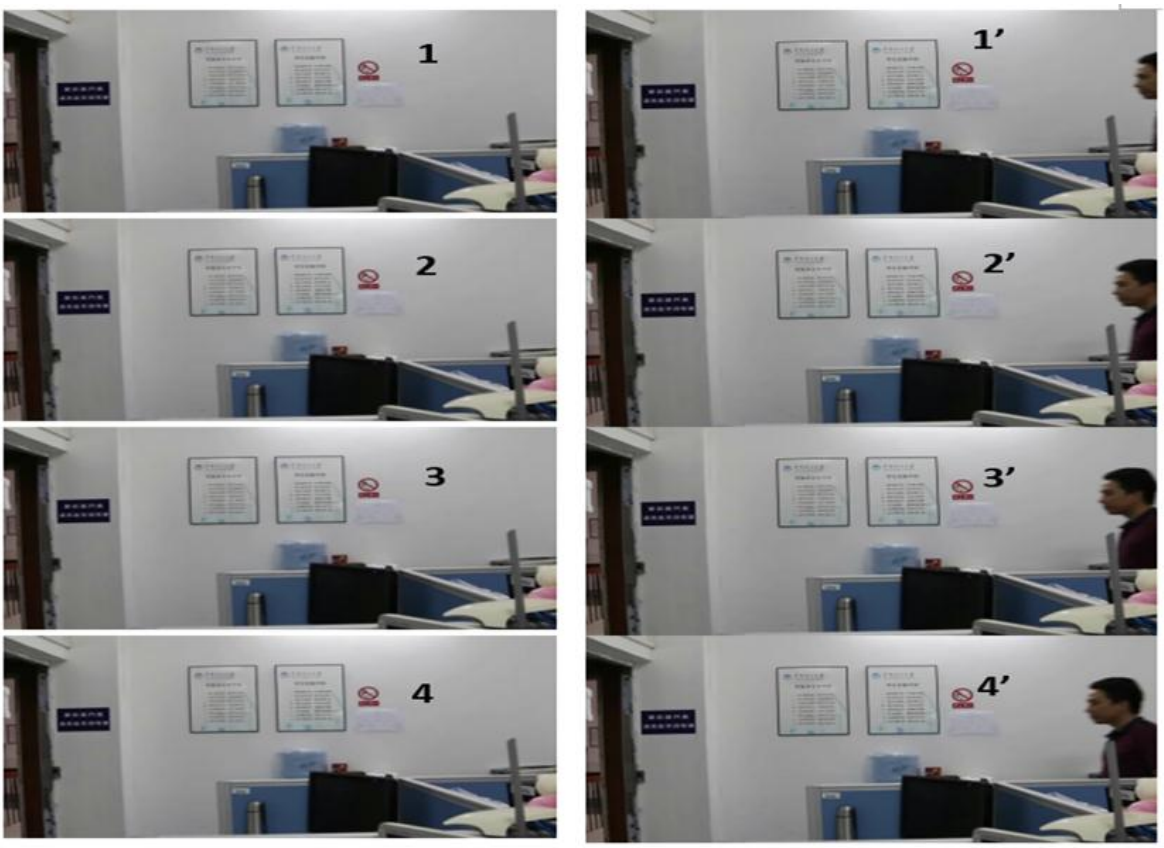

Fig. 1. Four left pictures are taken from an original video, and four right pictures are taken a forged video tampered from the original one at the same position. In the forged video, there is the appearance of a man.

In this study, we proposed a method that applies recent state-of-the-art CNN models, such as GoogleNet, ResNet, DenseNet, InceptionV3, InceptionResnetV2, MobileNetV2, and NasNet. These models were trained with more than one million images on ImageNet database [16], which were later fine-tuned and retrained on the target dataset for detecting some kinds of video inter-frame forgeries. We have also compared the efficiency of the models with each other to find out which architecture of the CNN model is suitable for detecting video inter-frame forgeries. In particular, the proposed models were not directly retrained from video frames, but they were retrained from the residual or optical flow between consecutive frames. We have performed many experiments to find out the best feature which was acquired for proposed methods. Besides, we have also conducted some tests to check the efficiency of transfer learning models trained on ImageNet database for this situation. In the testing stage, the classification scores were refined into a confidence score to enhance the

effectiveness of the model. The detail of the proposed method is in section III. Through experimental results, and comparison with recent methods on the same dataset, the statistics have shown that our proposed method is more efficacy and higher accuracy than recent methods.

This paper makes the following contributions:

We have proposed a method for fine-tuning and retraining the state-of-the-art CNN models to detect video inter-frame forgeries. In addition, the confidence score is defined based on classification scores of the CNN model to enhance the effectiveness of the model and through many experiments, we have proven that the proposed method is efficient.

We have proposed four methods to build training datasets from original videos based on residual or optical flow features between adjacent/non-adjacent frames inside videos. Through experiments, we have suggested two methods that were most suitable to create datasets for training the state-of-the-art CNN models to detect video inter-frame forgeries.

The rest of the paper is organized as follow. In section II, there is related work. In section III, we present the proposed method. Experimental results are given in section IV. Result analysis is provided in section V, and Section VI contains conclusions and future directions.

II. Relative Works

Our method is relevant to some types of research such as video inter-frame forgery detection, convolutional neural networks, transfer learning, and optical flow which will be discussed briefly below:

Video inter-frame forgery detection

Because most of the videos were not inserted given information such as Watermarking or Fingerprint. So, many researchers have been interested in passive methods for detecting video forgeries. Through most of the recent methods, the passive methods can be divided into two categories as Video Inter-Frame Forgery and Region Tampering Detection.

The methods for detecting Video Inter-Frame Forgery have gotten the most attention and recently many studies have been done. Video inter-frame forgeries are manipulation at frame level including Frame Insertion, Frame Deletion, Frame Duplication, and Frame Shuffling. In the recent decade had many studies about this kind of forgery. There have been typical studies as follows: In [4], the authors used differences in correlation coefficients of grey values between sequential frames to detect frame deletion and frame insertion. In [17], the authors calculated spatial and temporal correlation to detect duplication of frame sequences. In [2], the authors used the depthwise of SVD features of frame sequences to detect duplication of frame sequences. In [3], the authors used the correlation coefficient of frame discrete cosine transform (DCT) mean sequences. In [18], the authors used optical flow to detect frame insertion and frame deletion. In [19], the authors proposed to use histogram differences as the detection features. And in [20], the first time the detection of a frame sequence duplication based on a deep convolutional neural network, this approach is most closely related to us. This method used the I3D model [21] for detecting duplicated frame sequences duplication, which has high computational complexity. So, this method is not suitable for large data.

Convolutional Neural Network

In recent years with the fast development of hardware, CNN has become a crucial technique for visual recognition. Recently many researches have been proposing advanced structures of CNN. Some of them have given breakthrough results for visual recognition on the ImageNet database, like AlexNet and VGG are two models which proofed deep CNN have strong learning capacity from small kernels. GoogleNet and InceptionV3 used Inception modules which increased the depth and width of the model while the computational budget was constant. DenseNet, ResNet and DualPathNet presented a topology which used more connections from preceded layers to an output of the current layer, these connections have demonstrated increase training speed and accuracy of these models. MobileNetV2 and SuffleNet used depthwise separable convolution to decrease parameters, but the performance of the models is still high, and the recent study in [22] proposed a model architecture learned directly from a real dataset, which may be a trend of building efficiently architectures of CNN from reality datasets. In this work, we have applied the-state-the-art of CNN models above to propose the method for detecting video inter-frame forgeries.

Transfer Learning

To overcome the lack of big training dataset, transfer learning aims to transfer related knowledge which was learned before from source dataset to target [23]. Transfer learning was applied in many scopes that got potential results [24]. Because of the shortage of large datasets of video forgery detection, we fine-tuned recent models [8, 9, 11, 13, 22, 25, 26] which were pre-trained on ImageNet database then they were retrained with a target video dataset for detecting video inter-frame forgeries.

Optical Flow

Optical flow is the distribution of apparent movement velocities of brightness patterns in videos [27]. The optical flow has been applied for object detection, movement detection, and action recognition, etc., by analyzing optical flow consistency in the video, which has achieved breakthrough results in these fields. Besides that, until now in the video forensic field, optical flow also was applied in a few works [28, 29] which have gotten potential results. So, in this study, we proposed a method that would get optical flow features in the videos and train them by the advance deep learning models for detecting video inter-frame forgeries, and it has given potential results which would be shown in section IV.

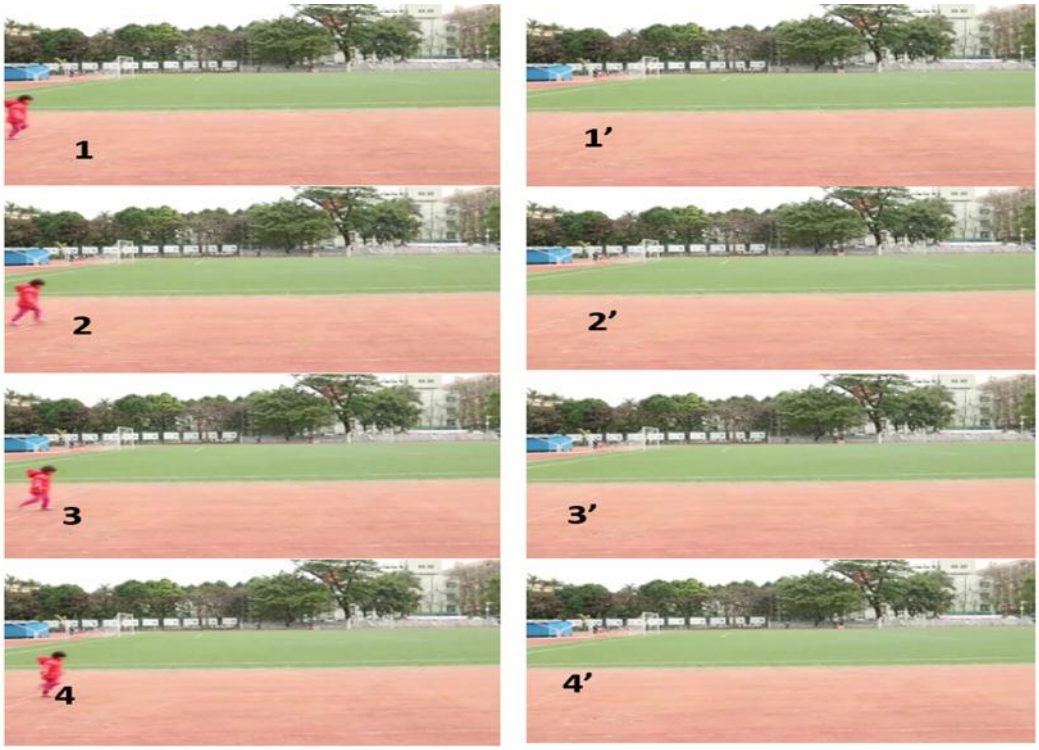

Fig. 2.Four left pictures are taken from an original video, and four right pictures are taken a forged video tampered from the original one in the same position. In the forged video, a baby was hidden.

III. Proposed Method

1. *Problem formulation*

The video inter-frame forgeries can include three types of forgeries shown in Fig. 3 as follows:

a) Deletion of a frames sequence shown in Fig. 3b, frames from 5^{th} to 9^{th} with dashed borderline were deleted to hide events inside a video.

b) Duplication of a frames sequence shown in Fig. 3c1 and c2; In Fig. 3c1 copied frames from 2^{nd} to 4^{th} then pasted at after 7^{th} frame in the same video with no frame deletion. This forgery is usually used to duplicate events in a video. Similarly, in Fig. 3c2 copied frames from 2^{nd} to 4^{th} then pasted at after 7^{th} but frames 8^{th} to 10^{th} were deleted. This forgery is usually used to duplicate an event while hiding another fact in a video.

c) Insertion of a frames sequence shown in Fig. 3d1 and d2; In Fig. 3d1 a sequence of frames from x1 to x4 copied from a different video then pasted at after 4^{th} frame with no frame deletion. This forgery is often used to add events from a different video into the video. Similarly, in Fig. 3d2 a sequence of frames from x1 to x4 copied from a different video then pasted at after 4^{th} frame, but frames from 5^{th} to 8^{th} were deleted. This forgery is often used to add an event from another video while hiding a fact inside the video.

All video forgeries above can easily manipulate videos by one of the video content editing software like Adobe Photoshop, Adobe After Effect, Video Editor and Window Movie Maker, etc. And those forged videos would present fingerprints, which are inconsistency in pixel values of temporal dimension between two consecutive frames at the manipulated position shown in Fig. 4. Those inconsistent pixel values are tough to detect because they can usually be very small inconsistencies when tampering videos sophisticatedly. To detect those fingerprints, we have proposed a method by applying the powerful state-of-the-art CNN models which are trained on ImageNet and fine-tuning then retrain them on the target dataset to detect those fingerprints. The detail of the proposed method is presented in the following section.

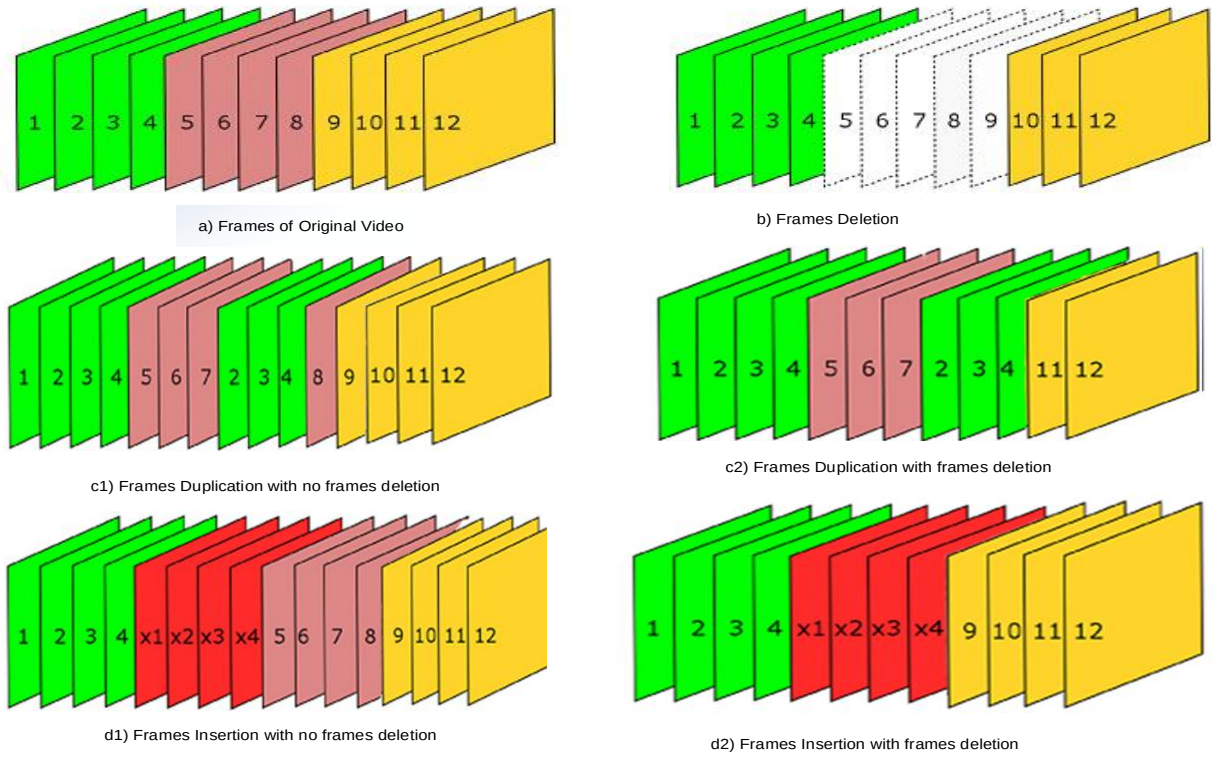

Fig. 3. Video inter-frame forgeries

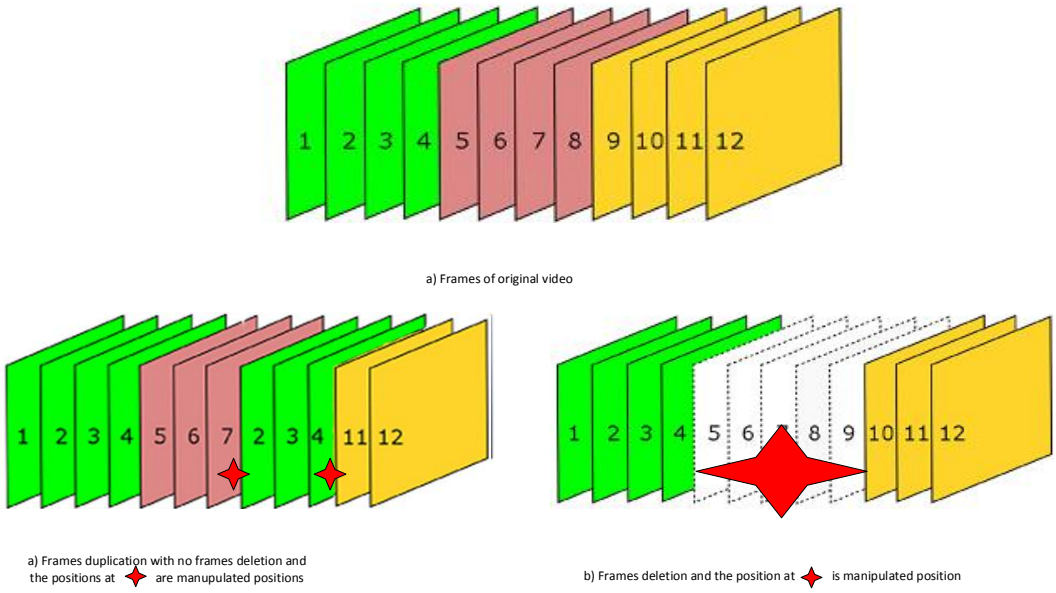

Fig. 4. The manipulated positions in video inter-frame forgeries

2. *Proposed method*

2.1 Pipeline of the proposed method

The proposed technique can detect video inter-frame forgeries based on the advanced CNN model to detect the fingerprints at the manipulated positions as shown in Fig. 4. The pipeline of the proposed model is shown in Fig. 5.

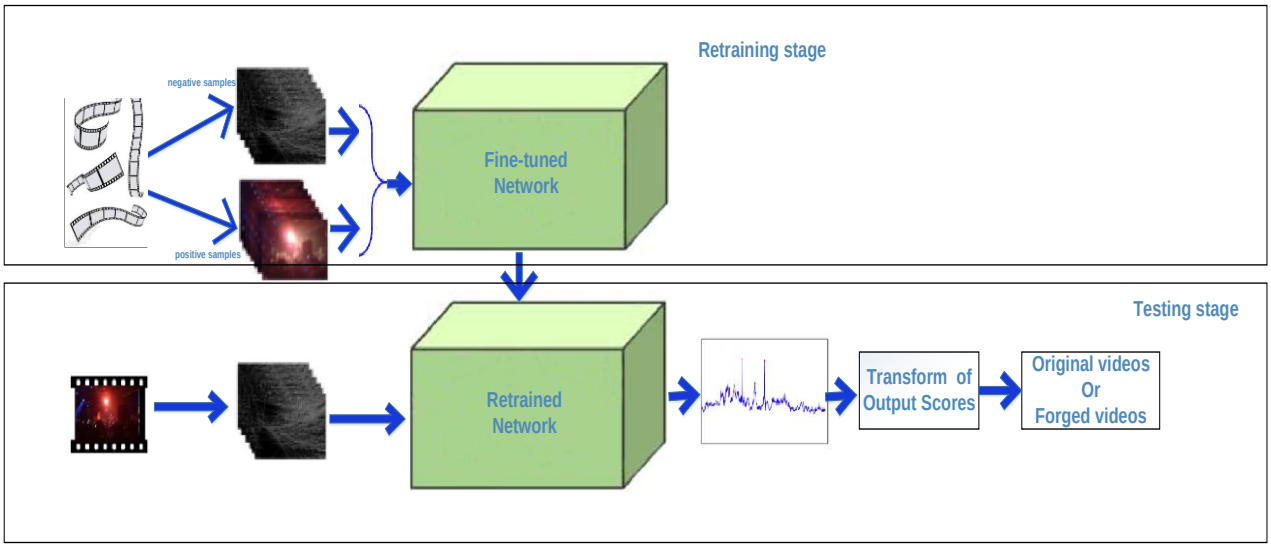

Fig. 5. The pipeline of the proposed method: In the retraining stage, the state-of-the-art CNN model was fine-tuned then re-trained on a target training dataset. In the testing stage, videos for testing were classified by retrained model and gave outputs that were later transformed into the confidence score.

We used recently state-of-the-art CNN models such as GoogleNet, ResNet, DenseNet, Inceptionv3, InceptionResnetv2, MobileNetv2, and NasNet, which were pre-trained on more than a million images from ImageNet database [16]. These models were fine-tuned then retrained again on the target dataset. The target dataset contains negative and positive samples that were created from the pristine videos dataset, the detail of creating a training dataset is discussed in the next subsection 2.2. The target dataset is input to the fine-tuned models for retraining and to build models, that expose features and highlight differences between negative and positive samples. Also, from the results of retrained models, we would compare each other to find out which architecture model is suitable for detecting video inter-frame forgeries.

The pipeline of the proposed method is shown in Fig. 5, which have two fine-tunings on the available model. Firstly, the available models were fine-tuned for classifying two categories (negative sample and positive sample) then retrained again on the target dataset. Secondly, we have transformed the output scores of models to a confidence score by temporal scaling to increase the testing efficiency of models as Eq (1). In the testing stage, videos for testing are separated into samples by following steps of creating negative samples which are discussed in subsection 2.2, then each sample is classified by retrained model and gives an output score f(i) which is in [0:1]. This score value is nearly '0' that means this sample is identified negative sample; otherwise almost '1' identified positive sample by retrained model. But to improve the classification result of the proposed model, the output score is transformed into the confidence score as follows:

$$f_{con}(i) = \min\left(f(i) - \sum_{l=1}^{10} \frac{1}{l} f(l), f(i) - \sum_{r=1}^{10} \frac{1}{r} f(r) \right) \quad (1)$$

Where l and r are the order of left and right sides respectively of i.

A video is original if $\max(f_{con}(i)) < Threshold$, where $i \in 1{:}T$, T is the number of samples in a video; otherwise it is forged. We have selected this *Threshold* equal to 0.5 in all of the experiments below.

2.2 Methods for creating training datasets

Because of training a model for detecting video forgeries that need a large number of videos including original videos and forged videos. To overcome the shortage of large video datasets for training models and to get advantage from the pre-trained models on ImageNet database, we have proposed four methods to build four different training datasets by basing on the residual or optical flow of adjacent or non-adjacent frames in original videos. The residual or optical flow of adjacent frames in the original video has consistency, which is used to create negative samples. Otherwise, The residual or optical flow of non-adjacent frames in the original video has inconsistency, which is used to create positive samples. In particular, these four methods create four datasets for retraining the model as follows: a) residual of two adjacent or non- adjacent frames, b) three residuals of grey value on four adjacent or non-adjacent frames, c) optical flow of two adjacent or non-adjacent frames, and d) three magnitudes of optical-flow on four adjacent or non-adjacent frames. Details of each method are as follows:

Let $X=\{x^t\}$ is an original video.
Where,
$t \in [1,T]$, T is the number of frames in the video.
x^t is the t^{th} frame in the video.

a. Creating a training dataset from the residuals of two adjacent or non-adjacent frames - Dataset_1

From the videos in original video dataset, negative samples were created by subtracting two adjacent frames, and positive samples were created by subtracting two non-adjacent frames, particularly in the following steps:

For creating negative samples:

Create R={r^t}, negative samples as follows: (2)

for t = 1 : T-1 do

$r^t = x^{(t+1)} - x^t$;

end;// for

Where,

x^t is the t^{th} frame in the video.

r^t is a residual of two adjacent frames as the difference between two adjacent frames, considered as a negative sample.

For creating positive samples:

Create $R^{'}$={ $r^{'t}$}, positive samples as follows: (3)

for i = 1 : T **do**

k = random(1:T);

/ The distance between two non-adjacent frames is at least 15 frames. So, randomly generated k until absolute of k greater than i-15*/*

while abs(k - i) <= 15 **do**

k = random(1:T);

end; //while

if k%2 <> 0

$r^{'t} = x^i - x^k$

else

$r^{'t} = x^k - x^i$

end; //if

end; //for

Where, $r^{'t}$ is a residual of two non-adjacent frames as the difference between two non-adjacent frames, considered a positive sample. Notably, in all experiments, we chose the distance between two non-adjacent frames inside the video at least 15 frames because in reality forgery manipulations at inter-frame usually tamper on length of frames sequence at least 15 frames.

b. Creating a training dataset from three residuals of grey values – Dataset_2

For creating negative samples

Create R={r^t }, negative samples as follows: (4)

for t = 1 : T-3 **do**

$r^t(:,:,1)$ = greyimage(x^{t+1}) - greyimage(x^t);

$r^t(:,:,2)$ = greyimage(x^{t+2}) - greyimage(x^{t+1});

$r^t(:,:,3)$ = greyimage(x^{t+3}) - greyimage(x^{t+2});

end; //for

Where,

x^t is the t^{th} frame in the video.

r^t is a sample including three residuals of grey values from four adjacent frames, considered a negative sample.

For creating positive samples:

Create $R^{'}$={ $r^{'t}$ }, positive samples as follows: (5)

for i = 2 : T **do**

k = random(1:T);

/ The distance between two non-adjacent frames is at least 15 frames. So, randomly generated k until absolute of k greater than i-15*/*

while abs(k-i) <= 15 **do**

k = random (1:T);

end; //while

if k%2 <> 0

$r^{'t}(:,:,1)$ = greyimage(x^i) - greyimage(x^{i-1});

$r^{'t}(:,:,2)$ = greyimage(x^k) - greyimage(x^i);

$r^{'t}(:,:,3)$ = greyimage(x^{k+1}) - greyimage(x^k);

else

$r^{'t}(:,:,1)$ = greyimage(x^k) - greyimage(x^{k-1});

$r^{'t}(:,:,2)$ = greyimage(x^i) - greyimage(x^k);

$r^{'t}(:,:,3)$ = greyimage(x^{i+1}) - greyimage(x^i)

end; //if

end; // for

Where, $r^{'t}$ is a sample including three residuals of grey values on four non-adjacent frames considered as a positive sample. Similarly, the distance between two non-adjacent frames inside the video is at least 15 frames.

c. Creating a training dataset from optical flow of two adjacent or non-adjacent frames – Dataset_3

Creating the Dataset_3 from the optical flow of two adjacent or non-adjacent frames is similar to creating the Dataset_1 by changing residuals to the optical flow of two adjacent or non-adjacent frames.

d. Creating a training dataset from three magnitudes of optical – Dataset_4

Creating the Dataset_4 from three magnitudes of the optical flow on four adjacent or non-adjacent frames is similar to creating the Dataset_2 by changing the grey values to the magnitude of the optical flow of four adjacent or non-adjacent frames.

IV. EXPERIMENTS

In this section, we present how to fine-tune state-of-the-art CNN models, configuration for retraining models on the target dataset, the testing results of the models and data preparation which we have collected and built to utilize in our experiments. Besides that, we have also compared the results with some latest researches which were performed on the same dataset.

1. Data preparation

Because of the shortage of large video inter-frame forgery dataset for training proposed CNN models, we have collected a dataset with 300 original videos from five surveillant cameras of VFDD dataset [30] which was taken from surveillant cameras in real life by our laboratory. This dataset was captured with diverse environments such as inside and outside school, offices, dormitories, streets, and buildings at the different light condition, daytime and night with light, and without light. The average length of videos is 10 seconds.

To create the training dataset, we have randomly selected 270 videos from this dataset, and followed the steps of creating datasets in section III, to build four training datasets. Finally, we have gotten four datasets, and each of them has about 143000 negative and positive samples. The summary of the training dataset has been shown in table 1.

Table 1. The summary of the training dataset.

Number of original videos	270
Average of videos length	10 seconds
Number of negative samples	71782
Number of positive samples	72026
Number of cameras were used for taking videos	5

To create a dataset for testing, we have used 30 remaining original videos from the 300 original videos dataset. We tampered these videos manually in different ways as following Fig. 3. By that way, we have 120 videos including 30 original videos and 90 forged videos, including the three types of video inter-frame forgeries above. An important note here, all the forged videos that we have tampered are not easily detectable with the naked eye. The summary of the testing dataset is shown in table 2. All of this dataset is published online at [31].

Table 2. The summary of the testing dataset.

Number of videos	120
Number of original videos	30
Number of duplication forgery	30
Number of deletion forgery	30
Number of insertion forgery	30
Average of videos length	10 seconds
Number of cameras were used for taking videos	5

2. Fine-tuning and retraining models

To retrain models on the target datasets, we have done fine-tuning of the state-of-the-art models by deleting the last three layers of those networks. Because the last three layers contain information on how to combine the features that the network extracts into class probabilities and original labels. Then add three new layers to the layer graph including a fully connected layer, a softMax layer, and a classification output layer. We have also set the final fully connected layer to have the same size as the number of classes in the target dataset (this case is 2). To learn faster in the new layers than in the transferred layers, we have set the learning rate of the fully connected layer equal to 5. Besides that, the rest of the training options were set as follows: Randomly selected 90% of the training dataset for retraining, 10% for validation. We used SGD optimization method which has momentum-contribution of the previous step is 0.9. The initial learning rate is 0.001, and the learning rate would drop 0.1 after 10 epochs; mini_batch_size is 10, max_epochs is 20 and shuffle at every epoch, L2 regularization is 0.0001.

3. Test results

For testing, each video in the testing dataset part would be followed by the steps of creating negative samples in section III. From that, we would have a set of samples from each video. This set of the samples are classified by the trained models above. Finally, each video is concluded to be a forgery or original video, which depends on the maximum of f_{con} values of that video. The video is original if max $(f_{con}(i)) <$ *Threshold*, where $i \in 1{:}T-1$, otherwise forged. And in all experiments, we set the *Threshold* equal to 0.5.

For performance measure, we depended on the following criteria:

Measures are used in this paper as follow True Positive (TP): forged video declared forged; False Positive (FP): original video declared forged; True Negative (TN): original video declared genuine; False Negative (FN): forged video declared genuine. Sensitivity or True Positive Rate (TPR); False Positive Rate (FPR) and Detection Accuracy (DA) as follow:

True Positive Rate:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

False Positive Rate:

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP}$$

Detection Accuracy:

$$DA = \frac{TP + TN}{N + P}$$

In the experiments, to find out which state-of-the-art CNN model is suitable to detect video inter-frame forgeries, we have fine-tuned then retrained all of them on the target Dataset_1. The average of three results from each model, as shown in table 3, and in Fig. 6,7 are examples show the progressing when training models.

Table 3. The average of three results from each retrained model on Dataset_1.

Fine-Tuned and Retrained Models	Detection Accuracy (%)	False Positive Rate (%)	True Positive Rate (%)	Parameters (millions)
SqueezeNet	94.17	10.0	95.56	1.24
MobileNetv2	96.67	6.67	97.78	3.5
Nasnetmobile	96.67	10.0	98.89	5.3
GoogleNet	92.5	16.67	95.56	7
ResNet18	97.5	3.33	97.78	11.7
DenseNet201	97.5	6.67	98.89	20
Xception	95.83	10.0	97.78	22.9
Inceptionv3	97.5	3.33	97.78	23.9
ResNet50	97.5	6.67	98.89	25.6
VGG16	95.83	6.67	96.67	138

Fig. 6. Progress of training the model based on Resnet18 model.

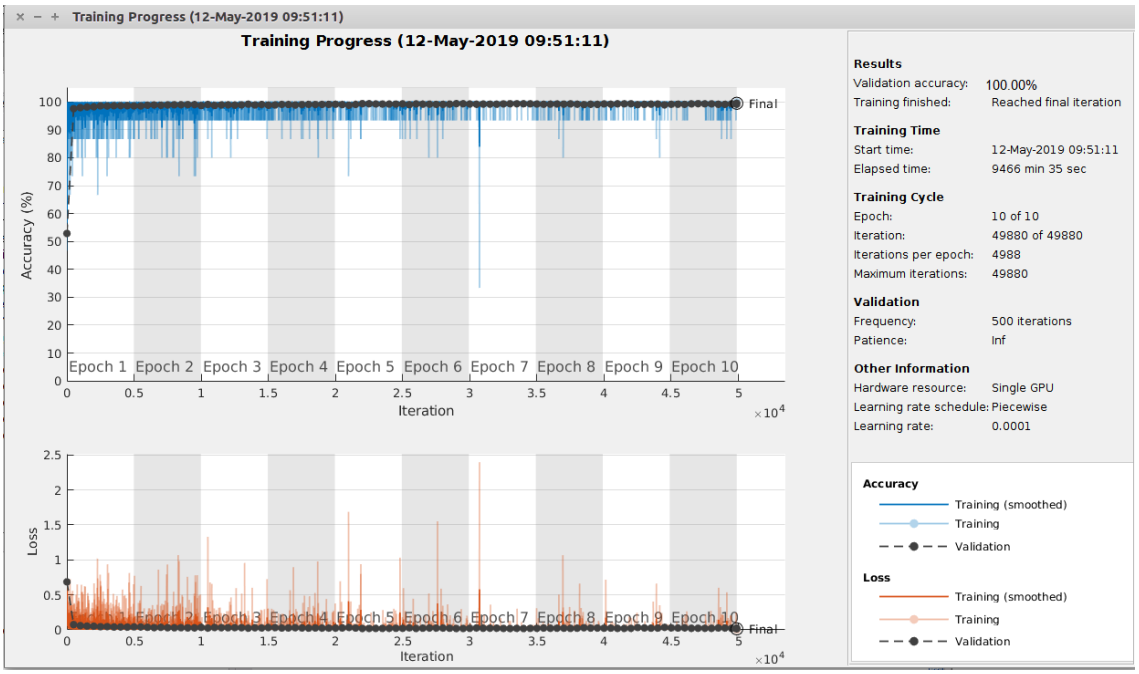

Fig. 7. Progress of training the model based on DenseNet201 model.

To compare the efficiency of the features in the 2.2 of section III which were used to create the training dataset. We have retrained on individual features and combinations of those features. All of the results were performed by the proposed method based on the pre-trained MobileNetv2 model and the results are shown in Table 4.

Table 4. The results of the proposed method based on MobileNetv2 when retrained on datasets built on different features.

Datasets	Detection Accuracy (%)	False Positive Rate (%)	True Positive Rate (%)
Dataset_1	**96.67**	**6.67**	**97.78**
Dataset_2	84.17	36.67	91.11
Dataset_3	83.33	13.33	82.22
Dataset_4	95.0	13.33	97.78
Dataset1 and Dataset4	**99.17**	**3.33**	**100**

To compare the efficiency between the transfer-leaning model and the model trained from scratch, whether it has a capability of transfer-learning on ImageNet database to detect video inter-frame forgeries. We have experimented on two models MobileNetv2 and Resnet18 by training from scratch and retraining from the pre-trained model on the same target dataset. The results are shown in Table 5.

Table 5. Comparing the results between transfer learning and training from scratch on MobileNetv2 and ResNet18 models on Dataset_1.

Methods	Detection Accuracy (%)	False Positive Rate (%)	True Positive Rate (%)
Mobilenetv2 (transfer learning)	96.67	6.67	97.78
Mobilenetv2 (trained from scratch)	84.17	33.33	90.0
ResNet18 (transfer learning)	**97.5**	**3.33**	**97.78**
ResNet18 (trained from the scratch)	93.33	16.67	96.67

To compare the efficiency between the proposed method and some recent methods. We have performed some experiments which were simulated from some recent methods on the same target dataset, and the results are shown in Table 6.

Table 6. Comparison between the proposed method and some recent methods on Dataset_1

Methods	Detection Accuracy (%)	False Positive Rate (%)	True Positive Rate (%)	Detection Types
Proposed Method (when combining Residual and Optical Flow features)	**99.17**	**3.333**	**100**	**Copy-move, insert, delete**
[18]	91.12	13.34	95.56	Insert, delete
[17]	86.67	16.67	92.23	Copy-move
[3]	93.34	10.0	96.67	Copy-move
[4]	88.89	16.67	94.45	Insert, delete

V. Results Analysis

The results in Table 3 show that usually the more parameters of models, the more accuracy of detecting video inter-frame forgeries. But besides that, there are exceptions that the models based on MobileNetv2 and Resnet18 have given quite good results as 96.67% and 97.5% accuracies respectively while the number of parameters is not too many. Accuracies of them are equal to some other models with more parameters such as DenseNet201 and Inceptionv3. So, the proposed method based on MobileNetv2 or ResNet18 may be suitable for applying to detect video inter-frame forgeries in situations that need fast processing speed or low hardware.

The results of the model trained from the four datasets built on different four features are shown in table 4. From these results, we have found that features like residuals of two adjacent or non-adjacent frames and three magnitudes of optical flow on four adjacent or non-adjacent frames are suitable for using to training and classificating in the proposed model. Especially, the accuracy would significantly increase to 99.17% by combining these two features.

Because of lacking large databases, so we have conducted some experiments to compare the model's efficiency between models trained from transfer learning and scratch. The results in Table 5 have proven transfer learning from the models that were pre-trained by ImageNet are more efficient in detecting video inter-frame forgeries. Notably, the accuracy of MobileNetv2 increased from 84.17% (when trained from scratch) to 96.67%, and ResNet18 risen from 93.33% (when trained from scratch) to 97.5% when trained from transfer learning models.

The results in Table 6 show that the proposed method for detecting video inter-frame forgeries have the accuracy as 99.17%, which is significantly better than of the recent methods. It has proved that the proposed method is significantly efficient in detecting video inter-frame forgeries.

VI. Conclusions And Future Directions

Nowadays with the rapidly developing hardware industry, especially the development of cameras, which were used for surveilling everywhere such as traffics, home, school, and office, etc. In addition, most people use smartphones which are also equipped with cameras. So, videos could be captured anywhere, manipulated anytime and spread quickly over the internet. The authentic video has great value as evidence. But until now although there are some methods for authenticating videos, they are either inefficient or very slow. In this study, we have proposed the method based on the state-of-the-art CNN models for detecting video inter-frame forgeries, which have shown the good and likely results, the accuracies are 97.5% and 99.17%. Through experiments, it has been proven that the proposed method has achieved significantly higher efficiency than the recent methods on the same dataset.

In the future, we will conduct in-depth research to propose suitable CNN architecture with fewer parameters and complexity for detecting and classifying the different types of video forgeries.

References

[1] Milani, S., et al., *An overview on video forensics.* APSIPA Transactions on Signal and Information Processing, 2012. 1.

[2] Yang, J., T. Huang, and L. Su, *Using similarity analysis to detect frame duplication forgery in videos.* Multimedia Tools and Applications, 2016. 75(4): p. 1793-1811.

[3] Singh, G. and K. Singh, Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation. Multimedia Tools and Applications, 2018: p. 1-36.

[4] Wang, Q., et al., Video inter-frame forgery identification based on consistency of correlation coefficients of gray values. Journal of Computer and Communications, 2014. 2(04): p. 51.

[5] Subramanyam, A. and S. Emmanuel. Video forgery detection using HOG features and compression properties. in 2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP). 2012. IEEE.

[6] Kobayashi, M., T. Okabe, and Y. Sato, *Detecting forgery from static-scene video based on inconsistency in noise level functions.* IEEE Transactions on Information Forensics and Security, 2010. 5(4): p. 883-892.

[7] Yu, L., et al., Exposing frame deletion by detecting abrupt changes in video streams. Neurocomputing, 2016. 205: p. 84-91.

[8] He, K., et al. Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Huang, G., et al. Densely connected convolutional networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[10] Chen, Y., et al. Dual path networks. in Advances in Neural Information Processing Systems. 2017.

[11] Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[12] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

[13] Szegedy, C., et al. Inception-v4, inception-resnet and the impact of residual connections on learning. in Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[14] Zoph, B. and Q.V. Le, *Neural architecture search with reinforcement learning.* arXiv preprint arXiv:1611.01578, 2016.

[15] Sandler, M., et al. Mobilenetv2: Inverted residuals and linear bottlenecks. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[16] Deng, J., et al. Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee.

[17] Wang, W. and H. Farid. Exposing digital forgeries in video by detecting duplication. in Proceedings of the 9th workshop on Multimedia & security. 2007. ACM.

[18] Chao, J., X. Jiang, and T. Sun. A novel video inter-frame forgery model detection scheme based on optical flow consistency. in International Workshop on Digital Watermarking. 2012. Springer.

[19] Liu, Y. and T. Huang, Exposing video inter-frame forgery by Zernike opponent chromaticity moments and coarseness analysis. Multimedia Systems, 2017. 23(2): p. 223-238.

[20] Long, C., A. Basharat, and A. 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.

[21] Carreira, J. and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[22] Zoph, B., et al. Learning transferable architectures for scalable image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[23] Oquab, M., et al. Learning and transferring mid-level image representations using convolutional neural networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[24] Weiss, K., T.M. Khoshgoftaar, and D. Wang, *A survey of transfer learning.* Journal of Big data, 2016. 3(1): p. 9.

[25] Chollet, F. Xception: Deep learning with depthwise separable convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[26] Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.

[27] Horn, B.K. and B.G. Schunck, *Determining optical flow.* Artificial intelligence, 1981. 17(1-3): p. 185-203.

[28] Jia, S., et al., Coarse-to-fine copy-move forgery detection for video forensics. IEEE Access, 2018. 6: p. 25323-25335.

[29] Kingra, S., N. Aggarwal, and R.D. Singh, *Inter-frame forgery detection in H. 264 videos using motion and brightness gradients.* Multimedia Tools and Applications, 2017. 76(24): p. 25767-25786.

[30] Al Hamidi, S., *VFDD (Video Forgery Detection Database) Version 1.0*. http://sites.scut.edu.cn/misip/main.psp, 2017.
[31] Xuan Hau, N. and H. Jongjian, *VIFFD - The data set for detecting video inter-frame forgeries.* Mendeley Data; http://dx.doi.org/10.17632/r3ss3v53sj.4, 2019. V4.

Authors' Profiles

Xuan Hau Nguyen was born in Phu Yen, Viet Nam, in 1981, received the B.E. degree in Computer Science from the Ha Noi Technology University, Viet Nam, in 2005. Xuan Hau Nguyen graduated the Master degree in Information System from Ho Chi Minh City National University, University of Natural Sciences, Viet Nam, in 2011 and is working toward Ph.D. degrees in electrical Information and Communication Engineering at the South China University of Technology, Guangzhou, P.R. China, from 2016.

In 2005, he worked at the Department of Computer Science, Mien Trung Industrial and Trade College, as a Lecturer, and in 2010 became the deputy of Computer Science Department at Mien Trung Industrial and Commercial College. He has published more than 6 peer reviewed papers. He currently research interests include multimedia security, machine learning and database system.

Yongjian Hu was born in Wuhan, Hubei, graduated from Xi'an Jiaotong University in 1990 with a master's degree in information and control engineering. Yongjian Hu received the Ph.D. degree in communication and information systems from South China University of Technology in 2002.

Now he works as full Professor in School of Electronic and Information Engineering at South China University of Technology. From 2011 to 2013, he worked as Marie Curie Fellow in the Department of Computer Science, University of Warwick, UK. From 2006 to 2008, he worked as Research Professor in the Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), South Korea. From 2005 to 2006, he worked as Research Professor in the School of Information and Communication Engineering, SungKyunKwan University, South Korea. Between 2000 and 2004, he visited the Department of Computer Science, City University of Hong Kong four times as a research assistant, senior research associate, and research fellow, respectively.

Dr. Hu is Senior Member of IEEE. He is also Senior Member of Chinese Institute of Electronics (CIE) and Senior Member of China Computer Federation (CCF). He has published more than 70 peer reviewed papers. His research interests include information hiding, multimedia security and machine learning.

Van Thinh Le was born in Phu Yen, Viet Nam, in 1976, received the B.E. degree in Computer Science from the Ha Noi Technology University, Viet Nam, in 2000, graduated the Master degree in Computer Science from Da Nang Technology University, Viet Nam, in 2011 and is working toward Ph.D. degrees in the School of Computer Science and Engineering, Southeast University, PR China, from 2014.

In 2002 he worked at the Department of Computer Science, Mien Trung Industrial and Commercial College, as a Lecturer. He has published more than 12 peer reviewed papers. He currently research interests include multimedia security, machine learning and database system.

Gohar Hayat Khan was born in Mianwali, Pakistan, in 1991. Received B.E. degree in Electrical and Electronics Engineering from University of Bradford United Kingdom, in 2014. Now is working Master degrees in electrical Information and Communication Engineering at the South China University of Technology, Guangzhou, P.R. China, from 2018.

Dinh-Tu Truong was born in Phu Yen, Viet Nam, in 1979**.** He received the Ph.D. degree in the School of Computer Science and Engineering from Southeast University, Nanjing, China in 2016.

He works at Ton Duc Thang University (TDTU), Ho Chi Minh City, Vietnam from 2016. His research area is network security, information hiding, network measurement, traffic sampling and machine learning.

Muhammad Ahmad Amin was born in Faisalabad, Pakistan, received his B.E. degree in Electrical and Electronics Engineering from The University of Faisalabad, Pakistan, in 2013. Muhamad Ahmad Amin received his Master's degree in Information and Communication Engineering from South China University of Technology, Guangzhou, P.R. China, in 2018 and is working towards his Ph.D. degree in Information and Communication Engineering at the South China University of Technology, Guangzhou, P.R. China, from 2018.

Currently, he is working as Researcher in Research Centre of Multimedia Information Security Detection and Intelligent Processing, School of Electronics and Information Engineering at the South China University of Technology from 2016. From 2017 to 2019, he worked as Researcher and Algorithm Engineer in R&D institute at GRG Banking, Guangzhou, P.R. China. His research interests include machine learning, pattern recognition, artificial intelligence, multimedia security, and forensics.

How to cite this paper: Xuan Hau Nguyen, Yongjian Hu, Muhmmad Ahmad Amin, Khan Gohar Hayat, Van Thinh Le, Dinh-Tu Truong, " Detecting Video Inter-Frame Forgeries Based on Convolutional Neural Network Model", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.12, No.3, pp. 1-12, 2020.DOI: 10.5815/ijigsp.2020.03.01

