Deep Learning-Based CNN Multi-Modal Camera Model Identification for Video Source Identification

Surjeet Singh, Vivek Kumar Sehgal
Department of Computer Science and Engineering and Information Technology,
Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India
E-mail: surjeetknmit@gmail.com, vivekseh@ieee.org

Keywords: convolutional neural networks, video forensics, audio forensics, camera model identification.

Received: September 12, 2022

Here is a high demand for multimedia forensics analysts to locate the original camera of photographs and videos that are being taken nowadays. There has been considerable progress in the technology of identifying the source of data, which has enabled conflict resolutions involving copyright infringements and identifying those responsible for serious offenses to be resolved. Video source identification is a challenging task nowadays due to easily available editing tools. This study focuses on the issue of identifying the camera model used to acquire video sequences used in this research that is, identifying the type of camera used to capture the video sequence under investigation. For this purpose, we created two distinct CNN-based camera model recognition techniques to be used in an innovative multi-modal setting. The proposed multi-modal methods combine audio and visual information in order to address the identification issue, which is superior to mono-modal methods which use only the visual or audio information from the investigated video to provide the identification information. According to legal standards of admissible evidence and criminal procedure, Forensic Science involves the application of science to the legal aspects of criminal and civil law, primarily during criminal investigations, in line with the standards of admissible evidence and criminal procedure in the law. It is responsible for collecting, preserving, and analyzing scientific evidence in the course of an investigation. It has become a critical part of criminology as a result of the rapid rise in crime rates over the last few decades. Our proposed methods were tested on a well-known dataset known as the Vision dataset, which contains about 2000 video sequences gathered from various devices of varying types. It is conducted experiments on social media platforms such as YouTube and WhatsApp as well as native videos directly obtained from their acquisition devices by the means of their acquisition devices. According to the results of the study, the multi-modal approaches suggest that they greatly outperform their mono-modal equivalents in addressing the challenge at hand, constituting an effective approach to address the challenge and offering the possibility of even more difficult circumstances in the future.

Povzetek: Razvita je metoda prepoznavanje izvornih kamer videoposnetkov s kombiniranjem zvočnih in vizualnih informacij z uporabo dveh DNN CNN tehnik.

1 Introduction

t should be noted that camera model identification has become increasingly important in multimedia forensic investigations, as digital multi-media content (including images, videos, audio sequences, etc.) is becoming more widespread and will continue to do so with the advancement of technology in the future. There is no doubt that a large part of this phenomenon can be attributed to the advent of the internet and social media, which have enabled a more rapid diffusion of digital content and, consequently, made it extremely challenging to trace their origins [28]. In forensic investigations, for instance, tracking the origins of digital content can be essential for identifying the perpetrators of such crimes as rape, drug trafficking, and acts of terrorism by tracing the origins of the digital content. There is also the possibility that certain private content may become viral through the internet, as has sadly happened in recent times with revenge porn, and there are other possibilities as well. It is therefore of fundamental importance to be able to retrieve the source of multimedia content in order to use it as a source [10]. The purpose of this paper is to determine the smartphone model used to acquire digital video sequences through the combined use of visual and audio information that has been extracted from the videos themselves. Due to the fact that there has been little work specifically done on identifying the video source in the forensic literature, we mainly focus on video source identification. In contrast, digital image analysis is one of the most commonly addressed aspects of digital imaging. Various peculiar traces left on the photograph when it was taken at the time when the image was taken can be used to identify

the camera model that was used to acquire the image [3]. The two main approaches that can be used to identify the model of an image camera are defined as model-based and data-driven approaches in this vein. In contrast, the model-based approach, on the other hand, focuses specifically on exploiting the traces that are released as a result of the process of taking a digital image, in order to be able to identify the type of camera from the traces as a result of being able to identify the information through the process of tracing. A significant number of other processing operations and defects using the same kinds of picture acquisition pipeline, including dust particles left on the sensor and noise patterns [11], have been demonstrated to be able to convey information and provide accurate information about a camera model that has been employed. In the last few years, the advent of digital data and computational resources led to the development of data-driven approaches that far outperform the solutions based on models. The data-driven approach is able to capture the model traces instead of focusing on a specific trace left by the image acquisition process, as is typical in model-based methodologies since the interaction of various components allows the approaches to capture model traces as well. Data-driven methodologies that have been most successful are those based on learned features, which in other words are methods that feed digital images directly to a deep-learning paradigm in order to learn model-related features and to associate images with the original source data [32].

The Convolutional Neural Networks (CNNs) are now becoming the most popular solutions in this field. As far as our knowledge goes, the only study that explores the problem of camera model identification on video sequences has been published. In this paper, we use advanced deep-learning approaches to develop effective methods for identifying camera models using video sequences in order to identify small patches from video frames, which they then fuse into a single accurate classification result for each video. In this paper, we use advanced deep-learning approaches to develop effective methods for camera model identification using video sequences. Specifically, we are proposing a method for recognizing videos by automatically extracting suitable features from the visual and audio content of the videos by using CNNs that are capable of classifying them by combining these features. Using a mixed-modal approach to solve the identification problem, we define the proposed strategy as multi-modal since we extract visual and audio information from a query video to solve the problem. It is important to note that, for visual content, we use patches cropped from the frames and, for audio content, we use patches cropped from the Log-Mel Spectrogram (LMS) of the audio track in the video that is used to solve the identification problem light of this, the method suggested by falls into the mono-modal category, since the authors rely solely on the visual content of a query video in order to determine its classification. In order to identify multi-modal camera models, we propose two distinct approaches based on this information [25]. With both approaches, we make use of CNNs and feed them with a pair of visual and audio patches in order to feed them with information. Our first approach consists of comparing and combining the individual scores obtained from a pair of CNNs that have been trained following a mono-modal strategy, that is, one CNN has been trained to deal with only visual data and the other CNN has been trained to deal with audio data only. The second approach involves training a single multi-input CNN that can be used simultaneously to process both visual and audio patches at the same time. For each of the proposed approaches, we examine three different network configurations and data pre-processing, which are based upon effective CNN architectures that are well known in the state of the art of video processing in order to maximize the level of performance. We evaluated the results in relation to the Vision dataset, which comprises approximately 650 native video sequences along with their related social media versions, which amounts to almost 2000 videos recorded by 35 modern smartphones [15]. The videos on which we conduct the experiments are not only the original native ones; we also use the videos that have been compressed by the algorithms of WhatsApp and YouTube in order to explore the effects of data recompression as well as to investigate challenging situations where the training and testing datasets do not share similar characteristics. To provide a baseline strategy for comparing the achieved results, we also investigate the mono-modal attribution problems. There is no doubt that the vast majority of state-of-the-art works in multimedia forensics in recent years have always dealt with video sequences by either exploiting their visual or audio content in a separate manner or by both. It has only been recently that both visual and audio cues have been used for multimedia forensics purposes, but they do not address the task of identifying the camera model used in those works. It is proposed that we evaluate the results obtained by exploiting only visual or audio patches in order to classify the query video sequence in a mono-modal manner [29]. Based on the results of the experimental campaign which was conducted, it can be concluded that the multi-modal methodology proposed is more effective than mono-modal approaches. Accordingly, the pursued multi-modal approaches have shown to be significantly more effective than standard mono-modal approaches in terms of solving the problem in a more efficient way. Moreover, we find that data that undergo stronger compression (e.g., videos uploaded to the WhatsApp application) are more difficult to classify than data that undergo a weaker compression (e.g., files uploaded to YouTube) [20]. In spite of this, we found that multi-modal strategies outperformed mono-modal strategies also in this complicated scenario". For the purpose of extracting feature descriptors from a sequence of images and categorizing them according to their descriptors, the algorithm for categorizing videos uses feature extractors such as convolutional neural networks (CNNs), which are comparable to feature extractors used for image classification. Using deep learning-based video categorization, it is possible to examine, categories, and keep track of activities in visual data sources such as video streams by examining, categorizing, and tracking these activities. In addition to surveillance, anomaly detection,

gesture recognition, and human activity recognition, video classification has many other applications as well.

1) For the purpose of classifying videos, the following steps can be taken as a guide to be taken as a guide.
2) Training materials should be created as part of the training process.
3) In order to classify videos, you need to select a classifier.
4) The classifier should be educated and assessed on a regular basis.
5) Using the classifier, you will be able to process the video data.
6) It is possible to train a classifier by using a large set of activity recognition video data sets, such as the Kinetics-400 Human Action Dataset, that are used for activity recognition.

A classifier can be trained by using a large-scale and high-quality set of activity recognition video data, such as the Kinetics-400 Human Action Dataset, which is a dataset collection composed of high-quality and large-scale activity recognition video data. Give tagged footage or video clips to the video classifier at the beginning of the process [39]. Using a deep learning video classifier that is composed of convolution neural networks, you may be able to forecast and categorize the videos based on the nature of the video input by using a deep learning video classifier that is constructed using deep learning techniques. As part of your process, you should ideally include evaluating your classifier as part of your analysis. It may also be possible to use the classifier to categorize activity based on a stream of live webcam video or a collection of video clips that are being streamed [17]. The Computer Vision Toolbox provides a variety of methods for training such as the slow and fast paths (Slow Fast), ResNet with (2+1) D convolutions, and two-stream Inflated-3D approaches as shown in Figure 1.

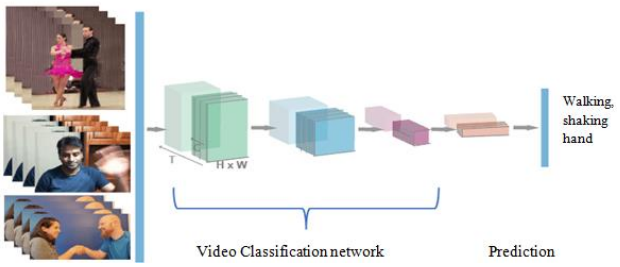

Figure 1: 3D techniques for training a classifier of video classification

1.1 An overview of camera calibration for DSLR cameras.

The manufacturers of DSLR cameras as well as other devices such as Canon, Nikon, and others often perform complex calibration algorithms before acquiring a scene image in their devices, which impacts the price of professional-level DSLR cameras considerably.

Therefore, it is necessary to develop effective, computationally less expensive, and affordable techniques of calibrating image-gathering equipment that are not inferior in quality to methods that are used abroad in order to make the process of calibrating equipment feasible and inexpensive for the masses [41]

1.2 Unique features of DSLR camera

There are a number of new digital forensic techniques that are being developed according to the unique characteristics of digital cameras that are closely related to the noise patterns from a few different kinds of DSLR cameras it is important to note that in order to solve the issues raised by the relevance of this work, it is necessary to limit one's attention to those kinds of noise and distortion that can be observed and detected, i.e., those that can be determined technically (experimentally) through the measurement of noise and distortion parameters obtained or those that can be observed by expert observation and subjective evaluation [23]. There is the possibility that other types of noise that were overlooked can also be ignored as they have little impact on the final noise component in the image due to their small impact.

This research is structured as follows: In Section 2 we briefly mention a related topic called the background of videos, whereas in Section 3 we describe methods that can be used to identify videos as sources of information. Section 4 explains the method of forensic video analysis, Section 5 outlines the problem statement, Section 6 explains the research methods, and Section 7 explains the results of the study. This paper provides an evaluation of the resolution method to be used with the Kaggle dataset as part of the resolution scheme we propose for using the Kaggle dataset. During the analysis that has been conducted, the results that have been obtained along with the analysis that has been conducted will be discussed. In the end, some conclusions are reached based on the findings of the study

2 Related works

It is possible to identify the camera model used to capture the photos and video frames shown in this article by using the numerous odd traces that have been left on the images and video frames during the shooting process that have been captured. It is here that we will provide the reader with some background information about the typical acquisition process of digital photographs so that, in the future, the reader will be able to better understand. In the next step, we will take a look at how we define the Mel scale, as well as the audio content of video sequences, in the next step. The author points out that the LMS is an excellent tool for studying how an audio track has changed over time, as well as how its spectral content has changed over time [14]. The issue of identifying image camera models over the past few decades has been addressed in a variety of ways over the course of the past few decades [9][21][13][35]. It is the aim of these approaches to derive noise pattern characteristics for each camera model from the images or videos that are supplied to them. The noise patterns or traces in these cameras are believed to be a

result of manufacturing defects and to be specific to each camera model [24].

2.1 Noise-based identification of digital video sources

In the field of multimedia forensics, there has been a great deal of attention paid to the task of blindly identifying the source device. By means of examining traces such as sensor dust and broken pixels, a number of strategies were put forth in order to identify the capturing device. When Lukas et al. first proposed the idea of utilizing Photo-Response Non-Uniformity (PRNU) noise to unambiguously define a camera sensor, they made a substantial advance in the understanding of the geometry of a camera sensor [7]. Because PRNU is a multiplicative noise, it cannot be effectively removed even by high-end equipment due to the fact that it is a multiplicative noise. The problem persists in the image even after JPEG compression at an average quality level has been applied to the quality level. In research on the viability of PRNU-based camera forensics for recovering photos from typical SMPs, it appears that alterations made to the photos by the user or the SMP could render the PRNU-based source identification useless

2.2 Analyzing the source of digitally identification videos

There is now digital identification technology built into new camera software to reduce the effects of unsteady hands on recorded footage caused by unsteady hands. In order to modify which pixels of the camcorder's image sensor are being utilized, this program evaluates the effect of user movement on which pixels on the image sensor are being utilized. It is generally true that image stabilization can be switched by the user on Android- based devices, but the camera software on iOS-based devices is not able to change this setting. In order to identify the source of videos shot with active digital identification using the PRNU fingerprint, the alignment of the fingerprints is disturbed during the identification process, which makes it impossible to identify the source of videos shot with active digital identification [30]. However, despite the fact that HSI has developed a reference side solution (which estimates the fingerprint from still photos), the problem still exists. Despite the fact that there are many variations in forensic video analysis techniques that could lead to the discovery of evidence, there are still many questions that need to be answered before they can be considered as being applicable. Additionally, forensic video analysis has shown to be more challenging than image analysis in terms of what it takes to make sense of the video's data. This is due to the fact that videos have more tightly compressed formats compared to picture formats [34]. An image frame is a series of images that make up the video that changes throughout time and evoke movement and change throughout time. A video is a video that contains a great deal of information that is encoded and decoded with the assistance of a mathematical technique called a codec, which encodes and decodes the information. In the multimedia file format, these previously encoded frames are wrapped up with tracks for the audio and metadata, as well as subtitles, and are known as multimedia files and are known as multimedia files.

3 Background

A number of strange traces were left on both the images and video frames that were captured during the shooting process. These traces have enabled researchers to determine the camera model that was used in order to capture these images and videos. We are trying to provide the reader with some background information about the typical digital picture collection pipeline in this section. In this way, they will be able to better comprehend the trace to which we refer in the next section. This will help them to understand it as well. After this, we define the Mel scale and the Log-Mel Spectrogram (LMS) of digital audio signals in order to be able to analyze the audio content of video sequences in the same way as we do the audio content of audio signals. LMS is a very valuable tool for examining the spectral and temporal evolution of an audio track. This is because it can be used to examine its spectral and temporal evolution based on its spectral and temporal characteristics.

3.1 A pipeline for acquiring digital images

In order to capture a picture with a digital camera or on a smartphone, we must initiate a complex process that involves numerous steps. This process involves numerous steps every time we use a digital camera. In a fraction of a second after pressing the shutter button, a short process begins which lasts only a fraction of a second. As soon as we are able to see the picture we have just taken, it stops. In general, the acquisition of a digital image does not follow a unique process.al image is not unique in most cases. There can be a significant variation in the vendors, the models of the devices, and the technologies that are onboard the devices. The picture acquisition pipeline can be thought of as a sequence of standard stages [42]. These are shown in Figure 2, which can be logically viewed as a sequence of standard steps.

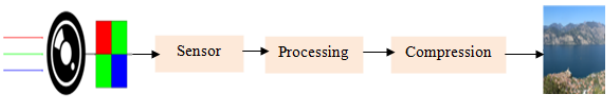

Figure 2: Acquisition of digital images.

3.2 A framework for analysis of forensic video

Compared to traditional photography-based evidence analysis in courts, forensic video analysis and the processing of multimedia evidence are still relatively novel fields compared to traditional photography-based evidence analysis in courts. It has become a growing trend over the last few years for a growing number of authoritative organizations, such as the Certified Forensic Video Analyst (CFVA) to recognize forensic video analysis as a significant objective norm, making its use in court more and more accepted. Forensic video

analysis can be classified into the following four categories: Law enforcement forensic video analysis,

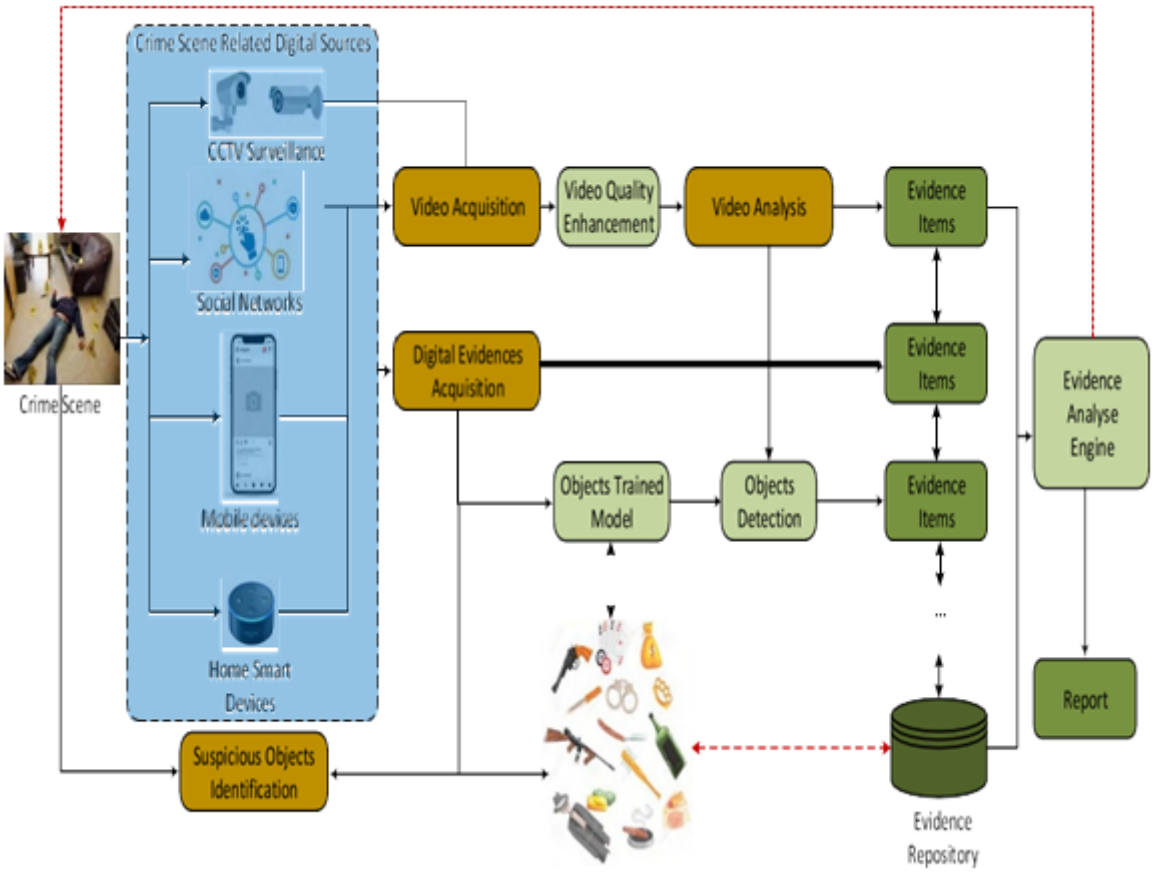

Figure 3: Enhanced forensic video analysis framework.

forensic video and multimedia analysis, image/video comparison, and enhanced forensic video analysis. These are the major factors that are being focused on by the newest forensic video analysis techniques [4].In our work, we focus on ”enhanced forensic video analysis,” i.e. the analysis of video and data using the most advanced video analysis tools. This enhanced forensic video analysis architecture, shown in Figure 3, is comprised of three fundamental parts: crime scene analysis, data collection, video enhancement and analysis, and presentation and enlargement of the findings”.

4 Method for the analysis of forensic videos

The preceding framework makes it obvious that there are two major categories of forensic video analysis that can be categorized in this manner an analysis of the content and type of video in a video. The retrieved pre-processed video is given to one or more CNNs in the CNN processing stage in order to extract unique characteristics among the many source camera models and categorize the original one[15].

4.1 A study of forensic video types and analysis

An obvious objective of forensic video analysis is to determine whether a video file has been unlawfully re-

produced or tampered with. In addition, it is critical to determine whether the video has been altered in any way. It is also possible to identify concealed information in this research by identifying the video source and analyzing the video steganography to identify concealed information. In particular, the identification of the video source is a key evidence source [19]. This is because it determines whether the video source is a camera or a device that tokens the video or image as shown in Fig.4.It has been confirmed that forensic audio analysis, forensic video analysis, image analysis, and computer forensics are all distinct fields of study as determined by the American Society of Crime Laboratory Directors Laboratory Accreditation Board (ASCLD/LAB). A large number of private, public, and state/local law enforcement organizations are now creating digital and multi-media sections within their organizations that may cover some or all of these disciplines. There are some agencies where the same person may conduct examinations for different agencies. It is quite common for examiners in large agencies, at the federal and state levels, and in one field to

specialize after years of training to become subject matter experts in their area. There are a number of ways in which video evidence can be enhanced [40]. It is very critical to submit the highest quality video recording in order to receive the most effective results from the enhancement process. A digital file or analogue copy that has been compressed with extra compression, if sent in for examination, may not be able to undergo the enhancement process. This is because it has been compressed with extra compression.

4.2 Enhancement of videos techniques

In order to achieve this goal, a wide variety of approaches has been used over the past decade to improve the quality of video. Several of these approaches have been developed for video monitoring systems intelligent highway systems, safety-monitoring systems, and a variety of other applications. As an example, [36]. have developed a method for identifying luggage from low-quality video footage by incorporating color information into the video footage. In order to identify the moving direction of an object, human-like temporal templates can be constructed and aligned with the appropriate parameters in order to identify the direction in which the object is moving. A number of authors have suggested that a system for detecting luggage should be created. As stated in Chuang et al., the purpose of the study was to detect missing colors using a ratio histogram. This variable is the ratio of the color histograms [31]. To find the missing colors, a tracking model should be used. From low-quality videos, forensics' primary goal is to extract as much information as possible from them in order to assist in the investigation process. It is the purpose of this section to present strategies for improving videos so that more information can be obtained from them. In low-quality videos/images, the likelihood of detecting additional information can be significantly enhanced using histogram equalization (HE)-based approaches compared to conventional approaches. Here is an example of how a webcam can be used to recognize objects using the suggested technique shown in Figure 4.

5 Problem formulation

In the present paper, we focus on the problem of identifying camera models from video sequences based on video content. As a primary focus of our research, we plan on identifying the source camera model from digital video sequences [33]. This has been attributed to the fact that digital image analysis has been extensively investigated in the forensic literature, without- standing results. In this study, we specifically work with video sequences that have been captured from a variety of smartphone models. This paper describes a novel method for combining informational and auditory information of videos under con- sideration to provide a comprehensive analysis of the videos under consideration [8]. We will first look at the classic mono-modal issue that seeks to identify the source camera model of a video sequence based on only visual or aural information, which will be discussed in the following sections. Next, we present the actual multi-modal problem identified in this research, which uses both visual and aural cues to identify the source of the sound.

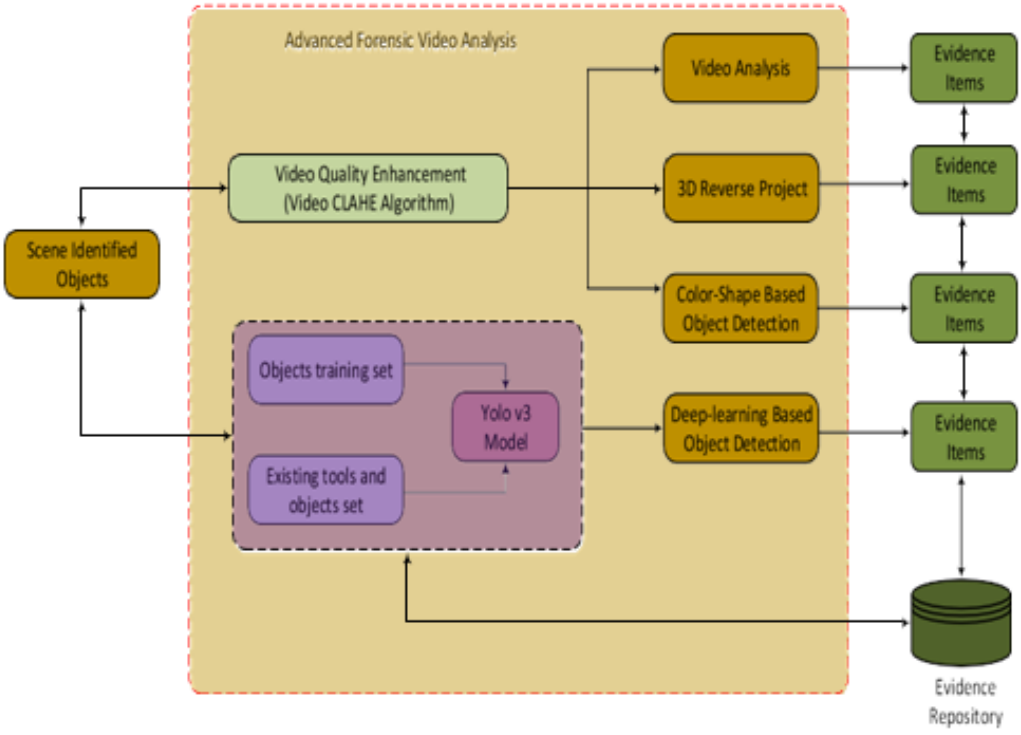

Figure 4: Video analysis procedures for advanced forensics.

5.1 Mono-Modal camera model identification

As a result, the problem is identified in the form of the device model, which was designed to acquire a particular media type in a single modality. When, for instance, an image has been captured, it is useful to know the model of the camera that was used to capture it. This is so that we can trace it back to its origin. In addition, if you have an audio recording, please include the model of the recorder that was used, along with the recording [26]. According to the mono-modal model attribution, in the context of a video, which is the situation we're interested in, the attribution of the device type that shot the video is identified solely based on the visual or auditory information contained within it.

5.2 Multi-Modal camera model identification

In the case of a video sequence, the challenge of multi-modal camera model identification is reduced to identifying the model of the device that recorded the video, taking both visual and aural information from the video sequence as input. In this example, we will consider a closed-set identification process that involves determining the camera model used to shoot a video sequence from a set of known devices that have been utilized in the past [38]. Assuming that the video being studied has been captured using a device from a device family familiar to the investigator, the investigator will assume that the video has been captured with a device of that device family. There is a possibility that the investigator will incorrectly assign a video to one of those devices if it does not originate from one of those devices.

6 Methodology

In this study, we present a method for identifying closed-set multi-modal camera models on video sequences that can be applauded in further research. In Figure 5 shows the main scheme of the proposal approach. Based on the visual and aural content of the video under consideration, we can determine the type of smartphone model that was used to capture the video. Using visual and auditory cues extracted from query video sequences, we input them into one or more CNNs that are capable of detecting the differences between different camera models used in the source video cameras based on their visual and auditory cues [2]. Two major steps comprise the proposed strategy, briefly:

1) Preprocessing and content extraction: The extraction of visual and auditory information from the videos under investigation, as well as the manipulation of the data before it is fed to CNNs, is referred to as pre-processing and content extraction.
2) There is a CNN processing block that consists of an extraction block that parses text into features and a classification block that consists of a CNN.

6.1 Content extraction and pre-processing

As part of the extraction and pre-processing step, visual and audio content is altered, as well as data standardization is performed.

There are three phases in this approach shown in Figure 6 that are involved in the extraction and pre-processing of visual content from the movie under analysis. These are:

1) It is possible to extract color frames from Nv that are equally distant in time and are spread out over a long period of time [12]. There are two sizes of video frames, which are Hv and Wv, and their sizes are determined by the resolution of the video being analyzed.
2) It is a raBy means of a random process, NPvcolour patches of the size HPV WPV are extracted at randomly to feed data into CNNs, patch normalization is carried out to ensure there is zero mean and unit variance.

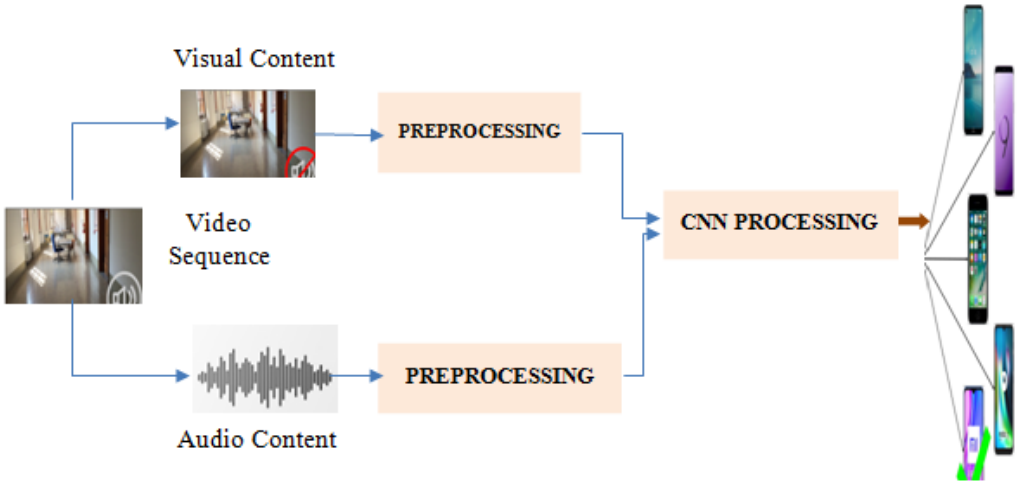

Figure 5: Pipeline of the proposed method.

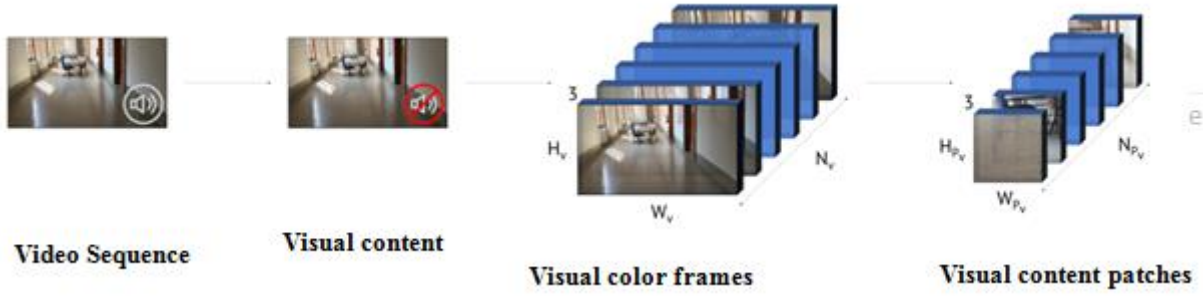

Figure 6: Process of creating a visual patch from a video stream.

There are three steps involved in the extraction and preparation of the audio material of the movie under examination shown Fig.7.

1) An extraction of audio content from the LMS L linked to the video sequence is performed. Considering this, it is clear that the LMS is a very useful tool for audio data and has been employed as a valuable feature for audio and speech classification and processing in a number of different studies. A number of audio characteristics were extracted from the magnitude and phase of the signal STFT during some exploratory experiments and it was determined that the LMS (based on the magnitude of the STFT signal) provided the best results. In the case of phase-based methods [1], LMS achieved an accuracy rate of less than 80%. As shown in the image below, the LMS L is a matrix of dimension Ha Wa, in which rows represent temporal information (which varies in length with the length of the video) and columns represent frequency content in Mel units.
2) Extraction of NPa patches of size HPaWPa randomly from L at random.
3) In order to achieve zero mean and unitary variance, patch normalization has been employed, as previously explained as described for the visual patches.

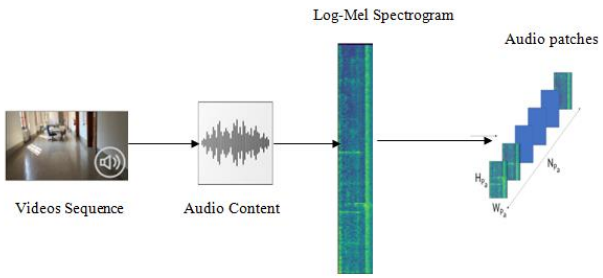

Figure 7: Audio patches extraction from a video sequence.

6.2 CNN processing

When the pre-processed information is retrieved, it is given to one or more CNNs in the CNN processing stage in order to extract distinct features based on the many source camera models and classify them accordingly demonstrate how it is possible to solve the mono-modal camera identification problem by feeding the retrieved visual or auditory data to a CNN [18]. In principle, any CNN architecture that is capable of classifying data could be employed at this stage; however, we discuss our choice in more detail in the next section.il in the next section. The final layer of the classification network is a fully connected layer with a number of nodes equal to the total number of models, M, where each node corre- sponds to a particular model of camera in the network. In this case, we are planning to produce an M-element vector with the name y, in which each element ym represents the likelihood that the model associated with the node was able to obtain input data. The node was able to obtain input data. We can extract it from the classification process by selecting the anticipated model m.

6.3 Early fusion methodology

As in the first method, the second method, called Early Fusion, involves combining two CNNs together to create a CNN with multiple inputs. In order to form the union, the final fully-connected layers of the two networks are concatenated, and three fully-connected layers are added until the prediction is formed As a result, the camera type is determined by the layer's dimensionality shown in Fig.8.

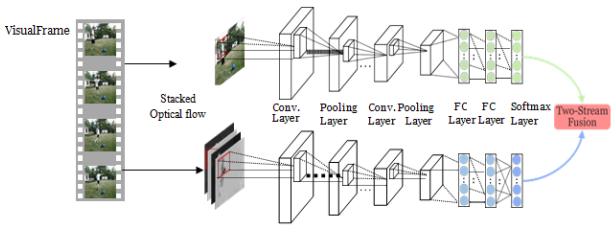

Figure 8: Early Fusion method pipeline.

Using the visual and audio patch pair, each Early Fusion forecasts the estimated camera model based on its estimated camera model in the final fully connected layer, yEF is the score obtained as a result of the final fully connected layer [37]. In the training phase, we use visual and audio patch pairs as a means of training the entire network. It is important to note that this is not the case with Late Fusion, since there is no separate training for the visual and audio branches. Similarly, both the training and testing phases are similar to those of the monomodal technique, except that we are distributing visual and audio patch pairs to the entire network this time instead of single patches (e.g., limited to visual or audio content). As shown in Figure8., the Early Fusion method's workflow is depicted in a flow chart. The size of the fully-connected layers' input and output features are also provided in order

to facilitate the design [16]. In addition, it is worthwhile to mention that the output feature at the final layer of the network has a size equal to M, which is the number of camera models that have been evaluated

6.4 CNN architectures

A CNN called EfficientNetB0 and a CNN called VGGish are the two CNNs we are using in order to solve this problem.

EfficientNetB0 is a member of the recently proposed Efficient Net family of CNN models. It has demonstrated excellent performance in multimedia forensics tasks and is one of the most promising models within the family We chose this Efficient Net model as it is the most basic model that we could use. As a result, we have a lot more time to experiment with different evaluation configurations as it enables faster training phases. It has also been demonstrated, also through preliminary experiments, that there is no evidence of a significant change if one uses parameters like This is an evaluation of EfficientNetB0's performance when compared to computationally heavier network models with more parameters that require more computation [6]. There are a number of CNNs being used for audio classification, including the VGGish CNN, which has been inspired by the well-known VGG networks used in image classification. In order to solve this problem, we are employing two CNNs, one referred to as EfficientNetB0 and the other referred to as VGGish. In the recently proposed Efficient Net family of CNN models, EfficientNetB0 is one of the members of the Efficient Net model family. Among the highest performing models within the family, it has demonstrated excellent performance in multimedia forensics tasks, and is one of the most promising models within the family [27]. We chose the Efficient Net model because it is the most basic model, we can apply to achieve our goals.ls. Therefore, we have a lot more time to experiment with different evaluation configurations. This is because have a much faster training phase due to the fact that we have more time to play around. As we have already seen through preliminary experiments, it has also been demonstrated that there is no evidence of a significant difference if one uses parameters like this is an evaluation of EfficientNetB0's performance when compared to computationally heavier models with more parameters that require more computation than EfficientNetB0 [5].

A number of CNNs are being used for audio classification, including the VGGish CNN, which is based on the well-known VGG network that is used for image classification, that has been inspired by the well-known CNNs that are used for audio classification. After exploring the dataset, you need to create the training set and the validation set. The training set will be used to train the model, while the validation set will be used to assess the model that has been trained. It is recommended to ex- tract frames from each video that is part of the training set and the validation set. After preprocessing these frames, train a model on the training set of frames after the preprocessed frames have been used. For the purpose of evaluating the model, use the frames from the validation set as input. In

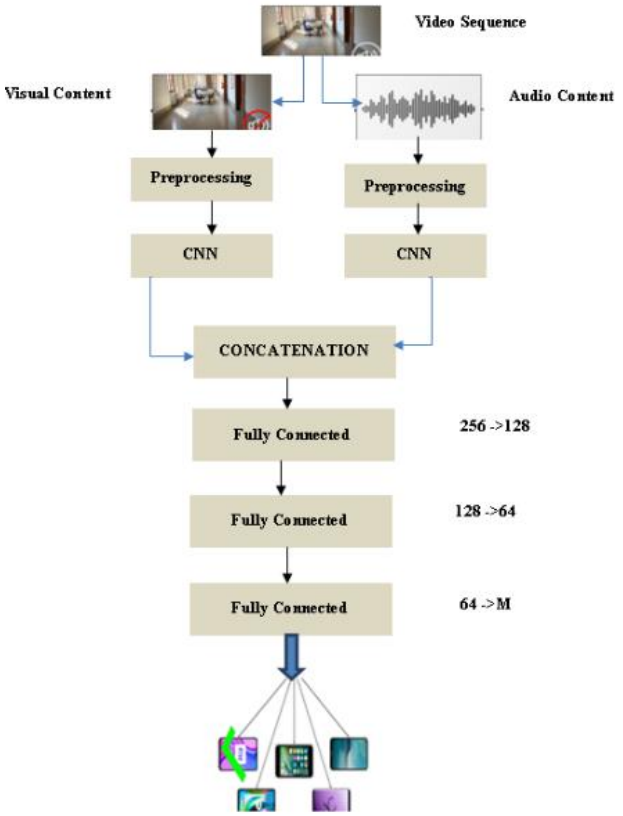

Figure 9: Processing pipeline for CNN's two-stream feature extraction.

the case that the performance on the validation set is satisfactory, we can use the trained model to categorize additional videos. According to Figure 9, the top portion of the figure shows the flow of the spatial stream's processing data. The CNN used for categorizing pictures is built in a similar way to a conventional deep CNN used for image categorization. In this method, each video frame is used as the input to the network, and then on top of that are added a number of convolutional layers, pooling layers, and fully connected (FC) layers.

7 Results

In this section, the dataset is processes first for experimental setup (i.e., the network training parameters and the configurations that we use in order to train the network). It is then reported what the evaluation metrics were, along with comments on what the results achieved.

7.1 Dataset

This study uses video sequences that are part of the Vision dataset. This is a recently released picture and video collection that has been created specifically for multimedia forensics investigations. Approximately 650 native video sequences were captured by 35 current smartphones and tablets, as well as their social media counterparts, as part of the Vision dataset. There are around 2000 video sequences in the collection, each of which has a clear indication of the source

device from which it was captured. In our trials, we selected non-flat movies (that is, movies displaying natural situations with objects) both from the original source (that is, videos that are obtained through the camera on a smartphone without any post-processing) and those that have been compressed by WhatsApp and YouTube.

In order to achieve the granularity, we seek in our analysis, we aggregate movies from different devices that belong to the same model. This allows us to analyze them at the model level. The videos taken from the device D04, D12, and D17 As per the Vision dataset nomenclature provided in this publication, lines D21 and D22 have been omitted because they cause problems with the extraction of frames or audio tracks. In addition, we exclude original videos that are not available on WhatsApp or YouTube in a compressed form WhatsApp or YouTube.

Unlike most other video analysis services out there, we don't just focus on high- resolution videos: while the majority of native videos have resolutions equal to or greater than 720p, we also examine native sequences with resolutions as low as 640480. As a result, we have 1110 videos that are around one minute in length, which were captured by 25 different cameras. In order to test the classification performance of the suggested technique, we use the available information about the model of the source camera as the ground truth for each video sequence. We extract 50 frames from each video sequence, equally distant in time and dispersed throughout the entire duration of the video sequence, in order to obtain the visual content. As a result, we extract 10 patches per frame (taken in random positions) for a total of NPv = 500 color patches per video. With 256 256 pixels as the patch size, we are able to achieve good results. Kaggle's dataset with ten classes and 275 instances may have been used as the basis for the feature extraction process. This could have resulted in issues such as overfitting and a decrease in the accuracy of the prediction. This was the reason why we constructed a fresh dataset with 1300 cases from three classes in order to overcome these situations: iPhone 6s, Xiaomi Note 4x, and Samsung Galaxy J7. Our next step was to introduce two new classes into the system. There are 275 Samsung Galaxy Note 3 and HTC One M7 examples included in the Kaggle dataset is shown in Table 1. In order to extract the features of the proposed model, the dataset was given to the model and the features were extracted to the model and the features were extracted. We categorized the camera models based on the characteristics retrieved from the retrieved data.

Table 1: Details of the dataset.

Model Name	**Number of Instances**	**Acquired From**
IPhone 6s	1500	self
Xiaomi Note 4x	1560	self
Samsung Galaxy j7	1600	self
Samsung Galaxy Note 3	1000	Kaggle
HTC One M7	550	Kaggle

According to Table 2, we present the error rate and the average confidence score for the test split of the patch dataset for different values of which have been found to lead to high misclassifications of adversarial instances while FGSM has not resulted in meaningful visual changes for untargeted attacks. Based on the patch test split, we discover that using = 0.005 provides the best compromise between error rate and apparent changes in the image, with the result that the trained DenseNet model detector has an average error rate of 93.1 percent and an average confidence level of 95.3 percent. When the value of increases, it should be noted that the manipulations become more visible as the value of rises.

This table displays our trained DenseNet model's error rate and confidence score following an untargeted FGSM assault to the test split. The second experiment, which is the CFA interpolation, is performed by simply taking the second set of features alone, which is the second set of features. According to the last analysis, the accuracy of the result was 86.93%. It is considered acceptable, but not enough, and it is still less than the result of the first experiment of co-occurrences alone, which was considered acceptable, but not enough. In order to achieve 97.81% accuracy on average, we combined the two feature sets into one and implemented them together. The average score achieved by all three sets was 98.75%.

Table 2: DenseNet model's error rate and confidence score.

Value	Error Rate (%)	Confidence Score (%)
0.01	97.3	97.8
0.02	94.8	91.0
0.03	92.6	93.9
0.04	93.7	92.8
0.05	98.4	94.8
0.06	96.7	98.6
0.07	91.5	99.4
0.08	90.6	97.1
0.09	92.0	92.0
0.11	91.4	91.2

According to Table 3, all the experiments mentioned above along with their accuracy rates are shown. The results of these experiments are presented in Table 3. The table below displays both the overall test accuracy as well as the test accuracy for each ConvNet for each of the three settings (flat, indoor, and outdoor) and each of the three compression types (native (NA), WhatsApp (WA), and YouTube (YT). Furthermore, these results are in agreement with tests that were conducted using N I-frames per movie for both training and testing. On the basis of PRNU, the best accuracy in the trials exceeds that of the limited counterparts by a large margin in each of the scenarios and compression types that were tested. On the VISION data set. As a comparison, we also conducted

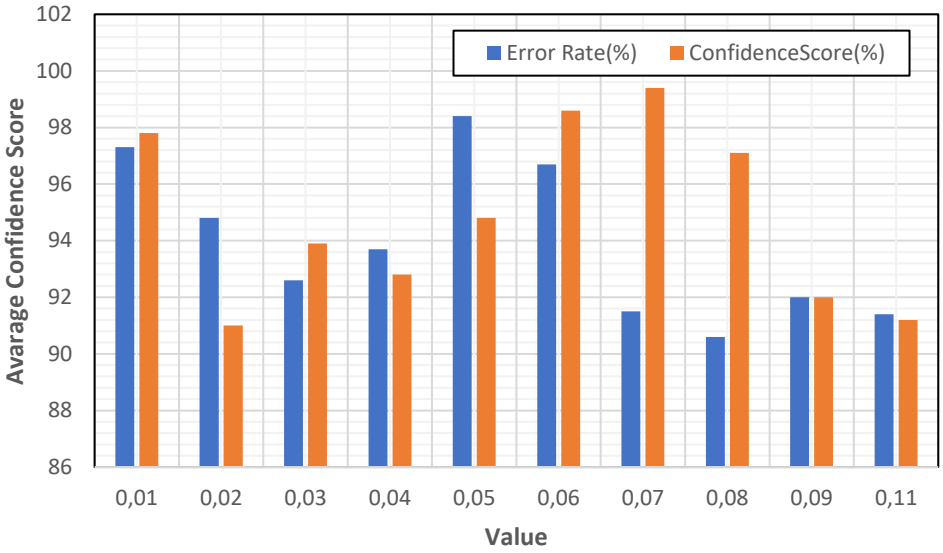

Figure 10: Comparison of the proposed method with other methods.

Table 3: Classification accuracy based on VISION data.

Model	N	Constraint Type	Overall	Flat	Indoor	Outdoor	WA	YT	NA
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
ResNet50	60	Conv	55.20	64.81	50.74	41.71	55.10	51.60	62.80
MobileNet	60	None	71.57	85.32	62.87	75.45	78.66	67.96	71.66
MobileNet	60	Conv	56.18	64.74	47.21	56.51	53.60	46.20	53.00
MobileNet	60	PRNU	62.70	63.96	53.11	61.12	58.80	63.50	67.30
MobileNet	60	None	75.87	76.92	64.62	75.02	74.84	77.68	75.90
MobileNet	60	PRNU	61.74	65.96	54.14	67.14	57.81	65.54	68.31

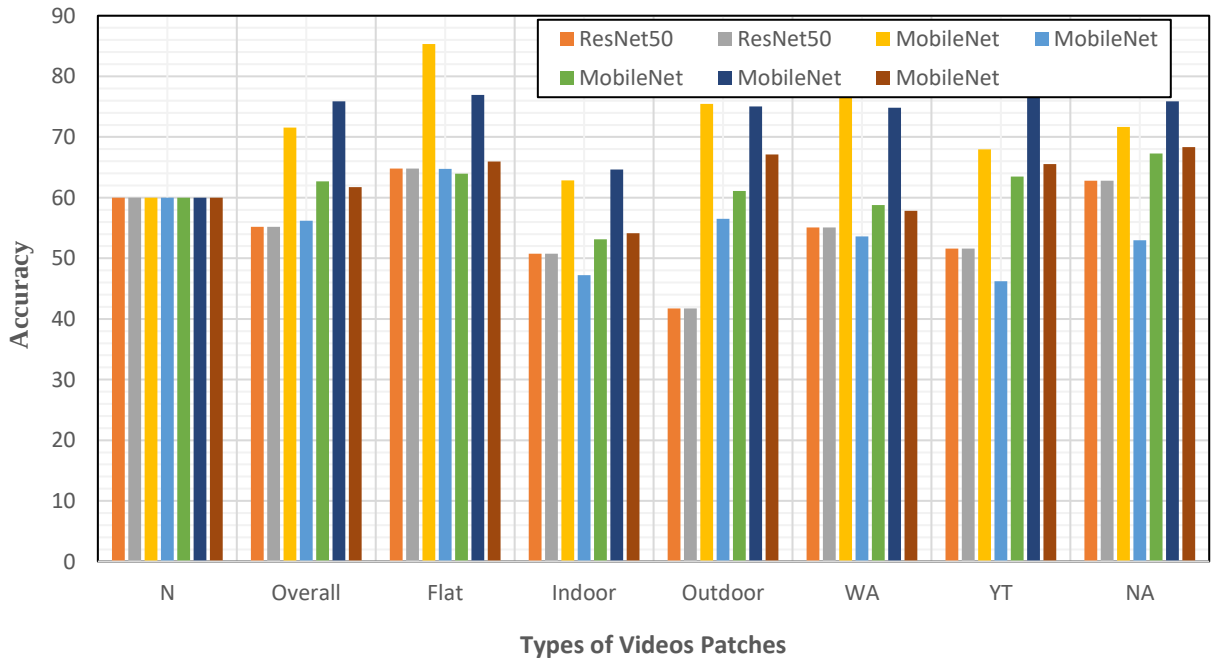

Figure 11: Classification accuracy of camera for proposed method.

Table 4: Compares the accuracy of MobileNet when it is compared to different counts of I-frames per video (I-fpv).

I-fpv	Overall	Flat	Indoor	Outdoor
1	69.12	71.1	57.5	76.5
5	72.31	79.8	59.6	75.4
30	74.10	82.1	62.3	76.0
50	73.51	81.5	61.6	75.4
100	73.71	82.1	61.6	75.4

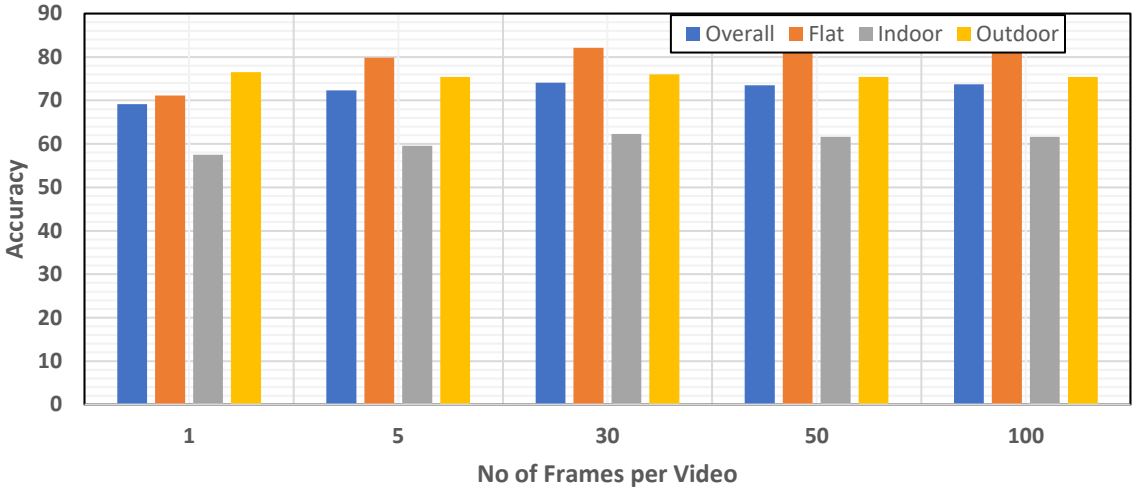

Figure 12: Test accuracy of mobile, net frames per videos.

the same experiment using the I-frames and the results are shown in Table 4. The results of the study show that the model achieves a high level of accuracy even when only a small number of tests I-frames are used. In addition, due to the short length of the movies included in the VISION data set, there are fewer I-frames available. Thus, even though we try to extract more I-frames, our accuracy remains the same, despite extracting more I-frames. As a result of our experience, we believe that the most effective overall strategy would be to apply the Late Fusion methodology in conjunction with configuring the EE192 according to our experience. With regard to native video sequences as well as YouTube video sequences, it consistently reports the most accurate results, regardless of whether it is a cross-test or not, and regardless of whether the test is a non-cross test or a cross-test. It is interesting to note that the cross-test results, including WhatsApp data, are on par with those of the other two configurations, if not a bit below. As a result of the fact that the trained CNNs. in this configuration are very adaptable to the data that they are shown during the training phase (i.e., patches selected from native or YouTube video sequences), they become less general and highly sensitive to significant data compression, such as that applied by WhatsApp, explaining the poor performance.

8 Conclusions and future works

The outcomes demonstrate that the suggested multi-modal methods are much more productive than traditional mono-modal methods. This research proposes a brand-new multi-modal methodology for identifying closed-set cameras models from digital video sequences that can be applied to digital video sequences. The overall objective of this research is to identify the smartphone model used to capture a query video by using visual as well as audio data from the video itself. Based on CNNs, the proposed method is devised to classify videos based on visual and aural information that can be extracted from the content of the video. The visual content of a video is derived from patches cropped from its video frames and the audio content is derived from patches cropped from the audio track's Log-Mel Spectrogram. To classify the query video, we use the Late Fusion method where we combine the scores obtained from two mono-modal networks (one working with visual patches and the other working with audio patches), and feed them into one multi-input network with visual/audio patch pairs extracted from the query video. The Early Fusion method uses a single multi-input network that is fed by visual/audio patch pairs extracted from the query video. It is important to note that both of these approaches are multi-modal methods of identifying camera models. Our study aims to examine three different topologies for each approach, with the use of various architectures and data pre-processing methods to do so. Using video clips that were taken from the Vision dataset, we assess the effectiveness of our experimental campaign. The videos we test are not just the original native ones that were captured by the smartphone camera directly, but we also test other videos as well. The purpose of this videos is to investigate a variety of training and testing configurations, as well as to come up with a way to simulate real-world scenarios in which it is necessary for

us to categorize data compressed through internet services. In order to achieve these goals, we also use movies compressed using WhatsApp and YouTube algorithms (for example, social media, and upload sites). In addition, we compare the multi-modal attribution strategy we propose to the traditional mono-modal attribution strategy as well as other suggested techniques [22]. On average, the Late Fusion technique provides the best outcomes of the var- ious multi-modal approaches and significantly outperforms traditional mono-modal approaches; the data confirms that the multi-modal approaches outperform mono- modal approaches. There are generally fewer than 99 percent chances that we will be able to correctly distinguish an original video sequence from a YouTube video sequence.99 percent. There are still some videos that are difficult to model, mainly because of the extreme compression used in WhatsApp, which may have something to do with the difficulty. It is obvious that this opens up possibilities for new problems and advancements centered around the identification of the originating camera model for videos that are posted (or shared repeatedly) on social media. Additionally, it is important to note that the suggested multi-modal solutions can be applied easily to a hypothetical situation where there are more than two data modalities being used. As a result of using the Late Fusion approach, the CNNs would only have to be trained independently on each target". When films share sequential data, one potential option would be to look into how neighbouring frames might be utilized for scene suppression and boosting the separation of camera noise [10].

References

[1] Abdali, S. (2022). Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. http://arxiv.org/abs/2203.13883

[2] Abdullakutty, F., Johnston, P., & Elyan, E. (2022). Fusion methods for Face Presentation Attack Detection. https://doi.org/10.3390/s22145196

[3] Akbari, Y., Al-Maadeed, S., Al-Maadeed, N., Najeeb, A. A., Al-Ali, A., Khelifi, F., & Lawgaly, A. (2022). A New Forensic Video Database for Source Smartphone Identification: Description and Analysis. IEEE Access, 10, 20080–20091. https://doi.org/10.1109/ACCESS.2022.3151406

[4] Akilan, T., Jonathan Wu, Q. M., Jiang, W., Safaei, A., & Huo, J. (2019). New trend in video foreground detection using deep learning. Midwest Symposium on Circuits and Systems, 2018-August (Cv), 889–892. https://doi.org/10.1109/MWSCAS.2018.8623825

[5] Al Banna, M. H., Ali Haider, M., Al Nahian, M. J., Islam, M. M., Taher, K. A., & Kaiser, M. S. (2019). Camera model identification using deep CNN and transfer learning approach. 1st International Conference on Robotics, Electrical and Signal Processing Techniques, ICREST 2019, January, 626–630. https://doi.org/10.1109/ICREST.2019.8644194

[6] Amerini, I., Anagnostopoulos, A., Maiano, L., & Celsi, L. R. (2021). Deep Learning for Multimedia Forensics. In Deep Learning for Multimedia Forensics. https://doi.org/10.1561/9781680838558

[7] Ashraf, A., Gunawan, T. S., Riza, B. S., Haryanto, E. V., & Janin, Z. (2020). On the review of image and video-based depression detection using machine learning. Indonesian Journal of Electrical Engineering and Computer Science, 19(3), 1677–1684. https://doi.org/10.11591/ijeecs.v19.i3.pp1677-1684

[8] Athanasiadou, E., Geradts, Z., & Van Eijk, E. (2018). Camera recognition with deep learning. Forensic Sciences Research, 3(3), 210–218. https://doi.org/10.1080/20961790.2018.1485198

[9] Bennabhaktula, G. S., Timmerman, D., Alegre, E., & Azzopardi, G. (2022). Source Camera Device Identification from Videos. SN Computer Science, 3(4), 1–15. https://doi.org/10.1007/s42979-022-01202-0.

[10] Bhatti, M. T., Khan, M. G., Aslam, M., & Fiaz, M. J. (2021). Weapon Detection in Real-Time CCTV Videos Using Deep Learning. IEEE Access, 9, 34366–34382. https://doi.org/10.1109/ACCESS.2021.3059170

[11] Blasch, E., Liu, Z., & Zheng, Y. (2022). Advances in deep learning for infrared image processing and exploitation. May, 56. https://doi.org/10.1117/12.2619140

[12] Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. Neurocomputing, 381, 61–88. https://doi.org/10.1016/j.neucom.2019.11.023

[13] Dal Cortivo, D., Mandelli, S., Bestagini, P., & Tubaro, S. (2021). CNN-based multi-modal camera model identification on video sequences. Journal of Imaging, 7(8). https://doi.org/10.3390/jimaging7080135

[14] Fan, H., Murrell, T., Wang, H., Alwala, K. V., Li, Y., Li, Y., Xiong, B., Ravi, N., Li, M., Yang, H., Malik, J., Girshick, R., Feiszli, M., Adcock, A., Lo, W. Y., & Feichtenhofer, C. (2021). PyTorchVideo: A Deep Learning Library for Video Understanding. MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, 3783–3786. https://doi.org/10.1145/3474085.3478329

[15] Gona, A., & Subramoniam, M. (2022). Convolutional neural network with improved feature ranking for robust multi-modal biometric system. Computers and Electrical Engineering, 101(November 2021), 108096. https://doi.org/10.1016/j.compeleceng.2022.108096

[16] Guera, D., Wang, Y., Bondi, L., Bestagini, P., Tubaro, S., & Delp, E. J. (2017). A Counter-Forensic Method for CNN-Based Camera Model Identification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017-July, 1840–1847. https://doi.org/10.1109/CVPRW.2017.230

[17] Hosler, B., Mayer, O., Bayar, B., Zhao, X., Chen, C., Shackleford, J. A., & Stamm, M. C. (2019). A Video Camera Model Identification System Using Deep

Learning and Fusion. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May, 8271–8275. https://doi.org/10.1109/ICASSP.2019.8682608

[18] Huynh, V. N., & Nguyen, H. H. (2021). Fast pornographic video detection using Deep Learning. Proceedings - 2021 RIVF International Conference on Computing and Communication Technologies, RIVF 2021. https://doi.org/10.1109/RIVF51545.2021.9642154

[19] Jagannath Patro, S., & M, N. V. (2019). Real Time Video Analytics for Object Detection and Face Identification using Deep Learning. 8(05), 462–467. www.ijert.org

[20] Maiano, L., Amerini, I., Ricciardi Celsi, L., & Anagnostopoulos, A. (2021). Identification of social-media platform of videos through the use of shared features. Journal of Imaging, 7(8). https://doi.org/10.3390/jimaging7080140

[21] Member, S., & Member, S. (2021). MMHAR-EnsemNet : A Multi-Modal Human. 21(10), 11569–11576.

[22] Ott, J., Atchison, A., Harnack, P., Bergh, A., & Linstead, E. (2018). A deep learning approach to identifying source code in images and video. Proceedings - International Conference on Software Engineering, 376–386. https://doi.org/10.1145/3196398.3196402

[23] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887–2905. https://doi.org/10.1007/s11042-020-08836-3

[24] Phan, T., Phan, A., & Cao, H. (2022). applied sciences Content-Based Video Big Data Retrieval with Extensive Features and Deep Learning. 1–26.

[25] Ramos Lopez, R., Almaraz Luengo, E., Sandoval Orozco, A. L., & Villalba, L. J. G. (2020). Digital video source identification based on container's structure analysis. IEEE Access, 8, 36363–36375. https://doi.org/10.1109/ACCESS.2020.2971785

[26] Salido, J., Lomas, V., Ruiz-Santaquiteria, J., & Deniz, O. (2021). Automatic handgun detection with deep learning in video surveillance images. Applied Sciences (Switzerland), 11(13). https://doi.org/10.3390/app11136085

[27] Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. Science Advances, 5(9), 1–10. https://doi.org/10.1126/sciadv.aaw0736

[28] Shi, Y., & Biswas, S. (2019). A Deep-Learning Enabled Traffic Analysis Engine for Video Source Identification. 2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019, 2061, 15–21. https://doi.org/10.1109/COMSNETS.2019.8711478

[29] Shi, Y., Feng, D., Cheng, Y., & Biswas, S. (2021). A natural language-inspired multilabel video streaming source identification method based on deep neural networks. Signal, Image and Video Processing, 15(6), 1161–1168. https://doi.org/10.1007/s11760-020-01844-8

[30] Shojaei-Hashemi, A., Nasiopoulos, P., Little, J. J., & Pourazad, M. T. (2018). Video-based Human Fall Detection in Smart Homes Using Deep Learning. Proceedings - IEEE International Symposium on Circuits and Systems, 2018-May, 0–4. https://doi.org/10.1109/ISCAS.2018.8351648

[31] Sreenu, G., & Saleem Durai, M. A. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data, 6(1), 1–27. https://doi.org/10.1186/s40537-019-0212-5

[32] Uddin, M. A., Joolee, J. B., & Sohn, K. A. (2022). Deep Multi-Modal Network Based Automated Depression Severity Estimation. IEEE Transactions on Affective Computing, 14(8). https://doi.org/10.1109/TAFFC.2022.3179478

[33] Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., & Chen, Y. (2022). Learning Two-Stream CNN for Multi-Modal Age-related Macular Degeneration Categorization. IEEE Journal of Biomedical and Health Informatics, X(X), 1–12. https://doi.org/10.1109/JBHI.2022.3171523

[34] Wang, Y., Sun, Q., Rong, D., Li, S., & Xu, L. Da. (2021). Image Source Identification Using Convolutional Neural Networks in IoT Environment. Wireless Communications and Mobile Computing, 2021. https://doi.org/10.1155/2021/5804665

[35] Wodajo, D., & Atnafu, S. (2021). Deepfake Video Detection Using Convolutional Vision Transformer. http://arxiv.org/abs/2102.11126

[36] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. IEEE Transactions on Neural Networks