Computer Assisted Methods in Engineering and Science
30(2): 111–130, 2023, doi: 10.24423/cames.447

This article belongs to the *Special Issue on Scientific Computing and Learning Analytics for Smart Healthcare Systems* edited by Dr. C. Chakraborty, Dr. S. Barbosa and Dr. L. Garg

Soft Computing Techniques-based Digital Video Forensics for Fraud Medical Anomaly Detection

Sunpreet Kaur NANDA[1], Deepika GHAI[1],
P.V. INGOLE[2], Sagar PANDE[3]*

[1] *Lovely Professional University, Phagwara, Punjab, India,*
e-mails: sunpreetkaurbedi@gmail.com, deepika.21507@lpu.co.in

[2] *Prof. Ram Meghe Institute of Technology & Research, Amravati, India*

[3] *School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India*

* *Corresponding Author e-mail: sagarpande30@gmail.com*

The current pandemic situation has made it important for everyone to wear masks. Digital image forensics plays an important role in preventing medical fraud and in object detection. It is helpful in avoiding the high-risk situations related to the health and security of the individuals or the society, including getting the proper evidence for identifying the people who are not wearing masks. A smart system can be developed based on the proposed soft computing technique, which can be helpful to detect precisely and quickly whether a person wears a mask or not and whether he/she is carrying a gun. The proposed method gave 100% accurate results in videos used to test such situations. The system was able to precisely differentiate between those wearing a mask and those not wearing a mask. It also effectively detects guns, which can be used in many applications where security plays an important role, such as the military, banks, etc.

Keywords: smart healthcare system, medical imaging, healthcare frauds, MRI imaging, digital image forensics, object detection, YOLO architecture, customized CNN.

1. Introduction

Digital and mobile cameras are commonly used as authenticated sources in forensic investigations. In examining the evidence of crimes, the video and photos extracted from such equipment are extensively employed, which may provide the main forensic evidence, aid evidence available or match evidence aspects to specific settings. Closed-circuit television (CCTV) devices are usually used in malls, banks, parking, retail, and even residential crossroads and areas, where video footage may be used as evidence even more than before. Audio and video-based pieces of evidence, alongside smart devices, can be easily accessible during research. The strategies for image improvement have been suggested [1–5] in the last several decades, most of which may be divided into various spatial and frequency domain approaches. Several approaches demonstrate strong potential to enhance picture quality. However, only a number of devices/methods, such as CCTV, video clips, etc., can be employed due to poor quality. Most monitoring systems in the CCTV industry export images of low-quality and need to be reformatted or transformed into a format that is easier and better to study. Images' low quality could also frequently lead to less reliability and loss of data, making them hard to examine.

Anomaly detection, also called aberration identification, is a method for identifying unexpected patterns that deviate considerably from data samples. Bank accounts, financial or medical scams, political contribution abnormalities, malignant tumors in an MRI. and even new universes and celestial objects can all be detected via anomaly detection. In medical fields, deep learning can be implemented for medical imaging anomaly detection.

Digital forensics footage is widely used for comparative study, involving forensic analysis, for comparing photographs of persons, cars, clothes, and armed people interrogated, with expert evaluation of the results [6–8]. In many current CCTV devices, internet offenders or other offenders are identified by face identification methods [9]. In recent years, other processes have been studied, for instance, motion identification, facial and body recognition, etc. In certain unfavorable scenarios with poor detection, it is quite hard to recognize persons from the face, body, stillness, etc. Although various approaches for the processing of images have been developed in recent decades, most do not benefit from the recognition of the face, body, etc.

When an individual or a corporation willfully misinterprets or mischaracterizes anything about the kind, extent, or quality of the medical care or services offered so that unlawful compensation is made, this is known as healthcare fraud. Misreporting aimed at defrauding taxpayer-funded health care is common and growing. This represents the enormous scale of the healthcare administration, the inadequacy of monitoring and transparency in the claims process, and the enormous sums of cash involved. All of these factors come together to make universal public healthcare an unstoppable objective.

To answer judicial system issues expertise from various science sectors is used, for example, forensic science. Increasingly, specialized areas, including computer science, engineering, and economics. are engaged in advancing illegal operations. Computer forensics covers the study disciplines of forensics and computer science, and includes research approaches based on hypotheses of a given problem through the usage of computers and electronic procedures. Nevertheless, computer forensics needs investigations and informatics to work jointly. Computer science join forces with forensic science in many aspects. Computer vision gives essential information in numerous researches to confront forensic issues, including ways for capturing, processing, analyzing, and comprehending real-world digital images. Computer vision and forensics combine data retrieval and analysis. More scientific assessment is accessible, and the actual application of such procedures continues to expand. Following difficulties have to be overcome in video-based forensic research. Identifying a suspect in forensics through images or video footage is always difficult due to the difference in the images in the gallery due to uneven face or weapon identification owing to the low quality of the image or video footage, angle of the camera, intensity of the image or video footage, etc. It requires developing more advanced approaches to enhance the image or video footage quality. To set up relationships, accessible document resources such as CCTV video, online image recordings, or historical trails of these artifacts among subjects in research situations are used. Because of growing innovations in the domains of social networks, the internet of things (IoT), smartphones, and advanced research methods are not just the end of everything for law enforcement applications but instruments exploiting any accessible information. Digital forensics locate relevant evidence from its current origins rapidly using certain smart methodologies utilizing the concepts of deep learning and artificial intelligence. For instance, in forensic research, for video footage-based face or weapon identification, the identification of unfamiliar faces encompasses aging of the faces, markings, forensic sketch identification, and near-infrared face or weapon identification, robust techniques for proof collection such as robust identification and the subject identification, etc.

Forensics tries to detect the elements based on small objects, whereas in medical image techniques entire image is required for training and analysis purposes. Healthcare data refers to information on an individual's or populations' health issues, fertility rates, reasons for mortality, and comfort of life. Medical measurements, as well as geographical, economic, and psychological data, are all included in healthcare data. When people engage with healthcare institutions, a variety of medical data is gathered and used. From a deep learning point of view. The medical image technique will take more time for training purposes compared to video forensics. Digital forensics experts have devised scientifically proven methods for identifying, collecting, preserving, validating, analyzing, interpreting, and presenting digital evidence derived from digital sources to facilitate the reconstruction of events that led to a breach. Indeed, with the help of digital forensics we can easily identify the suspected object, which can also help in securing medical data.

It is obvious that the digital inquiry of forensics into videos relies significantly on the quality of the video footage recorded, and that low quality considerably reduces the degree of trust in the process of investigation and does not thus prove to be worth submitting it to a court. In this study, we will try to resolve the issues mentioned above and present advanced methodologies for successful research in digital video-based forensics. The contribution of this article is as follows:

- The proposed framework deals with identifying guns and masks as a part of digital forensics.
- Identification of guns, masks and suspicious activity is implemented using a customized CNN architecture.
- Identification of guns, masks and suspicious person is implemented using the "you only look once" (YOLO) architecture.
- The results obtained from both the architectures are compared.

This article is divided into the following sections. The presented section dealt with introducing the concept related to video-based forensics. Section 2 highlights the discussion of related research work and the practices used in the forensic investigation. Section 3 discusses the methodology for the proposed framework. Section 4 deals with presenting and discussing the obtained results from the proposed framework. Finally, Sec. 5 presents the conclusion and possible future works based on the proposed framework.

2. Related work

Active research is going on in forensics related to computer vision as a part of artificial intelligence studies to protect society from unexpected violent scenarios. This became a major important aspect for most countries across the globe when some countries had to act strongly against terrorism. Jerian *et al.* [10] in 2007 presented a forensic application, an image processing-based software developed in the environment of MATLAB. The goal of the proposed framework was to counteract a few of the disadvantages that arise while using regular image processing applications applied to this study's particular scenario, such as the lack of control and documentary evidence over the functions conducted on the images, as well as the lack of advanced and more efficient methodologies the can enhance the performance and help bridge the gap in critical cases.

While the activity identification community has concentrated mainly on recognizing basic motions such as clapping, walking, running, identifying fighting and other hostile behaviors has received far less attention. Such a method might be quite beneficial in certain video monitoring scenarios, such as prisons, mental or elderly facilities, or even in camera phones. Nievas *et al.* [11] in 2011 employed the well-known bag-of-words structure for activity identification, as well as two of the finest activity descriptors currently offered: STIP and MoSIFT, to evaluate a well-defined bag-of-words structure to detect fight. They also provided an advanced video repository including 1000 frames classified into two classes: fighting and non-fighting, to assess and investigate video-based violence identification. Research on one repository and another containing fightings from action films demonstrate that fights can be identified with an accuracy of nearly 90%.

The procedure of analyzing a video, obtaining information, and analyzing the information to obtain domain-specific information is known as video analytics. Apart from evaluating any video for the data retrieval, live security footage analysis for identifying actions that occur inside its service coverage has grown increasingly significant in recent years. These mechanisms work in place in real-time. By utilizing a training model with the aid of an artificial neural network, automated face identification from surveillance footage becomes easier. Skin color estimate aids in-hand identification. From footage acquired by a security camera throughout tests, Gowsikhaa and Abirami [12] in 2012 identified suspicious behavior such as object transfer, new individual admittance, peeping at somebody else's work or test, and individual swap. This necessitated the use of face identification, hand identification, and the identification of touch between an individual's face and hands as well as between various people's faces and hands. The automation of suspicious activity identification will aid in lowering the error rate associated with manual surveillance.

Mobile devices are the primary means of communication in contemporary society, enabling users to send and receive messages, ideas, movies, and audio. There are a variety of instant messengers for mobile devices, which are a superior substitute to SMS technology. Increased use of instant messengers, on the other hand, has a negative influence, particularly as undesired behaviors related to cybercrime. On Android phones, the most popular instant messengers are WhatsApp and Viber. Lone *et al.* [13] in 2015 used forensic analytical techniques to acquire artifacts from the WhatsApp and Viber apps. Messages, contacts, chat history, attachments, and other artifacts from the mobile device's storage were the subject of their investigation. The authors shared their results after utilizing widely accessible applications and software to carry out forensic tests. The artifacts discovered in their inquiry can be used in a court of law in the criminal offense case.

Kamenicky *et al.* [14] provided a series of forensic image and video analysis techniques. These were created to assist in determining the reliability and source of images and videos, as well as to recover and improve image quality by reducing undesirable impression, noise, and other undesirable details. Their research originated from using photos and videos in criminal investigations, as such usage is considered to be among the best practices. Determining the image origin, verifying the image contents, and image recovery are considered the most essential difficult tasks in which automation may assist criminalists in their work.

Because digital recordings are routinely used for security reasons, most of research has focused on developing video forensic technologies. The main goals of forensic analysis of videos are: identifying evidence in a video and authenticating the actual video origin. Wan *et al.* [15] in 2017 presented an automatic jump-cut identification method to assess video manipulation and modification utilizing an innovative, reduced cost, and efficient video forensic methodology influenced by the human visual system, which identifies modifications that the naked eye may not be able to notice. The authors' aim was to determine the quality of digital video footage. The outcomes of their experiments show that their assessment is capable of reliable recognition and authorization in the applications of digital video forensic cases.

CCTV is an electronic camera used to capture surveillance footage and one of the most prevalent electronic camera techniques providing digital evidence for forensic investigation. The footage featuring the targeted individual or item is taken from the CCTV records for further study in the analysis of video forensics. Nevertheless, the performance quality of these records is frequently low due to various circumstances, including the type of cameras, setup, and location. The sharpness of the CCTV footage has a significant impact on the outcomes of forensic face identification. Low standard CCTV footage lowers the degree of trust in the facial recognition, making it ineffective, as proven by its use in examinations in courts of law. The aim of the study conducted by Senan *et al.* [16] was to provide a methodology for evaluating the CCTV quality data for use in forensic face identification analyses. There were two steps to this methodological study. The research was carried out using several CCTV cameras with various resolutions and distances between the person and the CCTV in the first step. The individuals' features were matched to the faces obtained during the enrolling process in the second step. The forensic face identification program's results were determined by camera resolutions, types of cameras, distances, and variations in ranking score upon implementing augmentation processes similar to the bicubic interpolation of the face images.

For the acquisition and evaluation of flexible data, a detailed understanding of both the staging and forensic instruments is essential. There are various tools available to make mobile phone crime scene inspection easier. In [17], the key tools available on the market to support Android devices in terms of their ability to extract and analyze data on a variety of factors are compared. Gathering information from a Google Android phone is an important issue in many criminal inspections. An Android phone may save a lot of information relevant to an inquiry. Standard Android phone data such as SMS, internet searches and history, phone log information, and mobile accounts are such information. The analysis, implementation and performance of the Google Android mobile forensic application and the findings are the topics of the research by Kaur and

Choudhary [17] in 2017. On completing forensic processes, the authors presented their study findings. Their study shows that the artifacts gathered during the research design can be used as proof in a trial of court against every criminal activity.

As an approach for multi-purpose detection of image manipulations under anti-forensic attacks, Chen *et al.* [18] in 2018 proposed a novel convolutional neural network technique. To increase the transmission of common characteristics relevant to image alteration identification, the dense interconnectivity patterns, which have greater parameters efficiency than the traditional patterns, were investigated. Their research shows that the suggested CNN design outperforms three other state-of-the-art approaches. The suggested approach can also increase resilience against JPEG compression, with maximum improvement of 13% accuracy under the low-quality JPEG compression technique.

Singh and Singh [19] in 2019 described a submissive blind methodologies consisting ot two different algorithms to detect video frame and region duplication forgeries in videos. The authors looked at the video frame duplication forgery in three different forms such as duplication of a pattern of a sequence of consecutive video frames at long running situation, duplication of many such sequences at many different locations, and duplication from other videos with different and similar dimensional aspects. It is a challenge to detect this copy-move forgery due to a slight change in pixel intensity values in the duplicated region and providing high correlation as authentic region. To compute the similarity among areas in two images or inside the impacted image, the second methodology in the suggested system detected these region duplication forgeries in videos by detecting the incorrect position using the deadline procedure. The experimental findings reveal that the suggested technique has a greater identification accuracy and execution time efficiency than the latest efficiency techniques.

In video surveillance settings, for example, train stations, gymnasiums, and mental health centers, it is very important to detect aggressive/violent conduct automatically. However, the previous detection methods extract descriptors in the spacetime locations/points of interest or statistic features in the motion areas, resulting in limited ability to effectively identify video-based aggression Zhou *et al.* [20] in 2018 offered a new way of detecting patterns of violence to overcome this problem. Initially, the activity areas are separated by optical flow domain distribution. Next, the authors suggested that two types of low-level characteristics in the activity areas to be extracted to reflect the appearance and dynamics of aggressive behaviors. The local histogram of oriented gradient (LHOG) description derived from colored images and the local histogram of optical flow (LHOF) description derived from optical flow images are the proposed low-level features. Then, the extracted features are encoded with the gag of words model to remove redundant data, and a specific-length vector is generated for each video clip. At last, the generated vectors are categorized by support vector machine (SVM). Research findings on three challenging benchmark datasets showed that the suggested methodology of identification is better than earlier techniques.

Ramzan *et al.* [21] in 2019 examined different state-of-the-art identification approaches. The methodologies of identification were divided into three groups based on classification methodologies: violence detection using traditional machine learning (ML), the usage of SVM, and deep learning. The feature extraction techniques and object detection techniques of the each single method were also presented. Furthermore, datasets and videos employed in the techniques, which play a crucial role in identification process were also discussed. The phases of the research methodologies were shown in an architecture diagram for better understanding. The results of the research have been discussed, which may help in finding the possible future studies in this field.

Aerial satellite imaging can be easily acquired and shared. Once the accessibility of powerful traditional and automated image editing technologies is given, the integrity of these types of images cannot be assumed anymore. Horváth *et al.* [22] proposed a deep learning-based algorithm to recognize and locate splicing alterations in overhead photos. Their method exploited recent advances in anomaly detection and required no previous knowledge of the type of manipulation that an opponent could insert in the satellite imagery. The authors compared their strategy with robust satellite-based manipulation detection systems. The authors also showed that the suggested strategy outperforms all existing techniques, especially when detecting small modifications.

To learn spatial-temporal data on video clips under different scenarios: subjective- and conceptual-based, Peixoto *et al.* [23] in 2019 studied several violence identification methodologies that rely on two deep neural networks (DNNs) frameworks. The proposed study considered deep feature representations for each specific concept, and then combined them by forming a superficial neural network to describe violence as a whole. Lastly, the authors demonstrated that employing more specific concepts is a straightforward and successful strategy, besides being complementary to producing a more precised description of violence.

The aim of Kaur and Jindal's study in 2020 was to provide an effective way of unfolding new features of forgeries for researchers working in the field of image and video forensics [24]. Their report used full research to help researchers face the different obstacles encountered in earlier studies. The paper examined the splicing and copy-moving methodologies of forgery detection in images and interframe and intra-frame forgery in videos, highlighting the commonly used datasets and thereby assisting new researchers in their work. The article's novelty was that such collaborative study was not conducted under one framework.

2.1. Research gaps based on the existing frameworks

The research gaps based on the existing frameworks are enlisted as follows:

1. Intelligent computer vision algorithms are available in state-of-art model [1, 7, 8, 15, 16]. However, it is required that video forensics detection should be automatically conducted along with it.
2. Classification results are available [5, 6, 12], but they must be high performance in – training, testing and validation stages.
3. It is required that high-performance hardware with a multi-core system should be available to train the deep learning models. In video forensics, traditional feature extraction methods [4] are time-consuming and low performance.
4. The state-of-art methods have generalized data sets, but today there is a need for the data sets to be equally balanced between types of anomalies (labeled as positive) and normal events (labeled as negative) [11].

2.2. Summary of the related work

The literature discussed so far can be used in identifying the current practices in video forensics and features that are used in the video forensics research. These days, video forensics is a very active research field, and in one way or the other, video forensics will play a vital role in preventing violence before it occurs to maintain the law and security in society [26–31].

3. Methodology

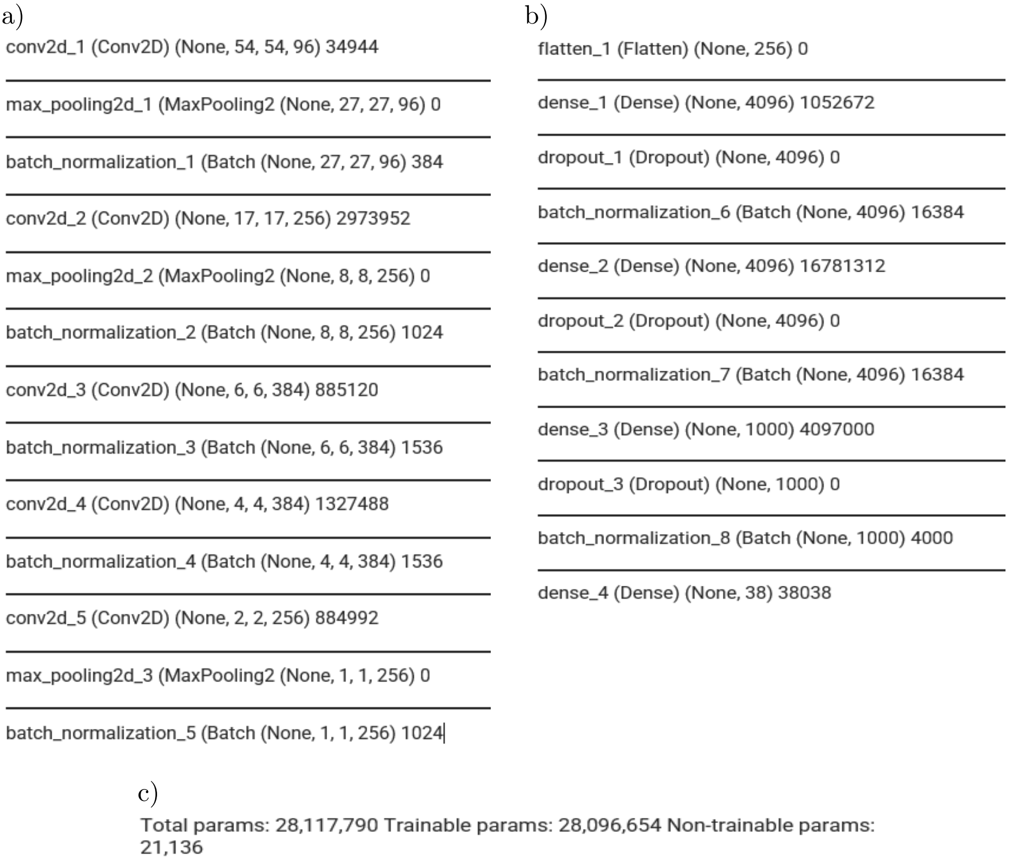
This section discusses the customized neural network used in this study. Also, the architecture of YOLO for object detection, its prerequisites, along with advantages and disadvantages, are discussed. These two architectures are used to compare the performances of object detection.

3.1. The customized CNN architecture

The CNN architecture is proposed to identify guns and masks. The proposed architecture consists of five convolutional layers, three maximum pooling layers, four fully connected layers, and, finally, batch normalization applied after each layer except the output layer. Thus, eight batch normalization layers are used. Batch normalization is very essential for stabilizing the learning within each unit of the deep learning network by standardizing the corresponding mean and variance. Due to these layers, the performance of CNN will be optimized. Besides these layers, dropout layers are also added to the proposed CNN architecture to prevent the overfitting of the model. Various parameters used for building the proposed CNN architecture are presented in Table 1, and the architecture of the proposed CNN model is presented with layerwise information in Fig. 1.

Table 1. The details of the parameters considered for the proposed CNN architecture.

Parameter	Details
Optimizer function	Stochastic Gradient Descent
Loss function	Categorical Cross-Entropy
Learning rate	0.001
Dropout	0.3
The activation function in intermediatory layers	ReLu
The activation function in the final layers	Sigmoid



a)

```
conv2d_1 (Conv2D) (None, 54, 54, 96) 34944
max_pooling2d_1 (MaxPooling2 (None, 27, 27, 96) 0
batch_normalization_1 (Batch (None, 27, 27, 96) 384
conv2d_2 (Conv2D) (None, 17, 17, 256) 2973952
max_pooling2d_2 (MaxPooling2 (None, 8, 8, 256) 0
batch_normalization_2 (Batch (None, 8, 8, 256) 1024
conv2d_3 (Conv2D) (None, 6, 6, 384) 885120
batch_normalization_3 (Batch (None, 6, 6, 384) 1536
conv2d_4 (Conv2D) (None, 4, 4, 384) 1327488
batch_normalization_4 (Batch (None, 4, 4, 384) 1536
conv2d_5 (Conv2D) (None, 2, 2, 256) 884992
max_pooling2d_3 (MaxPooling2 (None, 1, 1, 256) 0
batch_normalization_5 (Batch (None, 1, 1, 256) 1024
```

b)

```
flatten_1 (Flatten) (None, 256) 0
dense_1 (Dense) (None, 4096) 1052672
dropout_1 (Dropout) (None, 4096) 0
batch_normalization_6 (Batch (None, 4096) 16384
dense_2 (Dense) (None, 4096) 16781312
dropout_2 (Dropout) (None, 4096) 0
batch_normalization_7 (Batch (None, 4096) 16384
dense_3 (Dense) (None, 1000) 4097000
dropout_3 (Dropout) (None, 1000) 0
batch_normalization_8 (Batch (None, 1000) 4000
dense_4 (Dense) (None, 38) 38038
```

c)

```
Total params: 28,117,790 Trainable params: 28,096,654 Non-trainable params:
21,136
```

Fig. 1. Details of the proposed CNN architecture, a) the structure of convolutional layers, b) the structure of fully connected layers, c) the parameter details of the proposed CNN architecture.

3.2. YOLO architecture

Object recognition is a field that has greatly benefited from the latest advancements in deep learning methodologies. Recently, scientists have developed several methodologies, including YOLO, SSD, Mask RCNN, and RetinaNet, for object recognition. These days, YOLO stands out from many other methodologies in the field of object detection in terms of performance. So, YOLO architecture is considered a state-of-the-art (SOTA) architecture. Previously, fast R-CNN was one of the SOTA architectures used for object recognition; yet, this architecture has its drawbacks. To address them, the YOLO methodology was proposed to resolve the issues encountered in the object recognition models that detect different objects in the images or videos. Redmon *et al.* [25] in 2015 proposed the YOLO architecture. It is an object identifier detecting an object using characteristics learned by a deep CNN.

Object recognition is one of the conventional computer vision tasks in which identifying what and where – especially the type of object and the location of the object in the provided input image is conducted [26]. Object recognition is more complicated than its categorization, which also recognizes objects yet does not state the object's location. Moreover, the categorization of pictures does not work with more than one object [27]. YOLO is recognized for its great accuracy and the ability to drive in real-time. In the YOLO methodology, detection of the entire image is conducted in a single forward propagation throughout the neural network. It splits the input image into grids and each grid cell predicts the bounding boxes that highlight an object in an image following non maximal suppression. YOLO forecasts multiple bounding boxes and class probabilities for these boxes all at once using a single CNN. YOLO trains entire images and enhances recognition performance immediately. This methodology has several advantages over other techniques of object recognition. It is extremely fast. During training time and test time, this methodology examines the whole image. Therefore, it automatically integrates class and image data. This approach provides whole detection of objects; thus, the method surpasses other recognition methods when used for person detection in natural images and artwork.

YOLO differs from all previous approaches as it considers image recognition as a regression (object detection is framed as regression problem instead of classification as in all prior work where object detection uses classifiers to perform detection), and enables a single CNN to fulfill all of the aforementioned objectives. Combining all individual tasks in one network offers the following advantages:

Speed: This method is incredibly quick since it uses a unique CNN to identify objects compared to its competitors. The convolution only happens once on the whole input image to obtain the detections.

Fewer background faults: This method performs across the entire image instead of parts, and it encrypts background data based on the classes and their images. The prediction of background regions as objects produces fewer faults since such an approach examines the whole image and works globally.

Highly generalizable: This methodology learns generalized representations of objects suitable for use and less likely to break down when used in new domains and unexpected inputs [25].

YOLO's present methodology is built and tested on PASCAL VOC recognition datasets as a CNN. The proposed system data flow diagram is depicted in Fig. 2. There are 24 layers of convolution, followed by 2 fully linked layers. Based on their purpose, the layers are divided as follows:

- Initial 20 layers of convolution accompanied by an average pooling layer and a fully linked layer is pre-trained on the popular dataset of Imagenet with 1000 classes.
- Input image with resolution of 224 × 224 is used for pre-training for classification.
- The network is composed of 1 × 1 reduction layers and 3 × 3 convolutional layers.
- The network is trained for object detection in the last four convolutional layers accompanied by two fully linked layers.
- Object recognition needs additional granular information so that the data set resolution is increased to 448 × 448.
- The last layer predicts class probability and bounding box coordinates.

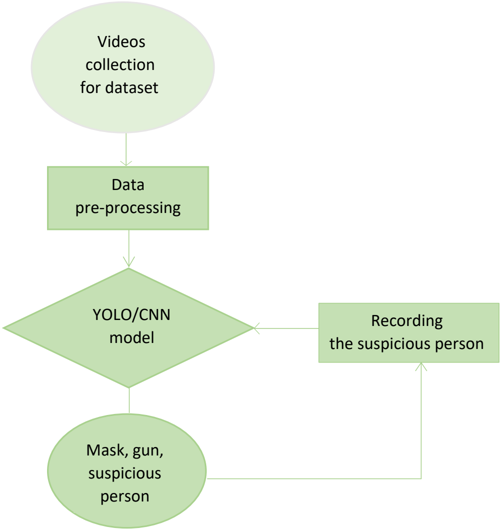

Videos collection for dataset
Data pre-processing
YOLO/CNN model
Recording the suspicious person
Mask, gun, suspicious person

Fig. 2. The proposed system flow diagram.

- The last convolutional layer is activated via linear function, while the previous layers are activated via leaky ReLU function.
- The input image is of size 448 × 448, and the output is the object in the bounding box.

All the elements detecting an object are united by a single neural network. They anticipate every bounding box using characteristics from the whole picture. Such a network concurrently forecasts every boundary box for an image in all categories, as the image is only seen once. YOLO was learned to detect twenty object categories. The image uploaded into the network is split into a grid of size $S \times S$. It is the responsibility of each grid cell to determine if the center of a 20-categorical object is enclosed in it. If the object center is in the grid cell, it says that the item will be enclosed by various bounding boxes of B. Every grid cell is necessary to create confidence scores for every bounding box besides creating various bounding boxes of B. Bound boxes are used to locate an object that exists in an image or video. Note that only one object may be bounded at a time in a box. The bounding box consists of four parameters: the center coordinates x and y referring to the grid cell's source, and the box's height and width. In general, the grid cell's source is specified in the upper left corner. There are five predictions for each bounding box: $y = [p_c,\, b_x,\, b_y,\, b_h,\, b_w]$,

$$\text{Box with confidence score: } p_c = p_0 \times \text{IOU}. \tag{1}$$

The confidence score of the bounding box indicates the model's confidence, and the accuracy of the bounding box it creates is in the object's prediction. (p_c) is calculated by multiplying the probability (p_0) in the box in which the object is identified and the cross-section or IOU (intersection-over-union) between the predicted box and the ground truth.

b_x and b_y indicate the center of the object coordinates, b_h and b_w indicate the bounding box's height and width, respectively. In addition to the bounding box, each grid cell carries a "C" class probability, that is, the conditions for the object contained in the bounding box given as $Pr(\text{class}|\text{object})$.

A grid cell can yield probability in the "C" class, but only one set of class probabilities per cell is predicted as the network is trained to recognize those classes of "C". The grid cell in YOLO can anticipate only one object such as a cat, dog, automobile, etc. It cannot predict several things, for example, a grid cell with cat and dog in it simultaneously. This is one of the main restrictions of YOLO. It cannot locate and recognize more than one object per cell grid. The output size produced by the grid of $S \times S$ is, therefore, $S \times S \times (B \times 5 + C)$.

The loss function consists of multiple sums of squared errors. This function is improved throughout training to enhance the network's predictions. The sum of squared error offers the advantage of using and optimizing different loss functions more readily. It should be noted that these functions are generally selected or constructed to provide easy optimization. For instance, a cross-entropy loss function is a smoother and convex negative logarithmic function. These two characteristics enable optimizing the learning time and outcomes simpler and faster. The generalized YOLO architecture is presented in Fig. 3.

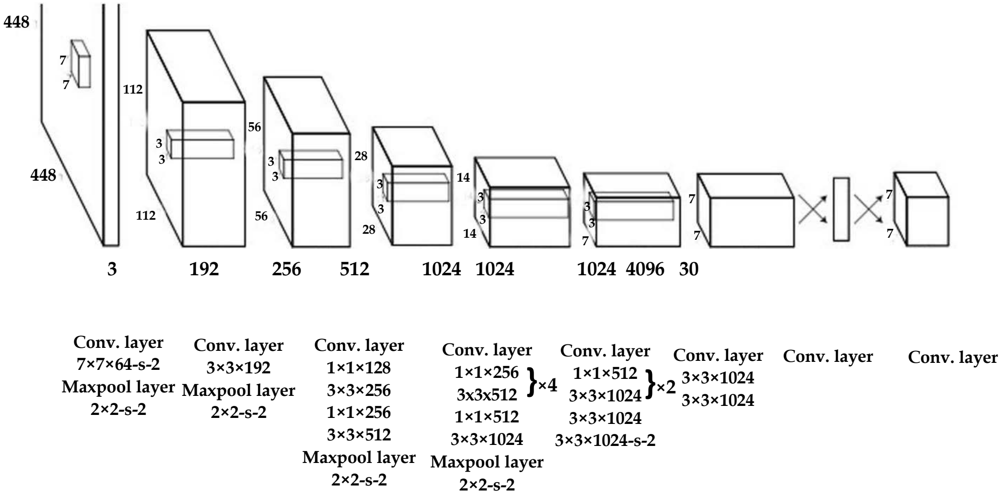


FIG. 3. The generalized YOLO architecture [25].

The above-mentioned formula has two aspects that are worth mentioning:

- Differential weights are used for confidence prediction of boxes that contain objects and boxes that do not contain object during training. The loss function only penalizes classification error in the boxes containing an object. There are no penalties for boxes without an object. The conditional likelihood of an object existing in a box, therefore, plays an essential role in determining the binding factor for the loss function.
- The square root of the height and width of the predicted boxes is used to detect large as well as tiny items. This demonstrates plainly that both are treated without any distinction, which is not optimal in the loss function.

It should be noted however that the loss function only penalizes categorization errors if the item is in the grid cell, and it only penalizes the binding coordinate error if the prediction has the greatest IOU in that grid cell, i.e., the ground truth label.

4. Dataset and results discussion

This section mainly discusses the dataset, hardware, and software details used to develop the proposed methodology. The results obtained through this methodology are presented, discussed and compared to the results obtained by the customized neural network and the YOLO network.

4.1. Dataset description

The dataset used for the proposed methodology is collected from various open sources as well as self-generated from [32–34]. The considered data have a size of 1346 images related to the gun. Similarly, the data collected for masks is collected from various open sources such as YouTube. The considered mask data have a size of 1043 images. Sample images of these two datasets are presented in Figs. 4 and 5.

FIG. 4. Samples of images from the gun dataset.

FIG. 5. Samples of images from the mask dataset.

The proposed work is implemented on a high architecture system whose operating system is Windows10, the processor is Intel® Core$^{\mathrm{TM}}$ i7-9750H with a base frequency 2.60 GHz and a max turbo frequency 4.50 GHz, and GPU was NVIDIA GeForce RTX 2060 of 6 GB.

4.2. Results discussion

The proposed work can be categorized into two phases. The first phase deals with detection of guns and the second deals with detection of masks. Both these sections are implemented using a customized CNN and the YOLO network. The identified gun and mask images are shown in Figs. 6 and 7, whereas the suspicious person carrying a gun is identified and shown in Fig. 8. “M”, “NM”, and “G” indicate “masked”, “not masked”, and “having hun”, respectively. The results for guns and masks are presented in Tables 2 and 3, respectively.

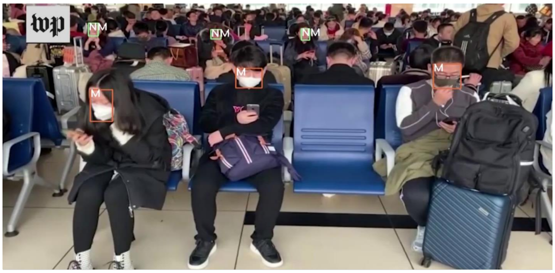


Fig. 6. The output image for detection of masks.

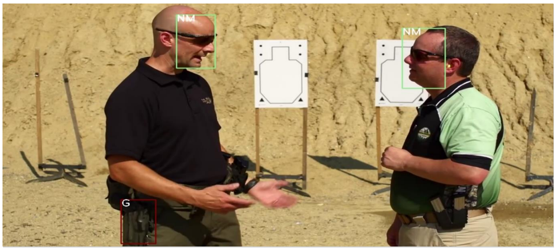


Fig. 7. The output image for detection of guns and masks.

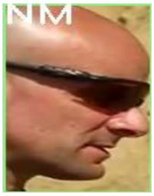


Fig. 8. The output image for detection of a suspicious person.

Table 2. The performance metrics for detection of guns.

Performance metrics [%]	YOLO architecture	Customized CNN
Accuracy	100.0	61.54
Precision	100.0	57.00
Recall	100.0	28.57
F1-score	100.0	44.44

Table 3. The performance metrics for detection of masks.

Serial number	Performance metrics	Accuracy [%]
1.	YOLO architecture	100
2.	Customized CNN	61.50
3.	CNN-LSTM [23]	61.00
4.	CNN [24]	97.54

In the above results, one can understand that the customized CNN is not even on-par comparison with the YOLO architecture. This indicates that the YOLO architecture performs better than other customized CNN architectures. The outcomes of this technique are shown and analyzed, and the results of the designed neural network and the YOLO network are compared. This study may be divided into two parts. The first part focuses on gun detection, while the second part focuses on mask detection. Both of these parts use of a modified CNN and the YOLO network. This due to two reasons: the initial weights are not initialized randomly, and the second is its architecture, as this architecture already includes a pre-training model on the Imagenet dataset [35].

5. Conclusion and future scope

A framework was proposed in this article as a part of digital image forensics. The presented framework has two segments: detection of guns and detection of masks using the YOLO architecture as well as the customized CNN architecture. The latter is strongly surpassed by the YOLO architecture due to uninitiating the weights randomly, and a pre-training section of its architecture. The YOLO architecture aids the user in identifying various objects in the image, and in this scenario, those are guns and masks. which is one more advantage to this architecture. The proposed framework only dealt with images as input, and this can be further extended by using videos as input. The performance of the YOLO architecture can also be improved by using a larger dataset. A larger dataset can be developed by rigorous collection from various sources, and if not, soft computing techniques become handy in improving the size of the dataset. The study can be further extended in the future. The proposed models can be implemented on other datasets such as banking and financial frauds, smart healthcare imaging systems for anomaly detection in various diseases and tumors, image processing systems, digital media, etc.

References

1. G. Gilboa, N. Sochen, Y.Y. Zeevi, Image enhancement and denoising by complex diffusion processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(8): 1020–1036, 2004, doi: 10.1109/TPAMI.2004.47.

2. S. Park, S. Yu, M. Kim, K. Park, J. Paik, Dual autoencoder network for retinex-based low-light image enhancement, *IEEE Access*, **6**: 22084–22093, 2018, doi: 10.1109/ACCESS.2018.2812809.

3. W. Fan, K. Wang, F. Cayre, Z. Xiong, Median filtered image quality enhancement and anti-forensics via variational deconvolution, *IEEE Transactions on Information Forensics and Security*, **10**(5): 1076–1091, 2015, doi: 10.1109/TIFS.2015.2398362.

4. C.Y. Li, J.C. Guo, R.M. Cong, Y.W. Pang, B. Wang, Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior, *IEEE Transactions on Image Processing*, **25**(12): 5664–5677, 2016, doi: 10.1109/TIP.2016.2612882.

5. S. Mandal, X.L. Deán-Ben, D. Razansky, Visual quality enhancement in optoacoustic tomography using active contour segmentation priors, *IEEE Transactions on Medical Imaging*, **35**(10): 2209–2217, 2016, doi: 10.1109/TMI.2016.2553156.

6. H. Walker, A. Tough, Facial comparison from CCTV footage: The competence and confidence of the jury, *Science & Justice*, **55**(6): 487–498, 2015, doi: 10.1016/j.scijus.2015.04.010.

7. E. Verolme, A. Mieremet, Application of forensic image analysis in accident investigations, *Forensic Science International*, **278**: 137–147, 2017, doi: 10.1016/j.forsciint.2017.06.039.

8. D. Seckiner, X. Mallett, C. Roux, D. Meuwly, P. Maynard, Forensic image analysis – CCTV distortion and artefacts, *Forensic Science International*, **285**: 77–85. 2018, doi: 10.1016/ j.forsciint.2018.01.024.

9. S. Li, K.K.R. Choo, Q. Sun, W.J. Buchanan, J. Cao, IoT forensics: Amazon echoes as a use case, *IEEE Internet of Things Journal*, **6**(4): 6487–6497, 2019, doi: 10.1109/JIOT.2019.2906946.

10. M. Jerian, S. Paolino, F. Cervelli, S. Carrato, A. Mattei, L. Garofano, A forensic image processing environment for the investigation of surveillance video, *Forensic Science International*, **167**(2–3): 207–212, 2007, doi: 10.1016/j.forsciint.2006.06.048.

11. E.B. Nievas, O.D. Suarez, G.B. García, R. Sukthankar, Violence detection in video using computer vision techniques, [in:] P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, W. Kropatsch [Eds.], *Computer Analysis of Images and Patterns, CAIP 2011, Lecture Notes in Computer Science*, vol. 6855, pp. 332–339, 2011, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-23678-5_39.

12. D. Gowsikhaa, S. Abirami, Suspicious human activity detection from surveillance videos, *International Journal on Internet & Distributed Computing Systems*, **2**(2): 141–148, 2012.

13. A.H. Lone, F.A. Badroo, K.R. Chudhary, A. Khalique, Implementation of forensic analysis procedures for WhatsApp and Viber Android applications, *International Journal of Computer Applications*, **128**(12): 26–33, 2015, doi: 10.5120/ijca2015906683.

14. J. Kamenicky *et al.*, PIZZARO: Forensic analysis and restoration of image and video data, *Forensic Science International*, **264**: 153–166, 2016, doi: 10.1016/j.forsciint.2016.04.027.

15. Q. Wan, K. Panetta, S. Agaian, A video forensic technique for detecting frame integrity using the human visual system-inspired measure, [in:] *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, pp. 1–6, 2017, doi: 10.1109/THS.2017.7943466.

16. M.F.E.M. Senan, S.N.H.S. Abdullah, W.M. Kharudin, N.A.M. Saupi, CCTV quality assessment for forensics facial recognition analysis, [in:] *2017 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence*, pp. 649–655, 2017, doi: 10.1109/CONFLUENCE.2017.7943232.

17. H. Kaur, K.R. Choudhary, Digital forensics: implementation and analysis for Google Android framework, [in:] I.M. Alsmadi, G. Karabatis, A. Aleroud [Eds.], *Information Fusion for Cyber-Security Analytics*, Springer International Publishing, pp. 307–331, 2017.

18. Y. Chen, X. Kang, Z.J. Wang, Q. Zhang, Densely connected convolutional neural network for multi-purpose image forensics under anti-forensic attacks, [in:] *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pp. 91–96, June 2018, doi: 10.1145/3206004.3206013.

19. G. Singh, K. Singh, Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation, *Multimedia Tools and Applications*, **78**(9): 11527–11562, 2019, doi: 10.1007/s11042-018-6585-1.

20. P. Zhou, Q. Ding, H. Luo, H. Hou, Violence detection in surveillance video using low-level features, *PLoS ONE*, **13**(10): e0203668, 2018, doi: 10.1371/journal.pone.0203668.

21. M. Ramzan *et al.*, A review on state-of-the-art violence detection techniques, *IEEE Access*, **7**: 107560–107575, 2019, doi: 10.1109/ACCESS.2019.2932114.

22. J. Horváth *et al.*, Anomaly-based manipulation detection in satellite images, [in:] *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 62–71, 2019.

23. B. Peixoto, B. Lavi, J.P.P. Martin, S. Avila, Z. Dias, A. Rocha, Toward subjective violence detection in videos, [in:] *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8276–8280, 2019, doi: 10.1109/ICASSP.2019.8682833.

24. H. Kaur, N. Jindal, Image and video forensics: A critical survey, *Wireless Personal Communications*, **112**: 1281–1302, 2020, doi: 10.1007/s11277-020-07102-x.

25. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, [in:] *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

26. Kanksha, B. Aman, P. Sagar, M. Rahul, K. Aditya, An intelligent unsupervised technique for fraud detection in health care systems, *Intelligent Decision Technologies*, **15**(1): 127–139, 2021, doi: 10.3233/IDT-200052.

27. S.K. Nanda, D. Ghai, S. Pande, VGG-16-based framework for identification of face-mask using video forensics, [in:] D. Gupta, Z. Polkowski, A. Khanna, S. Bhattacharyya, O. Castillo [Eds.], *Proceedings of Data Analytics and Management. Lecture Notes on Data Engineering and Communications Technologies*, Vol. 91, Springer, Singapore, 2022, doi: 10.1007/978-981-16-6285-0_54.

28. N. Yadav, S.M. Alfayeed, A. Khamparia, B. Pandey, D.N.H. Thanh, S. Pande, HSV model-based segmentation driven facial acne detection using deep learning, *Expert Systems*, **39**(3): e12760, 2022, doi: 10.1111/exsy.12760.

29. A. Kishor, C. Chakraborty, W. Jeberson, Reinforcement learning for medical information processing over heterogeneous networks, *Multimedia Tools and Applications*, **80**: 23983–24004, 2021, doi: 10.1007/s11042-021-10840-0.

30. A. Kishor, C. Chakraborty, W. Jeberson, A novel fog computing approach for minimization of latency in healthcare using machine learning, *International Journal of Interactive Multimedia and Artificial Intelligence*, **6**(7): 7–17, 2020, doi: 10.9781/ijimai.2020.12.004.

31. A. Kishor, C. Chakraborty, Artificial intelligence and internet of things based healthcare 4.0 monitoring system, *Wireless Personal Communications*, 2021, doi: 10.1007/s11277-021-08708-5.

32. Shershah Movie last scene/Captain Vikram Batra, https://www.youtube.com/watch?v=pbkAa7HTLyE.

33. Jai Jawan: Sushant Singh Rajput Heads For A Different Kind Of Shooting, NDTV, 2017, 2021, https://youtu.be/0sREKfYyiWE.

34. The Basics of Gun Handling | Shooting Tips from SIG SAUER Academy, NSSF – The Firearm Industry Trade Association, 2014, https://youtu.be/r6Nv74nvEWg.

35. S.K. Nanda, D. Ghai, Future of Video Forensics in IoT, [in:] S.L. Tripathi, S. Dwivedi [Eds.], *Electronic Devices and Circuit Design: Challenges and Applications in the Internet of Things*, Chapter 8, pp. 113–133, CRC Press, 2022.

Received December 25, 2021; revised version February 27, 2022;
accepted March 3, 2022.