Contents lists available at ScienceDirect

Measurement:Sensors

Journal home page:www.sciencedirect.com/journal/measurement-sensors

Remote Monitoring System Using Slow-Fast Deep Convolution Neural Network Model for Identifying Anti-Social Activities in Surveillance Applications

Edeh Michael Onyema[1], Sundaravadivazhagn Balasubaramanian[2], Kanimozhi Suguna S[3], Celestine Iwendi[4], B. V. V. Siva Prasad[5], Edeh Chinecherem Deborah[6] , Nwobodo Lois Onyejere[7]

[1]*Department of Vocational and Technical Education, Faculty of Education, Alex Ekwueme Federal University, Ndufu-Alike, Abakaliki, Nigeria; Department of Mathematics and Computer Science, Coal City University, Enugu, Nigeria.* mikedreamcometrue@gmail.com
ORCID: https://orcid.org/0000-0002-4067-3256

[2]*Department of Information Technology, University of Technology and Applied Sciences, AlMussanah Oman.* sundaravadi@act.edu.om

[3]*Department of Computer Applications, ArulmiguArthanareeswarar Arts and Science College,*
Tiruchengode, Tamil Nadu, India
dr.kanimozhisuguna@gmail.com

[4]School of Creative technologies, University of Bolton, United Kingdom, celestine.iwendi@ieee.org

[5]*Department of CSE, School of Engineering, Malla Reddy University, Hyderabad, India*
drbvvsivaprasad@gmail.com

ORCID: https://orcid.org/0000-0001-8650-3984

[6]*Faculty of law, Enugu State University of Science and Technology, Nigeria*
chinecheremdeborah@gmail.com

[7]Department of Computer Engineering, Enugu State University of Science and Technology, Nigeria.
lois.nwobodo@esut.edu.ng

ARTICLEINFO

Keywords:
Deep Learning
Convolutional Neural Network
Video Processing
Object Detection and Recognition
Abnormal Activity Detection
Surveillance Monitoring

ABSTRACT

Remote monitoring is the process that monitors and observes information from a distance utilizing sensors or electronic types of equipment. Remote monitoring is used in real-time applications like traffic, forest, military, shops, and hospitals to determine abnormal activities. Earlier research has done video processing methods based on computer vision techniques, but the computational complexity regarding time and memory is high. This paper designs and implements a novel Slow-Fast Convolution Neural Network (SF-CNN) to identify, detect, and classify abnormal behaviours from a surveillance video. The proposed CNN architecture learns the video frames automatically, obtains the most appropriate properties about various objects' behaviour from a large set of videos. The learning process of SF-CNN is carried out in two ways, such as slow learning and fast learning. The slow learning process is enabled when the frame rate is less, and the rapid learning process is enabled when the frame rate is high. Both the learning processes learn spatial and temporal information from the input video. Different objects, such as humans, vehicles, and animals, are detected and recognized according to their actions. All the videos have normal and abnormal activities that vary in various contexts. The proposed SF-CNN architecture provides an end-to-end solution to dealing with multiple constraints abnormal movements. The experiment is carried out on several benchmark datasets, and the performance of the SF-CNN architecture is evaluated. The proposed approach obtained 99.6% of accuracy, which is higher than the other existing techniques.

1. Introduction

Anit-Social activities are increasing day by day in innumerable fields. Theft, illegal, and other outlawed activities are considered anti-social and must be identified immediately, and protect the area as quickly as possible. It reduces the loss of data, things, and human death. The medical industry, forest, research centers, aerospace, vehicle stations, malls, most extensive building, etc., are some areas that need to be surveillance to avoid abnormal activities. The surveillance system records the activities using CCTV cameras and continuously records them as videos. The output of the surveillance system is a video. The video is processed, and the abnormal activity is identified using the object detection and recognition method. The movement of the objects is classified as normal or abnormal. Today, there are ten reasons business people need a video surveillance system. They are: Resolve Conflicts, Increasing employee productivity, Reducing Theft, providing Better experience, Real-Time Monitoring, Enhancing safety, Digital storage, Evidence making, Access control, and Business savings.

Though the surveillance monitoring system process is similar, the applications are different. The cameras' capacity and configuration and the surveillance systems have been changed based on the application. Some surveillance applications are given in Figure-1, which shows surveillance in the office, road, official building, and backside of a house. Different intrusion detection systems have been proposed for security provision earlier, but it detected after the abnormal event. There is no methodology, which can stop the strange activities automatically or manually. For preventing or controlling abnormal movements, location information is required with complete knowledge about the geo-region. It helps to identify the abnormal activity earlier and provides better security for the particular surveillance area. One of the best solutions to enhance the existing security is surveillance monitoring.

**Correspondingauthor: Edeh Michael Onyema*
*E-mailaddress:*mikedreamcometrue@gmail.com (EM Onyema).

Office [1], Road [2], Building [3], House [4]

Fig-1. Various Applications of Surveillance System

This paper aims to design and implement a deep learning-based abnormal activity detection using a convolution neural network to provide a better solution. Deep learning is a machine learning technique that enables computers to exhibit human-like behaviours that humans do as their second nature. For example, a particular deep learning technique is used to detect the street light signs, and another is taught to recognize a pedestrian from a cat, car, etc. It is also used to correct grammar, spellings, repetitive words, punctuations, and more in the given texts and automatically generates a new text with a nil error. More progress is achieved in various fields today than earlier [1]. The deep learning model is used for a computer to learn to do classification tasks over different kinds of data like voice, images, texts. The level of performance in deep learning exceeds human efficiency hence reaching state-of-the-art accuracy. Multi-layered neural network architecture and an enormous amount of labelled data are used to train these models.

Because of its accuracy, deep learning technology was widely used in a variety of applications at such a higher and more critical level [2]. These techniques are efficient in electronics that help to reach user expectations. The recent developments in deep learning outperform humans, thus used in safety and critical applications such as robotics, AI cars, image caption generation, etc. Though deep learning was speculated for the first time in the 1980s, its usage has become very popular only recently due to the following reasons:

Deep learning needs a massive amount of labeled data. i.e., in the case of Autonomous vehicles (driverless cars), to train the computer for that crucial job, millions of images and thousands of hours of footage are required.

So far, it is basic knowledge that deep learning needs enormous hardware and substantial computing power. It will require many weeks of deep learning networks to perform efficiently. So, to reduce the training period, the development teams enable, Hybrid Computing rather than cluster or cloud computing, i.e., using GPU's (Graphics Processing Unit) massive parallel processing power to boost up the performance

A deep learning model is also known as a deep neural network, as neural network architecture is used in primary deep learning methods. A deep neural network can contain a maximum of 150 hidden layers, whereas the traditional neural network can accommodate only up to 2-3 hidden layers. In a neural network, the number of hidden layers is denoted by the term "deep." Deep learning models are trained using a large number of labeled and neural networks capable of learning characteristics directly from the data without the need for manual feature extraction [3]. Convolutional neural networks (CNN) are famous deep neural network types. This architecture is most suitable for processing 2D data like images as CNN uses 2D convolution layers and the input data that convolutes learned features.

CNN does not require manual feature extraction to classify images, and therefore identifying features is eliminated. Feature extraction is done directly from the images using CNN and the structure of the neural model is represented in Figure 2. A collection of images is used to train the network, and the relevant features are learned simultaneously, eliminating the need for pre-training of the relevant features. In computer vision tasks like object classification, high accuracy is attained using automated feature extraction. An enormous number of hidden layers is required for the CNNs to identify various features of an image. Hidden layers raise the complexity of the image features that have been learned. For instance, the first and second hidden layers will identify the edges and complex shapes to recognize an object.

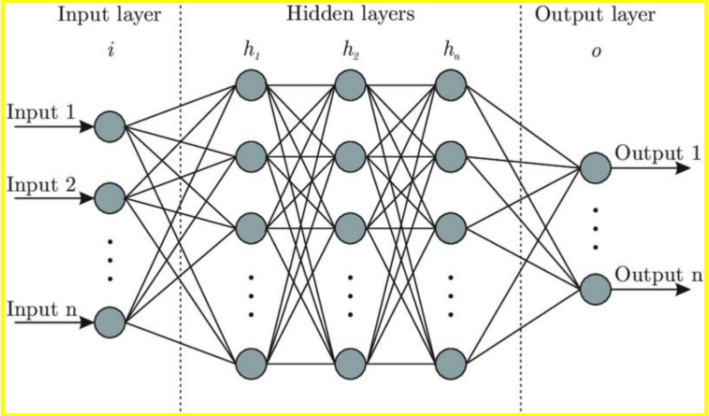


Fig-2. Structure of a Neural Networks

1.1 Machine Learning versus Deep Learning

Machine learning and deep learning are often interchangeable, but deep learning is specialized with human-like artificial intelligence, making it more efficient than machine learning. In machine learning, algorithms are manually fed to parse data. It learns the given data to derive images, videos, texts, and other information used to design a model that categorizes the object in the data. The model that performs a function uses its data and gets better at doing it over a period of time in machine learning.

The deep learning model uses a layered structure of algorithms known as Artificial Neural Network (ANN), inspired by the human brain's Neural System. ANN can analyze data continuously and draw precise decisions on its own, just like a human brain would do. Deep learning keeps improving with an increase in the size of the data and the sample layered architecture is given in Figure 3. Machine learning also provides various models and techniques, from which one can choose based on the application they need to sort. For the successful deep learning technique, an enormous amount of data, i.e., millions of images and hours of videos, is required to train the model. A Graphics Processing Unit is used for the fast processing of data. If the former and latter are not available, it is best to use machine learning algorithms rather than deep learning.

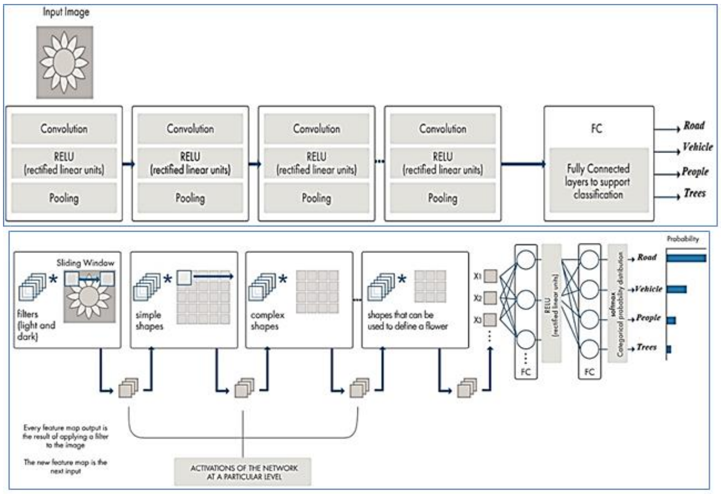

Input Image
Convolution
RELU (rectified linear units)
Pooling
Convolution
RELU (rectified linear units)
Pooling
Convolution
RELU (rectified linear units)
Pooling
Convolution
RELU (rectified linear units)
Pooling
FC
Fully Connected layers to support classification
Road
Vehicle
People
Trees
Sliding Window
filters (light and dark)
simple shapes
complex shapes
shapes that can be used to define a flower
X1
X2
X3
FC
RELU (rectified linear units)
FC
softmax
Categorical probability distribution
Probability
Road
Vehicle
People
Trees
Every feature map output is the result of applying a filter to the image
The new feature map is the next input
ACTIVATIONS OF THE NETWORK AT A PARTICULAR LEVEL

Figure 3: Layered Architecture of the proposed System

The paper's contribution is that it uses two learning processes to learn the video frames to increase object recognition accuracy. They are slowly learning and fast learning. The paper's novelty is that the deep learning model understands the video data and analyzes it according to the frame rate. If the frame rate is low, it analyses the frames slowly and obtains the spatial semantics. If the frame rate is high, it analyses the structures and gets the temporal semantics. Thus this proposed deep learning model learns the spatial and temporal information for video processing and objects recognition.

1.2 Related Works

Fig-3. A Sample Layer Architecture of CNN - Training and testing

[illegible] ıct a deep study on various existing methods, to understand the issues and challenges faced by the earlier works. For example, authors in [4] stated that pervasive computing uses CCTV cameras for video surveillance. So, it is considered a recent research work where the video data is analyzed using machine learning approaches. The devices used for video capturing are very common throughout the world, and the human resources used for video analysis are minimal but expensive. In most cases, surveillance cameras are used for surveillance monitoring. The authors in [5] stated that some human factors like tiredness and fatigue lead to lousy tracking. Also, humans working with CCTV monitoring suffer monotony. The reason is that in several cases, strange or unusual events occur rarely. The authors in [6] said that various kinds of methods like anomaly detection, abnormal detection [7], outlier detection [8] are big topics used in different real-time areas like intrusion detection in-network, medical diagnosis, marketing, and automatic surveillance.The authors in [9] said that providing a security algorithm could break the anomaly detection. Most of the research works focused on delivering security algorithms to avoid anomalies. Some of the research works focused on detecting abnormal events. The authors in [10] used the optical flow method and the Gaussian mixture model to identify and detect strange activities occurring in crowd scenes. The authors in [11] presented some significant research methods associated with abnormal behavior detection in crowd scenes. It can also be used in vehicle detection during traffic conditions in a congested area.

Thermal cameras were employed in several of the research studies to monitor surveillance. Simultaneously, the authors of [12] explained that its videos' sensitivity and quality are lower than those of the other videos. Thermal videos, according to the scientists, seem to be noisy and also have poor visual quality. Feature extraction, according to the authors in [13], is a method for extracting motion as well as spatial information from video frames. According to the authors in [14], the high - and low features are strongly integrated in inferring unusual behaviors, which aids the programmer in quickly identifying weird activities and analysing any complicated behaviourswith in video. The authors in [15] motivated to detect multiple objects in a video based on feature extraction and classification. The authors obtained 90% accuracy in object detection from the experiment, which further improved. Numerous researchers are still focusing on and researching surveillance monitoring-based applications. In [26], the authors have proposed a Cybrog intelligence for designing an intrusion detection system in a cloud network traffic and various security challenges were projected in [27]. The authors in [28] have implemented SF-CNN model for detecting the suspicious activity with the support of surveillance applications. RelativeNas is another advanced CNN model proposed by the authors in [29] which describes the performance of combined fast and slow learners. This model aids in searching of objects with reduced cost and error rates. Another application of SF-CNN is presented in the article [30] in which the model is tested for performing action recognition through detection procedures. The detailed description about this two-pathway CNN model is elaborated in [31]. However, SoFTNet which is a concept controlled DCNN and Attention Slow-Fast Fusion Networks presented in [32] and [33] are certain other applications that uses this SF-CNN model.

1.3 Limitation and Motivation

Though the accuracy of the object detection and classification regarding abnormal activities needs to be improved, some semi-automatic methods consumed more computational time, increasing the cost. Most of the earlier research works have been proposed for specific applications in surveillance monitoring. Also, the video learning process can extract a particular type of feature from the video, which cannot bring high accuracy in the recognition. The learning rate of the entire dataset is a variable, where it is changed during the program execution to tune the results. It takes more time complexity in video processing models. But the industry needs a fast and accurate video recognition model for surveillance applications. Hence, this paper aims to design and implement a slow-fast deep CNN model to learn the video frames with different learning rates, which retains various spatial and temporal features for automatically improving object recognition accuracy.

2. Deep Learning-Based Object Detection and Recognition

From the above discussions, it is clear that the proposed deep learning approach is highly suitable for video processing, object detection and recognition, pattern recognition, and speech recognition applications. This paper recommended using the Convolution Neural Network in Deep Learning for object detection and abnormality identification. Initially, the input video is obtained from the system/PC connected with the application's wired or wireless CCTV cameras. The video file is automatically stored on the PC if it is a wired connection. An intermediate device, such as a router, is used to send the video to the appropriate PC. In this work, some assumptions are established, which aid in a thorough understanding of the entire procedure.

1. Any number of surveillance cameras can be connected to the application network.

2. The PC interlinked with the camera is installed with the MATLAB software, and our proposed algorithm is executed automatically when the user clicks the run button.

3. Our proposed approach knows the video file's location, the video files' name, and the other meta information.

4. The model learns the video frame rate while streaming and act accordingly.

5. The spatial with temporal features was extracted automatically from the video frames to identify the object and its motions.

6. The abnormal activity occurs within midst of a video rather than at the start or end.

The deep learning approach is used in the same work to detect and identify aberrant actions in such a surveillance video. Deep learning is well recognised to be inspired by several neural network architectures with a deep framework for learning features while representing data. A specific neural network model comprises the input layer, output layer, different kinds, and a more significant number of hidden layers. Deep learning has a various number of large size networks with a more substantial number of layered networks. One famous and most applicable deep learning network is Convolutional Neural Network (CNN). This paper is aimed to use a CNN architecture for identifying abnormal activities. CNN learns and classifies the features automatically from the video/image data. This paper also analyses and evaluates the proposed CNN architecture's performance regarding the human, vehicle, and animal behaviours involving various backgrounds, which kind of multiple data processing is not carried out before.

2.1 Proposed Approach

The proposed CNN architecture is explained in this section. The video V is divided into frames (images) F, in which various objects and their activities are normal and abnormal. Some of the specific abnormal activities are different activities than the usual activities. For improving the efficiency of video/image processing, the images are initially applied for preprocessing using a moving 3×3 average filter, which removes the noises occurring in the images. It can be represented as in the Equation (1).

$$y_{ij} = \sum_{k=-m}^{m} \sum_{l=-m}^{m} w_{kl} x_{i+k,j+l} \qquad (1)$$

where the input image is represented as x_{ij}, (i,j) represents the pixels in the image, and y_{ij} represents the output image. Similarly, a linear filter with the size 3×3, is used on Equation (2).

$$(2m+1)\, x\, (2m+1) \qquad (2)$$

Equation (1) having the weights w_{kl} for every k and l from $-m$ to m, equal to 1.

Each video is made up of a greater number of frames and the video processing using DCNN is given in Figure 4. For example, a one-minute video has 100 to 110 video frames. The video frames are extracted from the input videos and called a video sequence by writing a computer program or performing manually. Similarly, one action has many frames where most of the frames have the same objects with the same activities, so the video processing time is a waste. Hence, it is essential to choose the video frames to speed up the process and provide better classification efficiency. For improving the classification accuracy, 30% of the video frames are initially applied for the training process, labeling the objects as "normal" or "abnormal." The number of abnormal activities is less than the normal activities. So, labeling the abnormal activities is sufficient for the classification, reducing the computational time, memory, and complexity. The time complexity can also be reduced by processing the selected frames and other frames labeled as normal. Only the abnormal frames and the labels are stored in a particular database for testing process references during the training process. The abnormal activities are identified as wrong features/odd features, which differ from the normal video sequences. The remaining 70% of the frames are used for the testing process, and the final extracted features are compared with the trained features.

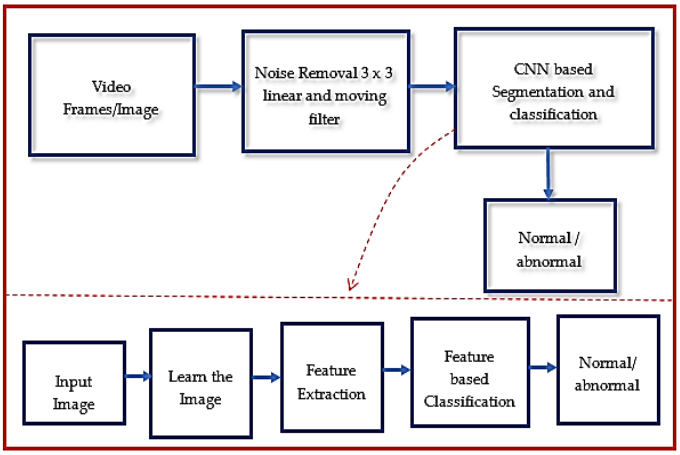
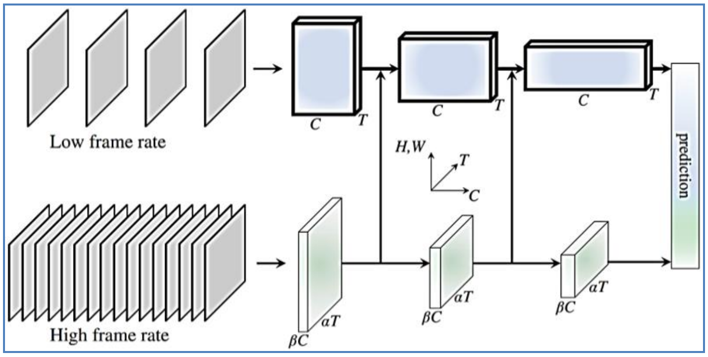
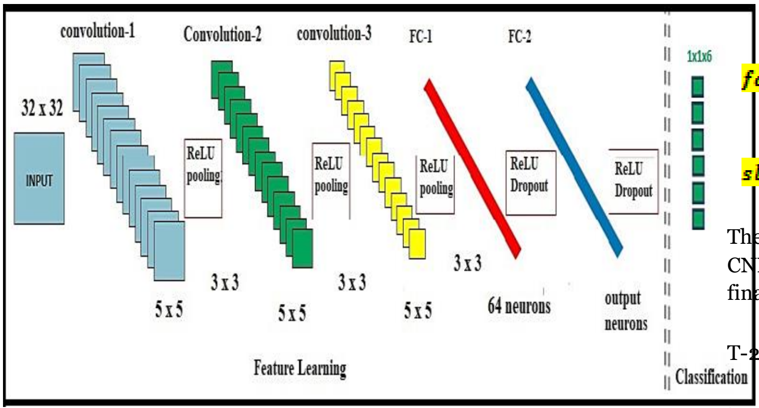

Video Frames/Image
Noise Removal 3 x 3 linear and moving filter
CNN based Segmentation and classification
Normal / abnormal
Input Image
Learn the Image
Feature Extraction
Feature based Classification
Normal/ abnormal

Fig-4. Deep Convolution Neural Network-Based Video Processing

The proposed CNN architecture comprises three essential components: input, output, and hidden layers. The hidden layers are also called convolutional, middle, or feature detection layers. The output layers are also called classification layers, which have two components: fully connected and SoftMax. The size of all the frames is resized as 32 x 32, increasing the time of the training process. The input layers read all the input images and send them to the middle layers. In the middle layer, three modes of operation are carried out: convolution and pooling with rectification. Rectified Linear Unit does this operation (ReLU). This proposed SF-CNN uses three convolutional, two completely connected with one SoftMax layer. The architecture of the proposed SF-CNN to be shown in Figure-5. Figure-5(a) shows how SF-CNN can change the learning method based on the frame rate. Figure-5(b) depicts the complete functionality of the CNN.



(a) Model from [34]



(b)

Fig-5. Proposed Deep Learning Architectures. (a). SF-CNN framework, (b). Functionality of CNN

Some features are extracted by convolution filter and activated in the input image, such as texture, edge, and corner features. These features are the most valuable ones that help identify the actions performed in the frames. The feature values that do not match the overall set represent the abnormal action. In the 1st convolution layer, 32 filters (5 x 5 x 3) are used. In 5 x 5 x 3, 3 represents the input and color images. Symmetrically, a two-pixel-pad is added to consider the image's edges that are also taken for the process. It is more important because it saves the edges from elimination in the CN network. Negative values are changed into zero in the ReLU layer to retain only the positive values for processing. The ReLU layer works faster in training the network than all the layers. ReLU layer always follows the pooling layer with a 3 x 3 spatial pooling region with 2 pixels as strides. Then the data size is down-sampled into 15 x 15 from the initial 32 x 32. All three convolutions with pooling and ReLU layers are repeatedly executed multiple times to extract the number of features and hidden information from the input image as much as possible.

The more significant number of pooling layers can be avoided to prevent the down-sampling data since essential features may be discarded in the earlier stage itself. Once the feature extraction process is completed successfully, the CNN performs the classification process. There are two different layers: 1. fully connected and 2—SoftMax layers used in the network and used for classification. The first fully connected layer comprises 64 neurons obtained from the input image size of 32 x 32, followed by a ReLU layer. Then the second fully connected layer generates the output signals to be classified. The entire CNN network combines the input, middle, and final layers. Finally, the SoftMax layer computes the probability of distribution over each class. The weight of the convolution layer is initialized by a random number 0.0001(standard deviation) for distribution, decreasing the loss during the learning process in the network.

3. SF-CNN

In this paper, deep CNN is incorporated with the Slow-Fast learning method for analyzing the video segments. It comprises two parallel CNN modesl for the same input video-a Slow learning and Fast Learning. Generally, video content contains two different data: static and dynamic. The static data will not be changed or slowly changed, but the dynamic data will continuously be changed (moving objects). According to Figure-5(a), the video frames obtained from fast streaming is input to slow frame rate learner since the slow learner can learn the output of the fast learning. The data format used in the SF learner is written as in the following Equation (3).

$$fastlearning = \{\alpha T, S^2, \beta C\}$$

$$slowlearning = \{T, S^2, \alpha\beta C\} \qquad (3)$$

The Equation (3) is used fused to create the SF learning process. The SF-CNN suggests a different methodology for transforming the data. The final one is the most efficient.

T-2-C (Time-2-Channel): data reshaping and transposing:$\{\alpha T, S^2, \beta C\} \rightarrow \{T, S^2, \alpha\beta C\}$, that is all the α frames as one frame to channel.

TSS (Time-Strided-Sampling): take each α frame as a sample, and it makes: $\{\alpha T, S^2, \beta C\} = \{T, S^2, \beta C\}$

TSC (Time-Strided-Convolution): It performs as a three-dimensional convolution of a 5 x 12 kernel with 2βCoutput channel and α as stride.

TSC (Time-strided-convolution): It performs as a three-dimensional convolution of a 5 x 12 kernel with 2βCoutput channel and α as stride.

Finally, both slow and fast learning are combined as SF learning to perform global pooling operator, efficiently reducing dimensionality. Then integrate both learners, and the output is fed to the FC layer to classify it. The FC layer uses the SoftMax layer to classify the object's behavior as normal or abnormal. The full functionality of the proposed SF-CNN architecture is given in the form of an algorithm. It is implemented and experimented with in any computer programming language, and the results are verified. The pseudocode of the SF-CNN architecture is given as Pseudocode-1.

Pseudocode *1:* Convolution Neural Network

```
Pseudocode_CNN( IMAGE, image-1)
{
Input: Input image is read in the matrix form of h x w x b dimensionFeed the input image into the convolution layer
Select set of parameters and apply filters
After the filter is used, the output received is:  fh x fw x b dimension
Filters, strides, and padding are applied if the dimension is high or required.
for i =1 to N number of convolution layers
    Convolution is applied on the image and apply ReLu function as: f(x) = max (0, x), activation to the input matrix.
    The pooling process is applied to reduce dimensionality.
end for
To flatten the result, apply the output from of the pooling layer to a fully -connected layers.
Output: produce a data in (h-fh+1) x (w-fw+1) x 1 dimension
Finally, the output class is classified using the activation function SoftMax, which classifies the images.
}
```

This SF-CNN can also be called as Dual-mode CNN for understanding the video. In Deep CNN, is used to identify patterns in the images and videos. In this DCNN, each frame of the video is treated as an image and the object or pattern identification is performed over it by considering one frame at a given time. Whereas, the SF-CNN can analyze multiple frames at a given time for collecting the static and dynamic data in the video.

4. Experimental Results and Discussion

In this paper, the algorithm is implemented and verified in MATLAB software, where it contains a built-in CNN module in the Image Processing toolbox. It provides more functionalities and enables various inbuilt algorithms like regression methods, decision-making methods, and it can be chosen for performance verification.

4.1 Dataset Used

In this paper, seven different kinds of datasets are used in the experiment. The full dataset details are given in Table-1, which comprises human activities, vehicle activities, and animal activities. Each frame is an RGB color image with different sizes in the number of pixels. For example, the size of the images from other datasets is 276 x 236, 352 x 240, 360 x 288, etc. Because of the frames' varying sizes, all the images are resized into 32 x 32 pixels. All the frames have a portion of positives and negative samples. The total videos taken for the training and testing process are 100 from the entire dataset. The total number of images used, which were extracted from 100 videos, is 12000 frames. Out of this, 5000 frames are normal, and other frames are considered abnormal.

Table-1. Dataset Information

Dataset	Videos	Total Frames	Frames-Normal	Frames-Abnormal
CMU Graphics Lab Motion [16]	11	2477	1209	1268
UT-Interaction Dataset (UTI) [17]	54	5069	2706	2903
Peliculas Dataset (PEL) [18]	2	368	100	268
Hockey Fighting Dataset (HOF) [19]	12	1800	900	900
Web Dataset (WED) [20]	10	1280	640	640
UCSD-AD [21]	5	600	480	120

Fig-6. Sample Normal and Abnormal Images collected from various datasets

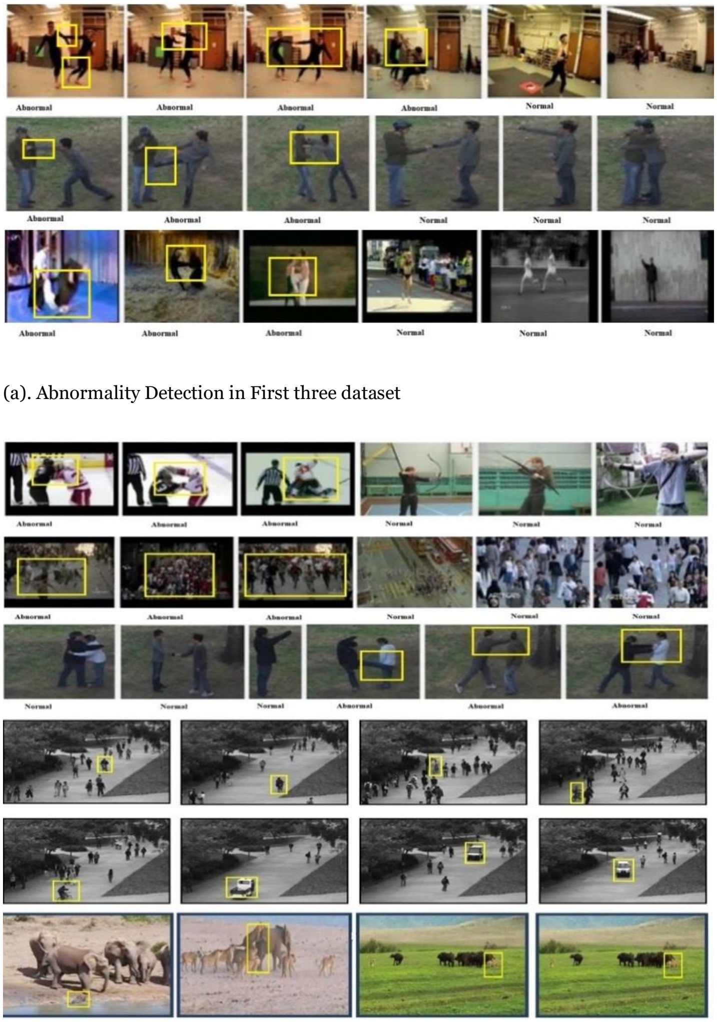
Fig-6. Sample Normal and Abnormal Images collected from various datasets

In the proposed framework, two different experiment stages are carried out. Initially, from experiment-1, the normal versus abnormal classes are classified. Then from experiment-2, all the abnormal classes are classified. For training the network, stochastic gradient descent with momentum method is used. All the parameters inside the network are tuned to obtain all the features which affect the output network. The number of epochs used in the experiment is 10 to 100, and the learning rate is 0.001 to 0.1. for fine-tuning the hyper-parameters of the network. One round of operation, including forwarding and backward passing in the training samples with the learning rate, is considered one epoch.

Deep learning needs more inputs to provide a high level of accuracy. In terms ofhardware requirements, deep learning needs a high-performance GPU. For experimenting with the proposed CNN, MATLAB-2017 software is installed in Intel core i7. The images are represented in binary for the classification process in the experiment. The normal and abnormal images are obtained from various datasets, and the abnormal activities are classified. The normal activities obtained from human-related videos are walking, pointing, hugging, and handshaking. The abnormal activities in human-related videos are kicking, pushing, and punching. The number of classes obtained using the proposed CNN is compared with the classes already labelled in the dataset, and the performance is evaluated. The results are given in terms of images, shown in Figure-6 it every row in the set of all normal and abnormal images as sample images.

From the experiment, using the proposed CNN, the classified normal and abnormal images are given in Figure-7. By comparing the images given in Figure-6 and Figure-7, it is easy to evaluate the performance of the proposed CNN. Abnormal activities in each frame are identified and detected using a binary classification model, and the abnormal activity classification is obtained using CNN. All the abnormal activity is highlighted using a yellow color bounding box in the frames and is shown in Figure-7. Figure-7 shows the abnormality as pushing, kicking, fighting, walking in the wrong path, and the crowd in wrong places, fighting, beating, car in the roadside, cycle, car, truck, and jeeps coming in pedestrian road. The abnormal identification and detection accuracy can be obtained by assigning the appropriate learning rate assigned in the experiment.

In both experiments, normal and abnormal, including all abnormal classes, are more accurate using the proposed CNN, which is understood from Figure-7. The performance of the proposed CNN is high and is evident by comparing both results given in Figure-6 and Figure-7. The abnormal detection accuracy is increased since the testing process is always compared with the training process. Hence, for human interacted classification, the training classes are highly accurate and are used in the testing process.



(a). Abnormality Detection in First three dataset

(b). Abnormality Detection in Four dataset

Fig-7. Abnormality Detection using proposed CNN

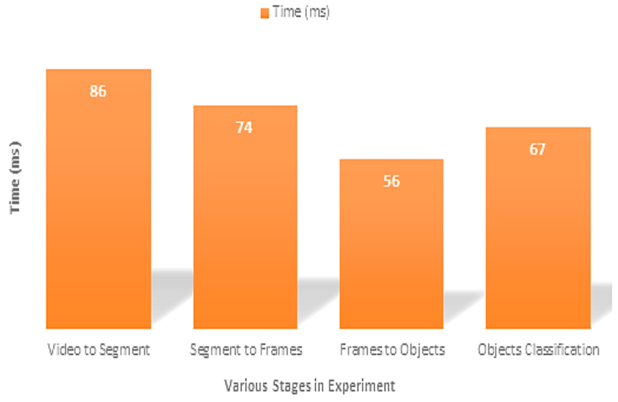


Fig-8. Time Complexity Calculation

The performance of the proposed CNN approach can be evaluated by verifying the time and computational complexity. The experiment's time complexity is initially calculated, and the obtained results are shown in Figure-8. The duration of each phase of the procedure, including video segmentation, frame segmentation, object recognition, and classification, is also displayed. From the results, it is understood that more time is required for video processing than the time needed for image processing. Hence, it is understood that once the video is converted into frames, the time required for object detection and classification is less.

Some errors may occur in the video data because of the sources, power failure or power fluctuations, and converting the video format. It spoils the entire object detection and classification process and degrades image processing performance like object detection and classification. Hence it is necessary to compute the errors present in the input frames in terms of frame-level and pixel-level. Figure-9 shows the error prediction results obtained using the proposed CNN method, including the existing methods as in [22], [23], [24], and [25]. The low frame-level and high pixel-level errors in all the methods are calculated and compared with other methods. From the comparison, it has been found that the proposed CNN method obtained lesser error in frames using pixel levels.

Once the error frames are identified, they are eliminated from the process, improving object detection and classification accuracy. In the experiment, 70% of the frames are applied to the testing process, and the classification accuracy is calculated. The object classification performance obtained from the experiment is given in Table-2. The total number of frames used in the investigation and the correctly classified frames are calculated and are given in Table-2. The total number of frames collected from all seven datasets is 12134, of which 6035 are normal. The remaining 6099 frames are abnormal, predefined already and verified by various earlier research works.

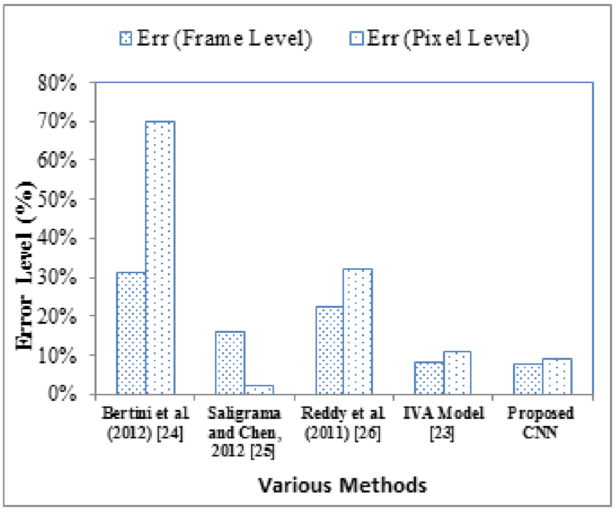
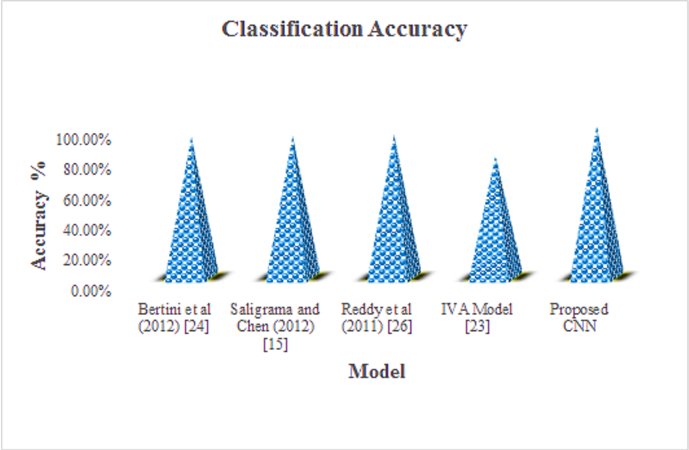


Fig-9. Dataset Based Error Analysis

Table-2. Performance Calculation

Data	Total Frames before the error	Total frames after error	Normal frames	Abnormal frames
Dataset	**12180**	**12134**	**6035**	**6099**
Proposed CNN	**12180**	**12134**	**6034**	**6098**

From Table-2, it is evident that 6034 frames are classified as normal



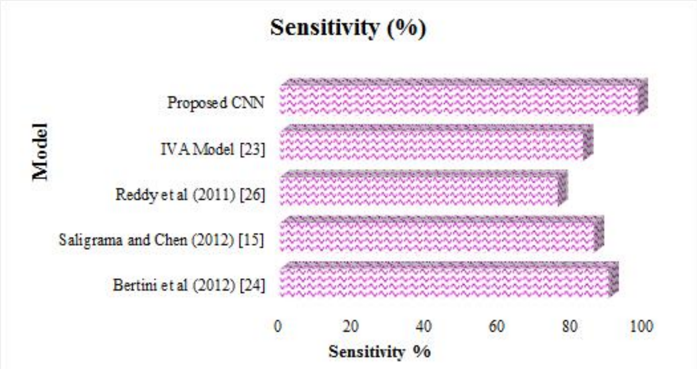
frames out of 6035, 6098 frames as abnormal frames out of 6099 frames using the proposed CNN architecture. For evaluating the performance, the set of all performance factors such as TP-True Positive, TN-True Negative, FP- False Positive, False-Negative, Sensitivity, Specificity, and accuracy are calculated from Table-2. The accuracy is calculated and compared with various existing approaches based on the performance measures. The accuracy is shown in Figure-10. From the comparison results, it is clearly understood that the proposed CNN architecture outperforms the other techniques. The proposed CNN obtained 99.6%, higher than the different approaches discussed in the literature survey.

Fig-10. Classification Accuracy

Table-3. Accuracy Based on Learning Rate

Learning Rate	Max. No. of. Epochs	Dataset Accuracy (%)					
		CMU	UTI	PEL	HOF	WED	UCS-AD
the0.001	10	99.66	54.9	90.38	58.5	89.21	99.9
	20	100	56.55	90.38	100	100	100
	30	100	57.74	90.38	100	100	100
	40	100	70.18	90.38	100	100	100
	50	100	99.15	90.38	100	100	100
0.01	10	100	54.9	90.38	100	100	100
	20	100	99.6	90.38	100	100	100
	30	100	99.75	87.98	100	100	100
	40	100	99.55	100	100	100	100
	50	100	99.8	100	100	100	100
0.1	10	46.64	54.9	9.62	100	100	100
	20	0	0	9.62	100	100	100
	30	0	54.9	0	100	100	100
	40	0	54.9	90.38	100	100	100
	50	0	0	9.62	100	100	100

The classification accuracy depends ona number of epochs with a learning rate. The learning rate of the CNN is changed as 0.001, 0.01, and 0.1 for various epochs varying from 10 to 50. The accuracy is calculated for different datasets under different learning rates given in different epochs. The obtained results are given in Table-3. The accuracy is increased according to the number of epochs. Hence, the proposed approach is executed with an increased number of epochs up to 100, increasing accuracy. According to the findings, the suggested CNN performs better than the other methods currently in use in regards of object recognition, classification accuracy, and time complexity.



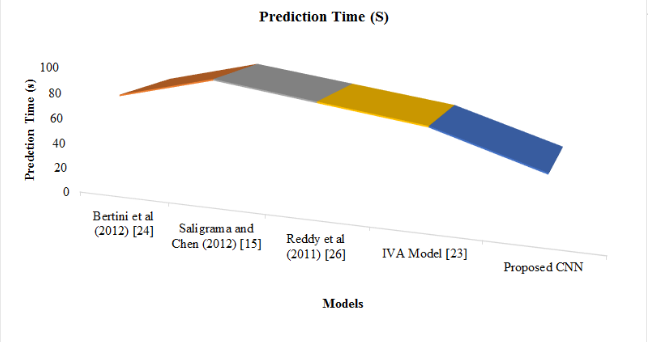


Figure -11. Sensitivity Analysis of the Models

Sensitivity analysis deals with the ability of the model to identify the frames with normal and abnormal conditions of the human activity. The analysis results is presented in the Figure 11, in which the results show that the proposed CNN models shows the highest ability towards the identification of the frames with 98% with a highest difference of 22% with Reddy et al (2011) [26] model.

Figure-12. Prediction Time (s)

Figure 12 is the representation of prediction time taken by the models in identifying or classifying the frames into normal or abnormal state for the occurrence of anti-social activities. The prediction time is measured in milli seconds (ms). The graph describes that the proposed CNN model takes least duration in predicting the abnormal human activities of 50ms, whereas the model utilized by Saligrama and Chen (2012) [15] takes highest duration of 97 ms for a random frame in a given video.

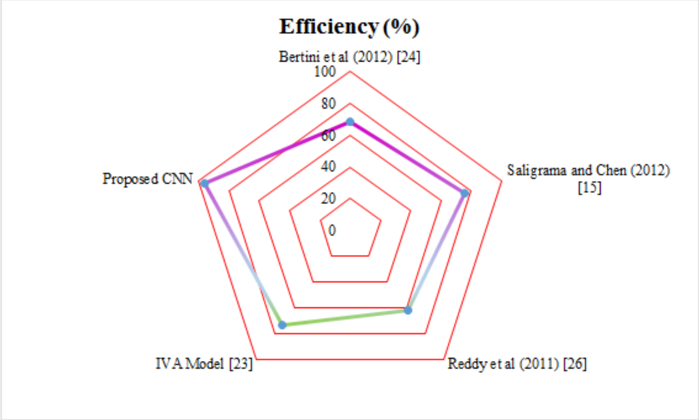


Figure-13. Efficiency of the implemented Model (%)

Efficiency analysis focuses on the measurement of models in achieving speed gain during the overall progress of any given task. In this research, efficiency parameter is considered for evaluating the model in classifying the frames in the video for normal and abnormal human activities. Results of this evaluation is presented in the Figure 13. Among the models considered for evaluation, the proposed CNN model have achieved the highest efficiency which makes a minimum increase of 20% when compared with the existing models.

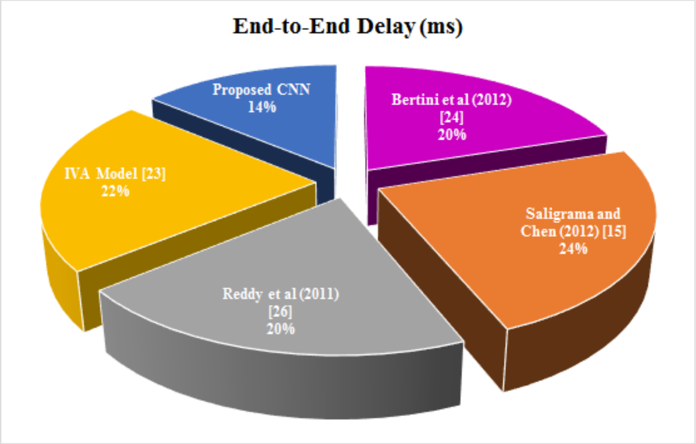


Figure-12. Analysis on End-to-End Delay (ms)

As this research concentrates on the collection of video data from the surveillance camera to the cloud system to perform the analysis on classifying the video frames, it is essential to evaluate the delay caused in data transmission. Delay of data transmission is calculated in milli seconds (ms). Figure 12 depicts this evaluation on delay and from the graph it can be observed that the proposed model shows the least delay of 14% than the other existing models.

5. Conclusion

Objective of this paper's primary goal is the design and implementation of a unique deep learning algorithm for surveillance system anomaly detection. Hence, this paper creates a CNN architecture for education, extracting information, and classifying the abnormality on surveillance of video frames. The specialty of this paper is that abnormality identification is carried out over several datasets. The main motto is to design a common abnormal identification system for any kind of surveillance application, including human, animal, and vehicle monitoring. The proposed deep learning model is trained for improving the accuracy level. The learning rate and the number of epochs are varied in the experiment, and the performances are verified. The increased number of epochs can tune the accuracy to high levels. The results of the experiments showed that the proposed CNN performs better than the alternative approaches. The obtained accuracy is 99.6% for abnormal activity classification. The various classes of the abnormality can be classified individually under different situations making the system fully automatic and suitable for any surveillance system. As a future enhancement of this research, the model can be tested with varying video categories with certain other parameters such as duration of the video, timestamp of the video recorded and also based on the number of persons available in each frame of the video.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: “The authors declare that they have no conflicts of interest to report regarding the present study.”

References

1. Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.
2. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., &Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1.
3. Balamurugan, S. P., &Duraisamy, M. Deep Convolution Neural Network with Gradient Boosting Tree for COVID-19 Diagnosis and Classification Model. European Journal of Molecular & Clinical Medicine, 7(11), 2020.
4. Sutrisno Ibrahim, "A comprehensive review on intelligent surveillance systems," Communications in Science and Technology, Vol. 1, No. 1, 2016.
5. Roberto Arroyo, J. Javier Yebes, Luis M. Bergasa, Ivan G. Daza and Javier Almazan, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls", Expert Systems with Applications, Vol. 42, No. 21, 2015, pp. 7991-8005.
6. Kun Wang, Stanley Langevin, Corey O'Hern, Michael Shattuck, Serenity Ogle, Adriana Forero, Juliet Morrison, Richard Slayden, Michael Katze, Michael Kirby, (2016), "Anomaly detection in host signaling pathways for the early prognosis of acute infection", PloS one, Vol. 11, No. 8, 2016.
7. Yudong Zhang, GenlinJi, Jiquan Yang, Shuihua Wang, Zhengchao Dong, Preetha Phillips, and Ping Sune, "Preliminary research on abnormal brain detection by wavelet energy and quantum-behaved PSO," Technology and Health Care, 2016, pp. 1-9.
8. Soumi Ray and Adam Wright, "Detecting Anomalies in Alert Firing within Clinical Decision Support Systems using Anomaly/Outlier Detection Techniques," Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, USA, 2016, pp. 185-190.
9. Wei Wang, Lin Chen, Kang Shin, and LingjieDuan, "Thwarting intelligent malicious behaviors in cooperative spectrum sensing," IEEE Transactions on Mobile Computing, Vol. 14, No. 11, 2015, pp. 2392-2405.
10. Oscar Rojas and ClesioTozzi, "Abnormal Crowd Behavior Detection Based on Gaussian Mixture Model," Proceedings of European Conference on Computer Vision, Springer International Publishing, Cham, 2016, pp. 668-675.
11. Andrea Pennisi, DomenicoBloisi, and Luca Iocchi, "Online real-time crowd behavior detection in video sequences," Computer Vision and Image Understanding, Vol. 144, 2016, pp. 166-176.
12. SupriyaMangale and MadhuriKhambete, " Camouflaged Target Detection and tracking using thermal infrared and visible spectrum imaging," Proceedings of International Symposium on Intelligent Systems Technologies and Applications, Springer International Publishing, Cham, 2016, pp. 193-207.
13. Yang Xian, XuejianRong, Xiaodong Yang, and YingliTian, "Evaluation of Low-Level Features for Real-World Surveillance Event Detection," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 27, No. 3, 2016, pp. 624-634.
14. SerhanCoşar, Giuseppe Donatiello, VaniaBogorny, Carolina Garate, Luis OtavioAlvares, and François Brémond, "Toward Abnormal Trajectory and Event Detection in Video Surveillance," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 27, No. 3, 2017, pp. 683-695.
15. SaritaChaudhary, MohdAamirKhana, CharulBhatnagar, (2018), "Multiple Anomalous Activity Detection in Videos," Procedia Computer Science, Vol. 125, pp. 336–345.
16. "CMU Graphics Lab Motion Capture Database.", http://mocap.cs.cmu.edu/, last accessed 2018/1/2.
17. Ryoo, MS, and Aggarwal, J.K.: "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," HTTP:// cvrc.ece.utexas.edu /SDHA2010 /Human_Interaction.html.
18. “Peliculas Movies Fight Detection Dataset”, http://academictorrents.com/ details /70e0794e2292fc051a13f05ea6f5b6c16f3d3635 /tech&h it=1&filelist=1, last accessed 2018/1/5.
19. E. Bermejo, O. Deniz, G. Bueno, R. Sukthankar, (2011), "Violence Detection in Video using Computer Vision Techniques," Proceedings of Computer Analysis of Images and Patterns.
20. "CRF Web Dataset," http://crcv.ucf.edu/projects/Abnormal_Crowd/#WebDataset, last accessed 2018/1/5.
21. http : // www . svcl .ucsd .edu / projects / anomaly / dataset.htm.
22. Balasundaram and Chellappan, (2018), "An intelligent video analytics model for abnormal event detection in online surveillance video," Journal of Real-Time Image Processing, Doi: 10.1007/s11554-018-0840-6.
23. M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," Compt. Vis. Image Und., 116(3):320–329, 2012.
24. V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size, and texture," In CVPR Workshops, pages 55–61, 2011.
25. V. Saligrama and C. Zhu, "Video anomaly detection based on local statistical aggregates," In CVPR, pages 2112–2119, 2012.
26. N Celandroni, E Ferro, A Gotta, et al. A survey of architectures and scenarios in statellite-based wireless sensor networks: system design aspects. International Journal of Satellite Communications and Networking. 2013; 31(1): 1-38.
27. Xia Weia, Yan Xijuna, Wei Xiaodong. Design of Wireless Sensor Networks for Monitoring at Construction Sites. Intelligent Automation & Soft Computing. 2012; 18(6): 635-646.
28. M. Agarwal, P. Parashar, A. Mathur, K. Utkarsh, and A. Sinha, “Suspicious Activity Detection in Surveillance Applications Using Slow-Fast Convolutional Neural Network,” in Advances in Data Computing, Communication and Security, Springer Nature Singapore, 2022, pp. 647–658.
29. H. Tan et al., “RelativeNAS: Relative Neural Architecture Search via Slow-Fast Learning,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–15, 2021, doi: 10.1109/tnnls.2021.3096658.
30. M.-H. Ha and O. T.-C. Chen, “Deep Neural Networks Using Residual Fast-Slow Refined Highway and Global Atomic Spatial Attention for Action Recognition and Detection,” IEEE Access, pp. 164887–164902, 2021, doi: 10.1109/access.2021.3134694.
31. Z. Jie, W. Muqing, and X. Weiyao, “A Two-Pathway Convolutional Neural Network with Temporal Pyramid Network for Action Recognition,” 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Dec. 2020, doi: 10.1109/iccc51575.2020.9345152.
32. T. Zia, N. Bashir, M. A. Ullah, and S. Murtaza, “SoFTNet: A concept-controlled deep learning architecture for interpretable image classification,” Knowledge-Based Systems, p. 108066, Mar. 2022, doi: 10.1016/j.knosys.2021.108066.
33. X. Zhang, Y. Tie, and L. Qi, “Multimodal Gesture Recognition Based on Attention Slow-Fast Fusion Networks,” Journal of Physics: Conference Series, no. 1, p. 012031, Jan. 2021, doi: 10.1088/1742-6596/1757/1/012031.
34. Stergiou, Alexandros. (2021). Efficient Modelling Across Time of Human Actions and Interactions. Available online via: https://arxiv.org/abs/2110.02120. Accessed 24 November, 2022.