Detecting Video Surveillance Using VGG19 Convolutional Neural Networks

Umair Muneer butt[1]
Department of Computer Sciences
University Sains Malysia
University of Lahore

Sukumar Letchmunan[2]
Department of Computer Sciences
University Sains Malysia

Fadratul Hafinaz Hassan[3]
Department of Computer Sciences
University Sains Malysia

Dr. Sultan Zia[4]
Department of Computer Sciences
University of Lahore
Chenab Campus, Pakistan

Anees Baqir[5]
Faculty of Computing & IT
University of Sialkot
Sialkot, Pakistan

***Abstract*—The meteoric growth of data over the internet from the last few years has created a challenge of mining and extracting useful patterns from a large dataset. In recent years, the growth of digital libraries and video databases makes it more challenging and important to extract useful information from raw data to prevent and detect the crimes from the database automatically. Street crime snatching and theft detection is the major challenge in video mining. The main target is to select features/objects which usually occurs at the time of snatching. The number of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 101; the videos in the dataset are further divided into frames. The frames are labelled and segmented for training. We applied the VGG19 Convolutional Neural Network architecture algorithm and extracted the features of objects and compared them with original video features and objects. The main contribution of our research is to create frames from the videos and then label the objects. The objects are selected from frames where we can detect anomalous activities. The proposed system is never used before for crime prediction, and it is computationally efficient and effective as compared to state-of-the-art systems. The proposed system outperformed with 81 % accuracy as compared to state-of-the-art systems.**

***Keywords*—Anomalous detection; surveillance video; VGG16; VGG19; ConvoNet; AlexNet**

I. Introduction

As the technology is growing rapidly, the crime ratio and strategies are also advancing. One of the major crimes faced by almost all over the world is street and theft crime [1]. One of the basic countermeasures is to do surveillance, i.e. monitoring the area, which is done by the CCTV cameras, it allows the user to watch what is going on in different places, and their footage can also be accessed remotely by the number of authenticated users and agencies. However, there is no intelligent method to identify or detect a specific object or person. The manual and common approach are to watch the lengthy videos carefully from one CCTV recording to another. It is quite difficult to detect abnormal activities through this CCTV footage. The picture quality, the motion and objects were identified through CCTV cameras [2] [3].

Here, the important question arises is that how we can detect the abnormal activities before it occurs. The basic challenge is to automatically and intelligently watch the video surveillance and detect the abnormal and anomalous events in rushy areas and protect the individual(s) at the spot. There are huge limitations which makes it challenging and tough to detect an anomalous event at the spot [4]. The selection of features is very tough because, through these features, we can detect anomalous activities. The features selected are responsible for detecting the moving object and has a significant impact on the analysis of behaviour and the performance of the system [5]. Figure 1 depicts the basic structure of a surveillance network.

Fig. 1. CCTV Surveillance Network

Nowadays, data mining is considered the most vigorous research field. By data mining, we mean the process of mining knowledge from the raw data and discovering fascinating pattern from a huge set of data. In data mining, most work is done on heterogeneous and unstructured data, i.e. videos, images, etc. [4]. A variety of technical tools are available for detecting video surveillance. LI Yi et al. [6] uses neural network architecture for segmentation and shape estimation. To achieve optimal performance, their architecture alternate between correspondence, deformation flow and segmentation

in an Inductively Coupled Plasma (ICP) like fashion. The important part is the induction algorithm, which successfully generalizes to new and unseen objects.

The most popular among all video surveillance system is a traffic surveillance system. Because the surveillance sensor and processors are available in the market at a very cheap rate and their decision-making capability is very much effective [7]. Usually, the majority of the system provides the facility of detecting motions and record the video when motions are detected; it reduces the processing and storage time of the video. It allows the users to remotely access the cameras from multiple devices and store the recorded videos in various formats. Figure 2 represents the basic structure of a traffic surveillance system.

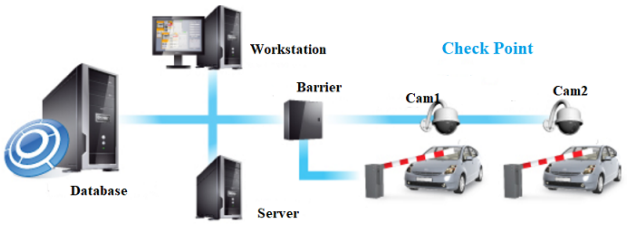

Fig. 2. Traffic Surveillance System

The main focus of this paper is detecting street crime i.e. snatching and theft via video surveillance. We mainly focuses on the object which usually occurs at the time of snatching [8] [9]. The amount of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 1.0[1], the video in the datasets are further divided into frames and from these frames, features are extracted. For this purpose, VGG19 algorithm has been employed in this paper and the results were found which compares the features from their original video [7]. The understand that the proposed method outperforms the state-of-the-art techniques, comparative analysis is performed and it is concluded based on the results that the proposed method outperformed them.

Moreover, the paper is organized as follows. Section II emphasis on the literature, while section III of the paper explained the methodology. In section IV, we present the results and outcomes of the research. Section V concludes the paper with future work.

II. Literature Review

Video surveillance is an essential part of our society to foresee criminal activities. Numerous efforts have been made in this area, but efficiency is still a big challenge as shwon in Table I. In [5] Appearance and Motion DeepNet (AMDN) based method is used to find out more Stacked Denoising Auto Encoders (SDAE) active video scene appearance and presentation of the motion. This new method of unsupervised learning is based on the depth study of anomalies architecture of video detection. This method examines the appearance, features and joint representation. There is an extensive experimental evaluation, taking into account three complex sets of social video data anomaly detection of the train, UC San Diego(UCSD) and subway, and demonstrated that the proposed method is reliable and effective. To detect additional unusual events using co-occurrence of more than one pattern, the AMDN method is beneficial.

The sparse method of representation is widely used in abnormal population detection; specifically, they represent dimensional movements. In [11], they proposed a method for detecting abnormal crowds. The proposed method includes two deep replacement processes, each of which uses a dynamic daily updated dictionary. Dynamic data update dictionary process dynamically adds normal test procedures to the dictionary while other dictionaries with exception samples are also resolved. The proposed model offers a wider vocabulary about normal and anomalous events. Additionally, abnormal events are more accurate than prior art methods. The results calculated from the experimental datasets depicted higher accuracy achieved by the proposed method as compared with the latest methods in local and global anomaly detection.

Dinesh et al. [12] use deep learning through bidirectional Long Short Term Memory (LSTM) for real-time violence detection in football stadiums. They used real-time video streams processed by the Spark framework along with Histogram of Oriented Gradients (HOG) to separate frames and features of each frame. Features are then labelled based on violence model and used for training to Bidirectional Long short term memory networks (BDLSTM) to recognize violent activities in the scenes. They validated this model with 94% accuracy in the detection of violent actions.

The video surveillance camera is increasingly challenging the video control system. The monitoring center needs monitoring tools to drive. Intelligent video perimeter protection solutions have to select and display cameras with evidence of these events, but background-based modelling systems only focus on the problem, whether or not an intrusion occurs. In [13], the authors recommended that you add a module based on machine learning and global functionality to adapt the video surveillance solution to identify problematic situations and provide the best priority. Instead of improving the robustness of a virtually impossible environment, the authors propose a way to solve problematic events based on global features.

The employment of surveillance cameras is increasing indoors and outdoors; it requires system intelligent enough to detect anomalous activities. Authors in this paper [14] investigated the most popular methods of extraction and description methods and presented an overview of the behaviour modelling classification methods and frames. Additionally, the authors presented a dataset and metric evaluation challenges for video systems. Finally, they introduced some intelligent real-world video systems.

Anpei et al. [15] introduced a novel algorithm based on the neural network called "Deep Surface Light Field" or DSLF for moderate sampling. Leveraging different patterns of sampling, DSLF fills in the missing data. They also addressed image registration. Aniqa et al. [16] proposed a framework which works by extracting the visual-based features from the frames of video by employing "Convolutional Neural Networks" (CNN). Furthermore, the framework passed the derived representations to the LSTM model. For the natural language description of

[1]https://sites.google.com/view/debadityaroy/datasets

TABLE I. A BRIEF COMPARISON OF THE TECHNIQUES TO DETECT VIDEO SURVEILLANCE

Reference	Techniques	Dataset	Description
[5]	Appearance and Motion DeepNet, SVM models	UCSD pedestrian dataset, consists of two subsets: Ped1 and Ped2	Depicted better performance as compared to existing methods. The basic advantage their approach was that it did not depend on any piror knowledge to design the features
[4]	FCN architecture	UCSD (Ucsd anomaly detection dataset, 2017) and Subway Benchmarks	The proposed method helped run a deep learning-based method at a speed of about 370 fps
[2]	SVM Model, Random Forest	Violent Flows- 246 videos, half of them being violent and other half being non violent collected from Youtube with 320*240 pixels resolution	To overcome the abnormal behaviors and limitations mentioned in the paper, the system proposed by the authors used huge amount of training data including all possible scenarios
[10]	Locality Sensitive Hashing Filters,Particle Swarm Optimization	UMN dataset consists of three different crowded scenes with a frame number of 1453, 4143, and 2143.	The abnormality degree of a new test sample is estimated by calculating the filter response of the test sample to its nearest filter. It was concluded that the proposed method is effective and robust
[8]	kernel support vector machine, binary support vector machine with graph kernel	UCSDped2 dataset, which contained that scenes of movements by pedestrian parallel to the plane of camera	Graph is used to represent the interaction and the co-relation of the motions of objects/entities. By using the graph kernel, to measure the similarity between two graphs provides robustness to slight deformations to the topological structures due to presence of noise in data

video frames, they used a fine-tuned CNN model.

For this reason, the scene transformations are sensitive to the characteristics that are robust and change the appearance of the object to attract the correlation. The selection of functions used to characterize floating objects is a lightweight job because it has a great impact on the description and analysis of behavior. In this review [14], different levels of the system of video surveillance were analyzed by the authors, which resulted in a behavioural representation and behavioural pattern.

In [17], the authors suggested another technique to classify suspicious events in video observation based on locality-sensitive hashing filters. Training tests are hashed into a rundown of pails, and each bucket's middle and radius is found to create location-sensitive hashing filters. The deviation from the normal level of another test is measured by the test's filter reaction to its closest filter. With new expectations, the locality-sensitive hashing filter is refreshed online. Test results demonstrate the adequacy and power of the proposed methodology on three datasets.

In this paper [18], they proposed a methodology for anomalous movement acknowledgment dependent on chart definition of video activists and graph kernel support vector machine. As a graph of geometric relationships between space-time intrigue points, the connection of the substances in a video is detailed. The graph vertices are spatio-worldly intrigue graph, and an edge represents the connection between the appearance and the elements surrounding the points of intrigue. For this, the chart details the improvements in video activities over the issue of detecting anomalies into a graph classification problem. Using the chart kernel to approximate the resemblance between two graphs gives the topological structure's power to minor mishappenings due to the nature of information noise.

An anomalous event detection technique has been proposed in [19], which was dependent on the unsupervised deep neural network. In particular, successful highlights of video events are thus omitted from 3D slopes to reflect both the appearance and the hint of movement. They use a deep Gaussian mixture model to lean ordinary event designs, which typically perform violent execution using a few parameters. Examinations on two open datasets indicate particular upgrades when compared with state of the art algorithms.

In this study [10], they introduced a structure for snatch theft detection in unconstrained videos utilizing activity credit demonstrating to take in all the activity traits in the snatch robberies, a huge Gaussian mixture model (GMM) called all universal attribute model (UAM) was prepared to utilize existing video datasets of human activities. For development, the authors presented a dataset called snatch 1.0 that contains snatch robberies in surveillance videos. It was demonstrated that activity vector pro video better discriminate portrayal for snatch robbery.

Wenqing et al. [20] presented a novel unsupervised deep feature learning algorithm for anomalous event detection. To fully utilize the Spatio-temporal information, the proposed system used a deep three-dimensional convolutional neural

network for feature extraction. To train the C3D network without any category labels, they used a sparse coding result of handcrafted features. The proposed system outperforms the state-of-the-art systems. Schuchao et al. [21] proposed a deep learning-based technique for tracking visual objects. They used CNN to rank the patches of the target objects based on how well it is centred. The promising patch is selected by the AlexNet framework using his matching function based on deep features.

Asghar et al. [22] introduced a novel algorithm for high-level feature extraction and used those features for classification and re-identification. Their proposed method is a two-tier approach. Firstly, they extract low-level features for identification and later use high-level features for classification and re-identification. In the end, they used a deep belief network to build a model based on the low and high-level features. Yonglong et al. [23] proposed a radio frequency-based fall monitoring system based on CNN. They introduce Aryokee, which is based on radio frequency to detect fall using CNN. The key idea behind this is to separate different sources of motion, which resulted in increased robustness. They achieved 94% recall and 92% precision in detecting falls.

Umair et al. [24] used a combination of HoG, and LBP features to extract features from the American Sign Language dataset (ASL). They used those features in an auto model feature of Rapid Miner and Weka software to train and test. Rapid Miner auto model performed with 99 percent accuracy. Debaditya et al. [25] proposed employing a GMM model on snatch thefts with a large number of attribute mixtures known as the universal attribute model. They used large human action data set UCF101 and HMDB51 to train the proposed model. They used factor analysis for low-level feature representation and evaluation; they used Snatch 1.0 data set. The proposed system performed well as compare to state-of-the-art systems.

III. Proposed Methodology

In this section, we present the detail of the dataset and pre-processing applied to it to enable it for further calculations. Moreover, proposed VGG19 convolutional neural network is discussed which is used for retrieving results as shown in figure 3.

A. Dataset Description

There is a large volume of datasets available for human activity recognition and object detection. But in the existing literature, there are no proper databases for street theft and motorbike theft. Hence, we have found a dataset named Snatch 1.0, consists of normal videos and snatching videos [10]. This dataset shows the normal behaviour of objects in a different place, e.g. roads, streets, markets etc. As well as the dataset depicts abnormal and anomalous behaviour of objects while snatching, for example, the position of the vehicle, body movement, facial expressions of snatcher and victim behaviour during the snatching. Few glimpses of the dataset with the aforementioned features are depicted in figure 4.

B. Convert Videos into Image Frames

The first step in detecting anomalous activities is to divide our videos into images. In this paper, we have used the data from more than 21 videos and generate the frames from these videos, which are more than 1000 images. The frames are generated by using MATLAB R2018b. These frames show different cases and behaviours before, during and after snatching, as shown in Figure 4. The following are the steps we performed while generating frames from videos.

1) Create a directory in Matlab R2018b and copied all the videos in the directory.
2) Write the commands mentioned in algorithm 1 in the Matlab R2018b and run to generate frames.

Algorithm 1: Frame generation from Videos

```
shuttleVideo = VideoReader('17.mp4');
workingDirectory = tempname;
mkdir(workingDirectory) ;
mkdir(workingDirectory,'images') ;
ii = 1;
while hasFrame(shuttleVideo) do
    image = readFrame(shuttleVideo);
    file_name = [sprintf('03d',ii) '.jpg'];
    full_name =
     (workingDirectory,'images',file_name);
    imwrite(image,full_name) ii = ii+1;
end
```

C. Snatching Scenarios

After generating frames, the next step is to select features and object on whose bases we will detect surveillance video. Few cases, i.e. Case1 in figure 5 and Case2 in figure 6, are described in the following set of figures and explain how the snatcher snatch the chains/ wallets and what are the victim's response. After getting through these scenarios, we can select the object on which we have to mainly focus on how we will further implement our algorithm to detect the surveillance video. In the scenario depicted in Case1, there are two people on the bike who came closer to the victim and snatched her chain and ran away. From this scenario, we extract different objects, e.g. motorbike, snatcher, women, empty roads etc. These objects may be further used in classifying and labelling the frames.

D. Image Labeling

The next step is to label the object and features which we have extracted from various surveillance videos. The objects and features we selected in our paper are snatchers, the vehicle used by the snatcher, the environment in which the anomalous event occurred, the victim's behaviour before and after the snatching. To label the images, we have used MatLab R2018b Image Labeler app to label the images. The process of labelling an image in MatLab is described as follow:

1) Load all the images from the given folder as depicted in figure 7
2) Define Region of Interest (ROI) and Scene Label definitions; in Matlab, we have to label the images either in pixel region or rectangular format, which is defined in the section of ROI Label as shown in figure 8. It consists of two basic parts, one is the name of

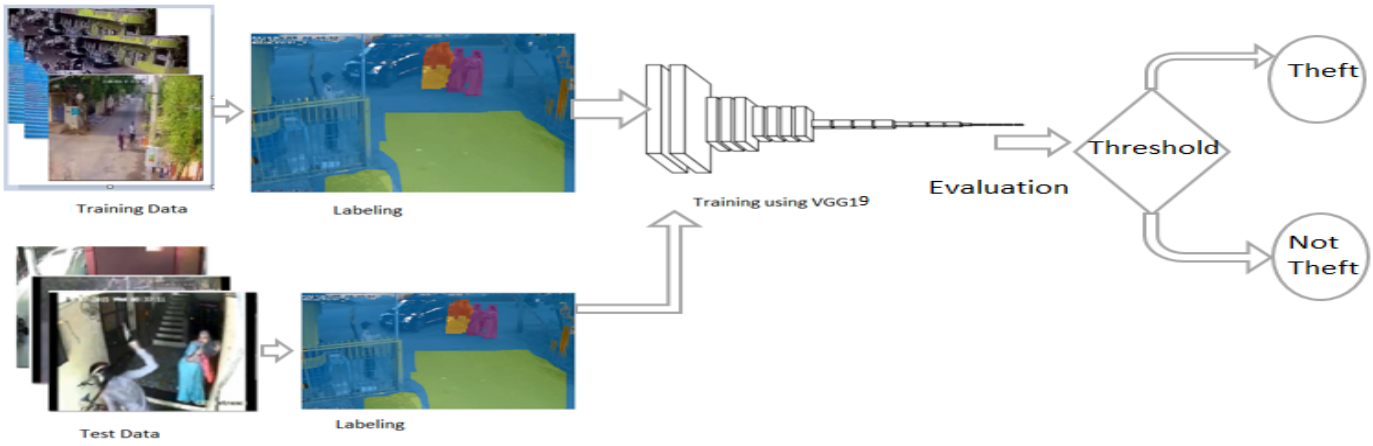

Fig. 3. Proposed Methodology

Fig. 4. Few Glimpse of Snatching from Data Set

the label, and the other one is the nature of the label (rectangular and pixel region). For example, our label is "snatcher" and the region you selected (rectangular or pixel region). The nature of the object is described in the Scene label format, such as the "background" and "environment". We also relate this label to our specific, defined frame.

3) Label the image objects either in rectangular or pixel format. We labeled all the images in pixel labeler format as show in figure 9.
4) The "green" color indicates "road" object, "pink" color defines women(victim), 'orange' color defines "snatcher", "yellow" color is for "vehicle (bike)" identification and "blue" color is for "background". We further creates classes of the mentioned objects and assign indexes to all the classes. In the end, we export label to the file or in the work space to save the labeled images using the **Export Labels** option shown in figure 10.

E. VGG19 Convolutional Neural Network

Convoluted networks (ConvNets) have been highly successful in recognition of large-scale images and videos, due to large-scale public storage depots (fast processing system such as GPU based operating system used and ImageNet or segmented the image into large clusters) [26]. In ConvNet depth measurements in fairy settings, our ConvNet Layer structure is designed with the same codes. In the training process, our fixed ConvNets input size is 112*112*128 RGB. Our only pre-processing is to reduce the average RGB value calculated for each pixel training group. The image is passed through a pile of convolutional layers, as shown in figure 11, and we use a very small cloud filter: 14*14*512 that is left to right, up to down, part of the concept [10].

We also use the max-pooling layer of 112*56*28*14*1, its nature is like a linear transformation input, but it is not linear. The first step of convolution is that 1 pixel is fixed, and the conversion space is filled. Layer input saves space resolution after convection, that is, for the conversion of 3x3, the padding is 1 pixel. Convolutional layers are stacked with different depths in different architectures following three layers Fully Connected (FC) layers: several channels in the first layer are 4096, and the last layer has 1000 channels. The last layer is a soft-max layer. We implement the same completed connection configuration for the overall system. All hidden layers are line-aligned, and our network is not standardization for Local

(a) Two persons on the bike (b) Snatching chain (c) Ran after snatching (d) Victim ran to catch him

Fig. 5. Case1: Snatching sequence

(a) Snatcher pretends (b) Snatching chain (c) Ran after snatching (d) Victim ran to catch him

Fig. 6. Case2: Snatching sequence without using the bike

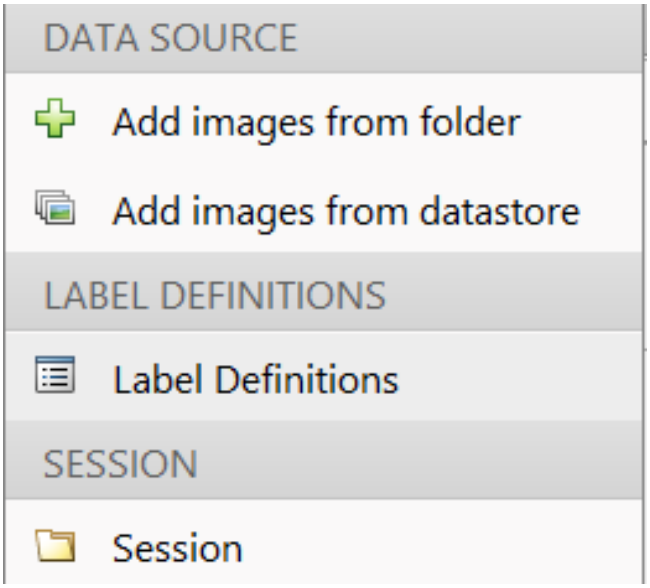

Fig. 7. Load (Video, Image Sequence or Custom Reader)

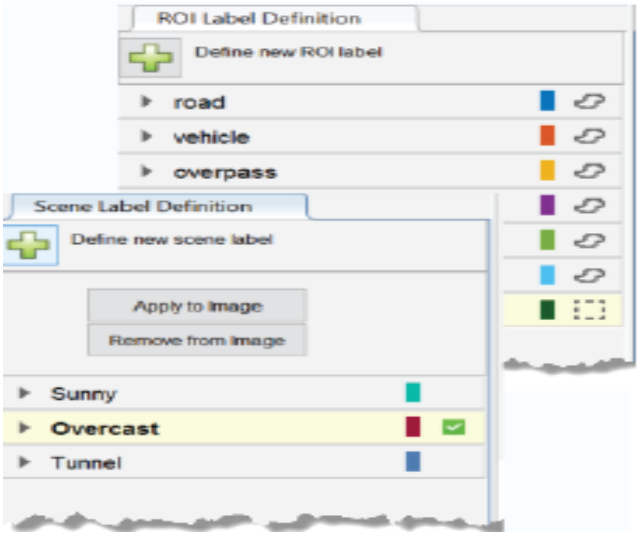

Fig. 8. Define ROI and Scene Labels

Response Normalization (LRN), which improves ILSVRC datasets but increases memory consumption and computing time [19]. Despite the great depth, the number of weights in our networks are not larger than the shallow net weight, with greater convolutional layer width and acceptable fields.

The training is done as [17] using the low-volume gradient to perform more impulse optimization for logistic regression .

Fig. 9. Labeled objects

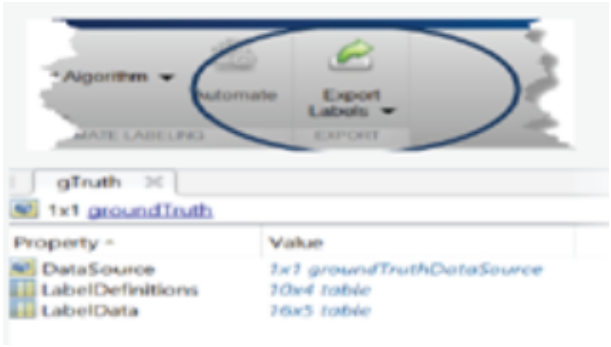

Fig. 10. Export Labeled

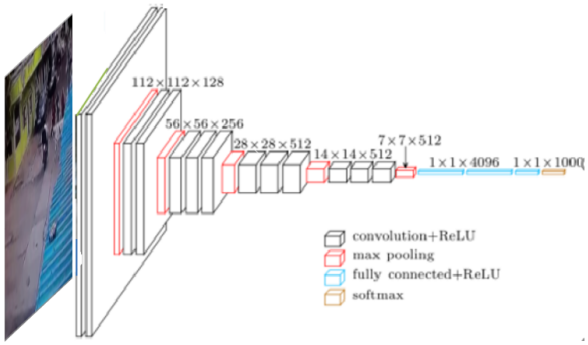

Fig. 11. VGG19 Architecture

The batch size goes to 256, and the torque is 0.9. Among the first two connecting layers, there are two ways to configure the training scale. The first one will change S, which one-scale training corresponds to the contents of the image can be represented by crop samples still multi-scale statistical image. In our experiments, we evaluated the two-scale model of training: S = 256 widely used before art, and S = 384 ConvNet setup, we used the first S = 256 Training network S = 384 networks, S = 256 pre-trained. We start with heavy weights, we have used a lesser initial 10-3-degree instruction, the second method is to configure a multi-scale S training, a specific range across each training, smax random samples, images individually readjusted S (stxikena = 256 and we use smax = 512) because the image may have different sizes for objects, so it is considered beneficial during the training period.

For many reasons, we prepare a multi-scale model with all scale layouts in the same configuration as [26] and then apply a fully uncropped image of the entire convolution network . The resulting category is graphical score; the size of the image depends on the numbers of classes as the number of channels and their variable resolution of space. In the end, to get a fixed size of the image scoring point, the category score is the average score. Since the entire computational network is applied to the whole image, multiple crops must be tested in different tests, which are more efficient, because it requires computation for each crop.

Our implementation from the public is derived from available C ++ Caffe tool [27], but it contains many important modifications that allow us to train. For training and evaluation of multiscale images we need to install multiple GPUs in one system as the authors did in this paper [10]. Multi-GPU training takes advantage of the data parallelism and is done by dividing each series of images for training on several GPU series and processing them parallel to each GPU. After calculating the GPU series of gradients, they concentrate to obtain a gradient of the complete series. Gradient calculations are synchronized between the GPU, so the results must be the same as the training model result on one GPU.

ILSVRC has been used for many years by the algorithm of one or more of the following tasks for image classification problem has algorithms generate a list of the categories of objects that are present in the image, localization algorithm explain the scale of the image and the axis of the bounding areas of the image [19], [26], [28]. Object detection has algorithms that create a list of object classes in the image along with a border-oriented box that indicates the position and size of each copy of each object class. For checking the accuracy of ILSVRC, we have to find precision and recall.

The major contribution of our paper, the methodology used in this paper VGG19 (Convent Neural Network), has never been implemented on this type of unstructured dataset, e.g. videos. As we have studied the previous research literature, human recognition is done on only images. In this paper, we proposed a method on a video dataset, generate frames and further label the objects which detect surveillance before it. The proposed method outperformed the state-of-the-art method with 81% accuracy.

IV. Result and Discussion

Video surveillance is an important area of research for the researchers and law enforcement agencies due to the widespread usage of cameras for abnormalities detection. Several techniques have been proposed to make a robust Video surveillance system for anomaly detection, but efficiency is still a big challenge for the researchers.

In this study, we used the VGG19 architecture of a convolutional neural network to predict video surveillance in snatching videos. The proposed architecture of VGG19 is particularly altered for video surveillance and detecting anomalies in the video. The combination of ConvNets, Convolutional layer and max-pooling in a proposed order found to be efficient and

effective, particularly in snatching detection. To the best of author knowledge, the proposed architecture has never been used for this purpose on an unstructured dataset.

We also compare the proposed system with the state-of-the-art system fined tuned models VGG16 and AlexNet. The earlier used models are not robust, not completely infallible and have false detections. To evaluate this process efficiently, a series of tests were carried out with the snatching theft videos that were not used in the training set. Another important aspect of our system is the processing time, which is far much better than the state-of-the-art systems with the same experimental setup. The processing time is a very important aspect of these kinds of real-world crime scenarios. The experiment was carried out on 300 video frames of the same data set. The results are shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF STATE-OF-THE-ART AND THE PROPOSED METHOD

Performance Measure	AlexNet	VGG16	VGG19
Positive Detection	219	231	239
Fails	81	69	61
Accuracy (%)	73	77	81
Frames Per Second (FPS)	0.4	0.04	0.025

V. CONCLUSION

A surveillance system is to detect and identify abnormal and anomalous events. This will only be possible when we select objects and features from the anomalous events. Certain weaknesses make it harder and more difficult. The amount of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 1.0; the video in the datasets are further divided into categories of normal and snatching videos. Then the videos are converted into image frames. The frames are labelled to identify the objects which we have selected for video surveillance detection.

For the implementation of the proposed method, we used VGG19 deep neural network and performed experiments on GPU based system. Later on, a comparison of experimental results was performed with the original video, and the accuracy and performance of the model were evaluated using the evaluation, as mentioned in the above techniques. The proposed system outperformed as compared to state-of-the-art systems with 81 % accuracy and 0.025 frames per second detection time.

In the future, we aim to work on further improving its accuracy and time efficiency by using ensemble methods along with Bidirectional Long Short Term Memory (BLSTM). We will also consider the demographic factors and crime statistics of the region to predict crime so that law enforcement agencies can take precautionary measures.

REFERENCES

[1] T. Manjunath, R. S. Hegadi, and G. Ravikumar, "A survey on multimedia data mining and its relevance today," *IJCSNS*, vol. 10, no. 11, pp. 165–170, 2010.

[2] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018.

[3] X. Chen and C. Zhang, "An interactive semantic video mining and retrieval platform–application in transportation surveillance video for incident detection," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 129–138.

[4] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, 2018.

[5] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.

[6] L. Yi, H. Huang, D. Liu, E. Kalogerakis, H. Su, and L. Guibas, "Deep part induction from articulated object pairs," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, p. 209, 2019.

[7] P. Thirumurugan and S. H. Hussain, "Event detection in videos using data mining techniques," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 2, pp. 3473–3475, 2012.

[8] D. Singh and C. K. Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognition*, vol. 65, pp. 265–272, 2017.

[9] E. Cermeño, A. Pérez, and J. A. Sigüenza, "Intelligent video surveillance beyond robust background modeling," *Expert Systems with Applications*, vol. 91, pp. 138–149, 2018.

[10] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognition*, vol. 59, pp. 302–311, 2016.

[11] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[12] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[13] A. V. Kate, P. Nikilav, S. Giriesh, R. Hari Prasath, and J. Naren, "Multimedia data mining-a survey," *International Journal Of Engineering And Computer Science*, vol. 3, no. 12, 2014.

[14] J. Oh, J. Lee, and S. Kote, "Real time video data mining for surveillance video streams," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2003, pp. 222–233.

[15] A. Chen, M. Wu, Y. Zhang, N. Li, J. Lu, S. Gao, and J. Yu, "Deep surface light fields," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 1, no. 1, p. 14, 2018.

[16] A. Dilawari, M. U. G. Khan, A. Farooq, Z.-U. Rehman, S. Rho, and I. Mehmood, "Natural language description of video streams using task-specific feature encoding," *IEEE Access*, vol. 6, pp. 16 639–16 645, 2018.

[17] T. Karthikeyan, B. Ragavan, and N. Poornima, "A comparative study of algorithms used for leukemia detection," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, vol. 5.

[18] D. Roy *et al.*, "Snatch theft detection in unconstrained surveillance videos using action attribute modelling," *Pattern Recognition Letters*, vol. 108, pp. 56–61, 2018.

[19] H. Xu, M. Fang, L. Li, Y. Tian, and Y. Li, "The value of data mining for surveillance video in big data era," in *Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on*. IEEE, 2017, pp. 202–206.

[20] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 246–255, 2018.

[21] S. Pang, J. J. del Coz, Z. Yu, O. Luaces, and J. Díez, "Deep learning to frame objects for visual target tracking," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 406–420, 2017.

[22] A. Feizi, "High-level feature extraction for classification and person re-identification," *IEEE Sensors Journal*, vol. 17, no. 21, pp. 7064–7073, 2017.

[23] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 137, 2018.

[24] U. M. Butt, B. Husnain, U. Ahmed, A. Tariq, I. Tariq, M. A. Butt, and M. S. Zia, "Feature based algorithmic analysis on american sign language dataset."

[25] D. Roy *et al.*, "Snatch theft detection in unconstrained surveillance videos using action attribute modelling," *Pattern Recognition Letters*, vol. 108, pp. 56–61, 2018.

[26] J. Oh, J. Lee, and S. Hwang, "Video data mining: Current status and challenges. encyclopedia of data warehousing and mining.(a book edited by dr. john wang)," *Idea Group Inc. and IRM Press*, 2005.

[27] A. Divakaran, K. Miyahara, K. A. Peker, R. Radhakrishnan, and Z. Xiong, "Video mining using combinations of unsupervised and supervised learning techniques," in *Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307. International Society for Optics and Photonics, 2003, pp. 235–244.

[28] D. Saravanan and S. Srinivasan, "Data mining framework for video data," in *Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010*. IEEE, 2010, pp. 167–170.