ISSN 2395-1621

Video-Based Abnormal Human Behavior Detection

[#1]Rohit Oswal, [#2]Ashish Nalawade, [#3]Pushkar Bhangale, [#4]Krishna Kishore,
[#5]Prof.B.S.Gayal

[#1234]BE, Department of E&TC,
[#5]Prof., Department of E&TC

Sinhgad Academy of Engineering, S.P.Pune University, Pune, India.

ABSTRACT

To develop a real time video surveillance system for security purposes, by detecting normal and abnormal behavior of human beings and making the security system automatic and real time. Modeling human behaviors and activity patterns for recognition or detection of special event has attracted significant research interest in recent years. Diverse methods that are abound for building intelligent vision systems aimed at scene understanding and making correct semantic inference from the observed dynamics of moving targets. Most applications are in surveillance, video content retrieval, and human–computer interfaces. This paper presents not only an update extending previous related surveys, but also a focus on contextual abnormal human behavior detection especially in video surveillance applications. The main purpose of this survey is to extensively identify existing methods and characterize the literature in a manner that brings key challenges to attention.

Keywords: Abnormal Behavior, Video Surveillance , Security, Vulgar Signs, Violent Behavior, Handshaking motions.

ARTICLE INFO

Article History

Received: 5th May 2019

Received in revised form :

5th May 2019

Accepted: 8th May 2019

Published online :

9th May 2019

I. INTRODUCTION

Right now around 7.5 billion people are living on this planet. The population of cities towns is increasing day by day. With the increasing number of population the need for human safety and monitoring security has become a major concern. In the last two decades, the number of surveillance cameras installed in private and public places has increased dramatically. It is happening because of the rising fear in people about crime and terrorism. For this reason Surveillance cameras are set at important places to ensure security like home, airports, banks, city centers and major important places. Nowadays the autonomous visual analysis of surveillance videos and images is a major field of research in computer vision. The main reason is the process that influences the behavior in the monitored scene due to the different variety of situations in real life. One important necessity is to detect the situation when unexpected things happen. There is an increasing desire and need to detect abnormal behavior from live video surveillance. Due to technological advancement image processing for observing abnormal behavior detection has become desired area for research. Successful detection can prevent unwanted events and which can save human lives, assets. If we can apply machine learning, to be more precise Convolutional Neural Networks concept with this the system will become more efficient and cost effective. It is possible that, in the near future this system will be able to give the amount of security an individual expects to get rid of the rising fears about crime, terrorism and personal security.

All of these resulted into a great loss of human lives. We wanted to build a system which will prevent these unknown and unexpected events in a reliable cost effective way. Previous works based on Neural Networks helped to classify objects, image segmentation and create models which is efficient to classify images. Moreover, in our abnormal behavior detection process we are using CNN because it provides a wide application in dynamic image analysis in the form of machine learning. First of all, we tried to implement our work by using manual algorithms which was bit costly to implement and later on we decided to go for Convolutional Neural Networks where CNN is way more cost effective and can be updated inexpensively later on. Using CNN for security and include its classification method for dynamic image analysis was our main concern. CNN consists of multi-layered based neural network where there is huge number of connections

between neurons. CNN is a self-learning model where the network is trained using a large dataset of dynamic images according to their weights. It helps the network to learn and then in testing phase system makes assumptions of the input appropriately and gives outputs as far as the output classes are concerned. Finally, this model will help us to build a highly advanced artificial Intelligence system which will be able to recognize abnormal behavior from a data set of multiple dynamic images for a particular domain area where abnormal behavior will be classified by the person who is in need of the security to detect the following actions in the video :

a) Vulgar signs
b) Violent Behavior
c) Handshake

II. REVIEW OF LITERATURE

Abnormal event detection is a major topic of research due to which a lot of people have come up with different algorithm to detect abnormality. We do a detailed study and comparison of each of them and come up with better approach if possible to include with the sparsity learning method to achieve better accuracy.

In[4]author uses clustering-based approach for detecting abnormalities in surveillance video requires the appropriate definition of similarity between events. The HMM-based similarity defined previously falls short in handling the overfitting problem. Author propose in this paper a multi-sample-based similarity measure, where HMM training and distance measuring are based on multiple samples. These multiple training data are acquired by a novel dynamic hierarchical clustering (DHC) method. By iteratively reclassifying and retraining the data groups at different clustering levels, the initial training and clustering errors due to overfitting will be sequentially corrected in later steps. Experimental results on real surveillance video show an improvement of the proposed method over a baseline method that uses single-sample-based similarity measure and spectral clustering.

A surveillance system that supports a human operatorbyautomatically detecting abandoned objects and drawing the operator's attention to such events as addressed in [5]. It consists of three major parts: foreground segmentation based onGaussian Mixture Models, a tracker based on blob association and a blob-based objectclassification system to identify abandoned objects. For foreground segmentation, weassume that video sequences of the background shot under different natural settingsare available a priori. The tracker uses a single-camera view and it does not differentiate between people and luggage. The classification is done using the shape ofdetected objects and temporal tracking results, to successfully categorize objects intobag and non-bag (human). If a potentially abandoned object is detected, the operatoris notified and the system provides the appropriate key frames for interpreting theincident.

Human activities taking place in an outdoor surveillance environment can be effectively detected using method described in [7]. Human tracks are provided in real time by the baseline video surveillance system. Given trajectory information, the event analysis module will attempt to determine whether or not a suspicious activity is currently being observed. However, due to real-time processing constrains, there might be false alarms generated by video image noise or non-human objects. It requires further intensive examination to filter out false event detections which can be processed in an off-line fashion. Author propose a hierarchical abnormal event detection system that takes care of real time and semi-real time as multi-tasking. In low level task, a trajectory-based method processes trajectory data and detects abnormal events in real time. In high level task, an intensive video analysis algorithm checks whether the detected abnormal event is triggered by actual humans or not.

In[8], author address the problem of unusual-event detection in a video sequence. Invariant subspace analysis (ISA) is used to extract features from the video, and the time-evolving properties of these features are modeled via an infinite hidden Markov model (iHMM). The iHMM retains a full posterior density function on all model parameters, including the number of underlying HMM states. Anomalies (unusual events) are detected subsequently if a low likelihood is observed when associated sequential features are submitted to the trained iHMM. A hierarchical Dirichlet process framework is employed in the formulation of the iHMM. The evaluation of posterior distributions for the iHMM is achieved in two ways: via Markov chain Monte Carlo and using a variational Bayes formulation. Comparisons are made to modeling based on conventional maximum-likelihood-based HMMs, as well as to Dirichlet-process-based Gaussian-mixture models.Automatic technique for detection of abnormal events in crowds is presented in [29].Crowd behavior is difficult to predict and might not be easily semantically translated. Moreover it is difficult to track individuals in the crowd using state of the art tracking algorithms. Therefore author characterize crowd behavior by observing the crowd optical flow and use unsupervised feature extraction to encode normal crowd behavior. The unsupervised feature extraction applies spectral clustering to find the optimal number of models to represent normal motion patterns. The motion models are HMMs to cope with the variable number of motion samples that might be present in each observation window. The results on simulated crowds demonstrate the effectiveness of the approach for detecting crowd emergency scenarios.

III. COLLECTION OF METHODS AND PROCESSES

Soft computing:
Soft computing provides an approach to problem-solving using means other than computers. With the human mind as a role model, soft computing is tolerant of partial truths, uncertainty, imprecision and approximation, unlike traditional computing models. The tolerance of soft computing allows researchers to approach some problems that traditional computing can't process.

Deep Learning:

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts. Deep learning models are vaguely inspired by information processing and communication patterns in biological nervous systems yet have various differences from the structural and functional properties of biological brains (especially human brains), which make them incompatible with neuroscience evidences.

Deep learning is a class of machine learning algorithms that:

1) Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
2) Learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.
3) Learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.

Convolution Neural Network :

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery.

CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

IV. STUDIES AND FINDINGS

Convolutional Neural Networks (ConvNets or CNNs) are a category of Neural Networks that have proven very effective in areas such as image recognition and classification. ConvNets have been successful in identifying faces, objects and traffic signs apart from powering vision in robots and self drivingcars.aAConvNet is able to recognize scenes and the system is able to suggest relevant captions ("a soccer player is kicking a soccer ball") whileConvNets being used for recognizing everyday objects, humans and animals. Lately, ConvNets have been effective in several Natural Language Processing tasks (such as sentence classification) as well.ConvNets, therefore, are an important tool for most machine learning practitioners today. However, understanding ConvNets and learning to use them for the first time can sometimes be an intimidating experience. The primary purpose of this blog post is to develop an understanding of how Convolutional Neural Networks work on images.If you are new to neural networks in general, I would recommend reading this short tutorial on Multi LayerPerceptrons to get an idea about how they work, before proceeding. Multi LayerPerceptrons are referred to as "Fully Connected Layers" in this post.

The LeNet Architecture (1990s)

LeNet was one of the very first convolutional neural networks which helped propel the field of Deep Learning. This pioneering work by YannLeCun was named LeNet5 after many previous successful iterations since the year 1988. At that time the LeNet architecture was used mainly for character recognition tasks such as reading zip codes, digits, etc.

Below, we will develop an intuition of how the LeNet architecture learns to recognize images. There have been several new architectures proposed in the recent years which are improvements over the LeNet, but they all use the main concepts from the LeNet and are relatively easier to understand if you have a clear understanding of the former.

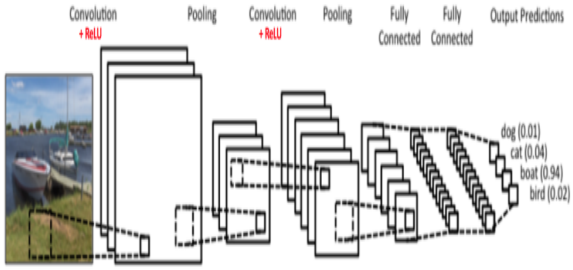


Fig.simple ConvNet.

The Convolutional Neural Network in Fig. is similar in architecture to the original LeNet and classifies an input image into four categories: dog, cat, boat or bird

There are four main operations in the ConvNet shown in fig. above:

1. Convolution
2. Non Linearity (ReLU)
3. Pooling or Sub Sampling
4. Classification (Fully Connected Layer)

An Image is a matrix of pixel values

Essentially, every image can be represented as a matrix of pixel values

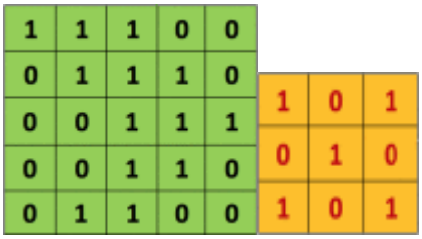
Channel is a conventional term used to refer to a certain component of an image. An image from a standard digital camera will have three channels – red, green and blue – you can imagine those as three 2d-matrices stacked over each other (one for each color), each having pixel values in the range 0 to 255.

The Convolution Step

ConvNets derive their name from the "convolution" operator. The primary purpose of Convolution in case of a ConvNet is to extract features from the input image. Convolution preserves the spatial relationship

between pixels by learning image features using small squares of input data. We will not go into the mathematical details of Convolution here, but will try to understand how it works over images

As we discussed above, every image can be considered as a matrix of pixel values. Consider a 5 x 5 image whose pixel values are only 0 and 1 (note that for a grayscale image, pixel values range from 0 to 255, the green matrix below is a special case where pixel values are only 0 and 1):



Also, consider another 3 x 3 matrix as shown.Then, the Convolution of the 5 x 5 image and the 3 x 3 matrix can be computed as shown in the animation in **Fig** below:

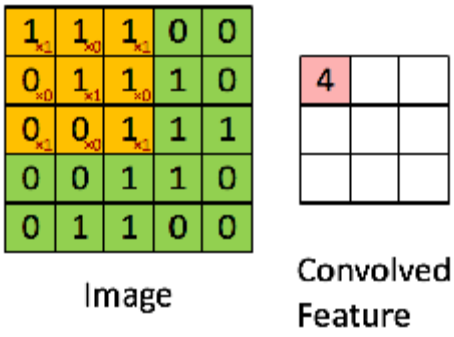
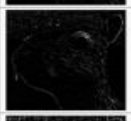
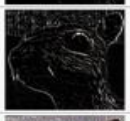


Fig : The Convolution operation. The output matrix is called Convolved Feature or Feature Map.

Take a moment to understand how the computation above is being done. We slide the orange matrix over our original image (green) by 1 pixel (also called 'stride') and for every position, we compute element wise multiplication (between the two matrices) and add the multiplication outputs to get the final integer which forms a single element of the output matrix (pink). Note that the 3×3 matrix "sees" only a part of the input image in each stride.

In CNN terminology, the 3×3 matrix is called a '**filter**' or 'kernel' or 'feature detector' and the matrix formed by sliding the filter over the image and computing the dot product is called the 'Convolved Feature' or 'Activation Map' or the '**Feature Map**'. It is important to note that filters acts as feature detectors from the original input image. It is evident from the animation above that different values of the filter matrix will produce different Feature Maps for the same input image. As an example, consider the following input image:

In the table below, we can see the effects of convolution of the above image with different filters. As shown, we can perform operations such as Edge Detection, Sharpen and Blur just by changing the numeric values of our filter matrix before the convolution operation– this means that different filters can detect different features from an image, for example edges, curves etc.

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$	
Edge detection	$\begin{bmatrix}1&0&-1\\0&0&0\\-1&0&1\end{bmatrix}$	
	$\begin{bmatrix}0&1&0\\1&-4&1\\0&1&0\end{bmatrix}$	
	$\begin{bmatrix}-1&-1&-1\\-1&8&-1\\-1&-1&-1\end{bmatrix}$	
Sharpen	$\begin{bmatrix}0&-1&0\\-1&5&-1\\0&-1&0\end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix}1&2&1\\2&4&2\\1&2&1\end{bmatrix}$	

Introducing Non Linearity (ReLU)

An additional operation called ReLU has been used after every Convolution operation in **Figure 3** above. ReLU stands for Rectified Linear Unit and is a non-linear operation. Its output is given by:

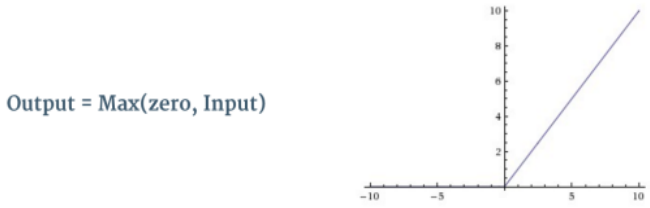


Fig: the ReLU operation

ReLU is an element wise operation (applied per pixel) and replaces all negative pixel values in the feature map by zero. The purpose of ReLU is to introduce non-linearity in our ConvNet, since most of the real-world data we would want our ConvNet to learn would be non-linear (Convolution is a linear operation – element wise matrix multiplication and addition, so we account for non-linearity by introducing a non-linear function like ReLU).

The Pooling Step

Spatial Pooling (also called subsampling or downsampling) reduces the dimensionality of each feature map but retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum etc.

In case of Max Pooling, we define a spatial neighborhood (for example, a 2×2 window) and take the largest element from the rectified feature map within that window. Instead of taking the largest element we could also take the average (Average Pooling) or sum of all elements in that window. In practice, Max Pooling has been shown to work better.

shows an example of Max Pooling operation on a Rectified Feature map (obtained after convolution + ReLU operation) by using a 2×2 window.

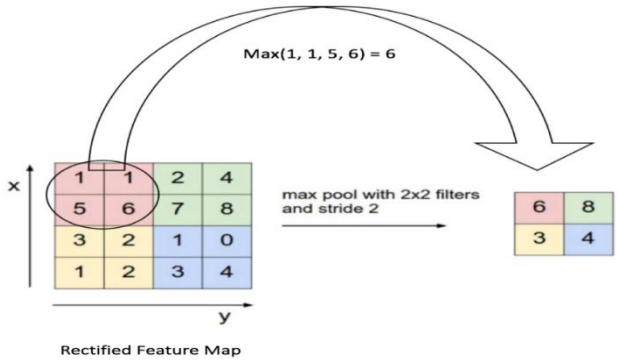


Fig: Max Pooling.

We slide our 2 x 2 window by 2 cells (also called 'stride') and take the maximum value in each region. As shown in **Figure 10**, this reduces the dimensionality of our feature map.

In the network shown in fig..

V. RESULTS AND DISCUSSION

1. UCSD Perl Dataset Benchmark

The UCSD Ped1 dataset [13] provides 34 short clips for training, and another 36 clips for testing. All testing clipshaveframe-levelgroundtruthlabels,and10clipshave pixel-level ground truth labels. There are 200 frames in each clip. Our configuration is similar to that of [13]. That is, the performance is evaluated on frame- and pixel-levels. We show the results via running time comparisons across various methods as referenced.

2 Running time comparison on UCSD dataset

We compare the running time in Table 6.1. The detection time per frame and working platforms of [13, 5, 2] are obtained from the original papers.

VII.CONCLUSION

The goal of this project is to develop a technique by which we will be able to detect abnormal behavior efficiently. It is clear that different techniques balance certain trade-offs between computational complexity, speed and accuracy of recognition and overall practicality and ease-of-use. Convolution is now a very good efficient and developing platform for image and video classification. Classification is the basic criteria of convolutional neural networks. Rather going for direct algorithm implementation which is bit costly and inefficient we used convolution to detect abnormal behavior. System with our work included is easy and efficient and can be implemented easily in any environment. Moreover, our system can be updated inexpensively anytime. Furthermore, our proposed system will be adaptable in any dynamic environment. If an environment changes for our system it will not be a problem to reinitiate the system again. We tried to ensure maximum security with the help of the recent technology along with the help of machine learning platform. Experiments will provide best accuracy in normal identification and abnormal behavior detection by using Convolutional neural networks.

VIII. FURTHER SCOPE

We have future improvement plan regarding this project. As the availability of different reliable datasets is increasing and thus more training on images of different scenarios will significantly increase the efficiency of our system in detection of abnormal behavior. Security is the most important need of our daily life. Our project is able to give any sort of protection and in future we will try to increase the percentage of detection accuracy. Further enhancement and moderation can be brought in the architecture of our system.

We would like to improve the following side of our project.

a) Advanced security system for Private property, Bank, Office building, Airport etc.

b) Real time video analysis for more accurate detection.

REFERENCES

[1]A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixedlocation monitors. IEEE TPAMI, 30(3):555–560, 2008.

[2] B. Antic and B. Ommer. Video parsing for abnormality detection. In ICCV, pages 2415–2422, 2011.

[3] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal cooccurrences. In CVPR, 2009.

[4] D. Bertsekas. Nonlinear programming. Athena Scientific Belmont, MA, 1999.

[5] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction costs for abnormal event detection. In CVPR, pages 3449–3456, 2011.

[6]X.Cui,Q.Liu,M.Gao,andD.Metaxas.Abnormaldetection using interaction energy potentials. In CVPR, pages 3161–3167, 2011.

[7]E.EhsanandR.Vidal. Sparsesub space clustering. In CVPR, 2009.

[8] F. Jianga, J. Yuan, S. A. Tsaftarisa, and A. K. Katsaggelosa. Anomalous video event detection using spatiotemporal context. Computer Vision and Image Understanding, 115(3):323–333, 2011.

[9] K.JouseokandL.Kyoungmu.Aunifiedframeworkforevent summarization and rare event detection. In CVPR, 2012.

[10] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In CVPR, pages 2921–2928, 2009.

[11] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In CVPR, pages 1446–1453, 2009.

[12] C. Lu, J. Shi, and J. Jia. Online robust dictionary learning. In CVPR, 2013.

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In CVPR, 2010.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research, 11:19–60, 2010.

[15] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. InCVPR,2009.

