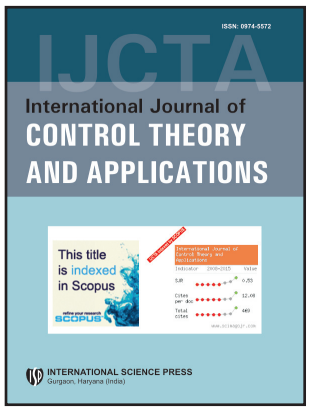

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 9 • Number 43 • 2016

Suspicious Activity Detection in Surveillance Video using Discriminative Deep Belief Network

Elizabeth Scaria[a], Aby Abahai T[b] and Elizabeth Isaac[c]

[a-c]Department of Computer Science and Engineering, M.A College of Engineering, Kothamangalam, Kerala, India. Email: [a]elizabethscaria24@gmail.com; [b]abytom@gmail.com; [c]elizabeth.issac@gmail.com

Abstract: Detection of suspicious human actions in automated video surveillance applications, is of great practical importance. Reliable classification of suspicious human movements can be very difficult due to the random nature of human movements. The primary aim of the project is to define an approach to the problem of automatically tracking people and detecting unusual or suspicious movements in Closed Circuit TV (CCTV) videos. Firstly, the videos are converted into frames. Then from the obtained frames, humans are detected from the video using a background subtraction method. Then the features are extracted using a convolutional neural network (CNN). The features thus extracted are fed to a Discriminative Deep Belief Network (DDBN). Labeled videos of some suspicious activities are also fed to the DDBN and their features are also extracted. Then the features extracted using Convolutional Neural Network (CNN) are compared against these features extracted from the labeled sample video of classified suspicious actions using a Discriminative Deep Belief Network (DDBN) and various suspicious activities are detected from the given video.

Keywords: Closed Circuit TV, Convolutional Neural Network, Discriminative Deep Belief Neural Network.

1. INTRODUCTION

The monitoring of behavior, activities, or other changing information, usually of people or places for the purpose of influencing, managing, directing, or protecting them is termed as surveillance. The surveillance methods can include observation from a distance by means of electronic equipment such as closed-circuit television (CCTV) cameras, or interception of electronically transmitted information such as Internet traffic or phone calls, and it can include simple, relatively low-technology methods such as human intelligence agents and postal interception. Many organizations and people are deploying video surveillance systems at their locations with Closed Circuit TV (CCTV) cameras for better security. The captured video data is useful to prevent the threats before the crime actually happens. These videos also become a good forensic evidence to identify criminals after the occurrence of crime. Traditionally, the video feed from CCTV cameras is monitored by human operators. These operators monitor multiple screens at a time searching for anomalous activities. This is an expensive and inefficient way of monitoring. The process is expensive because the operators are on a payroll of the organization and inefficient

because humans are prone to errors. A human operator cannot efficiently monitor multiple screens simultaneously. Also, concentration of an operator will reduce drastically as time passes. One of the methods to cope with this problem is to use automated video surveillance systems (video analytics) instead of human operators. Such a system can monitor multiple screens simultaneously without the disadvantage of dropping concentration.

The function of an automated surveillance system is to draw the attention of monitoring personnel to the occurrence of a user-defined suspicious behaviour or incident when it happens. Recognizing human actions in the real-world environment finds applications in a variety of domains including intelligent video surveillance, customer attributes, and shopping behaviour analysis. However, accurate recognition of actions is a highly challenging task due to many factors such as cluttered backgrounds, occlusions, and viewpoint variations, etc. Most of the current approaches make certain assumptions (e.g., small scale and viewpoint changes) about the circumstances under which the video was taken. However, such assumptions seldom hold in the real-world environment. In addition, most of the methods follow a two-step approach in which the first step computes features from raw video frames and the second step learns classifiers based on the obtained features. In real-world scenarios, it is rarely known what features are important for classifying the task or activity at hand since the choice of features is highly problem-dependent. Especially for human action recognition, different action classes may appear dramatically different in terms of their appearances and motion patterns as different action classes may look different when done by different personalities.

Deep learning models are a class of machines that can learn a hierarchy of features by building high-level features from low-level features. Such learning machines can be trained using either supervised or unsupervised approaches, and the resulting systems have been shown to yield competitive performance in various areas like visual object recognition, human action recognition, natural language processing, audio classification, brain-computer interaction, human tracking, image restoration, denoising, and segmentation tasks. The convolutional neural networks (CNNs) are a type of deep learning models in which trainable filters and local neighborhood pooling operations are applied alternatingly on the input images, resulting in a hierarchy of increasingly complex features. It has been shown that, when trained with appropriate regularization, CNNs can achieve superior performance on visual object recognition tasks. In addition, CNNs have been shown to be invariant to certain variations such as pose, lighting, and surrounding clutter.

2. RELATED WORK

Behavior recognition is a broad term that covers a number of categories of activities, which need different ways of detection. For example, crowd behavior, such as crowd movement, requires techniques that capture the overall characteristics of the crowd rather than the individuals in it. On the other hand, short-term human actions, such as gymnastic exercises and gestures, are often relatively simpler and even periodic. These are of a different nature and therefore require different detection techniques, involving body models and space–time shapes. This approach focuses on automatically flagging suspicious behavior in public transportation systems. These types of behavior may occur over a significant period of time. They often involve more than one object; therefore, such matters as finding trajectories, identity tracking, and object classification must be addressed.

Elhamod and Levine [1] proposed a complete semantics-based solution to the behavior detection problem that addresses the whole process from pixel to behavior level. Furthermore, the processing is achieved in real time. In this method it is assumed that the foreground blobs are extracted in each frame using a conventional background subtraction method. These blobs represent the silhouettes of animate (e.g., people) and inanimate (e.g., luggage) objects in the scene, which are the semantic entities associated with the events described. However, in practice, it is noted that a single blob will often represent multiple objects occluding or standing next to each other. After all blobs have been extracted, inferences are made to segment, track, and classify the objects that

they represent. Finally, the anomalous events must be labeled. Thus a 3-D object level information is obtained by detecting and tracking people and objects using blob-matching technique. Color histograms are used for blob matching. Kalman filter is used for tracking purpose. Then, the activities are classified into various predefined suspicious activities using semantics-based approach. The semantics-based approach replaces the need for training with a more straightforward process based on human reasoning and logic.

Kim [2] deal with detecting and tracking multiple moving objects through a single camera. The proposed method uses red-green-blue (RGB) color background modeling to extract moving regions. Blob labeling is used to group moving objects. This method is suitable for the real-time surveillance system because of the fast computation and is robust against the environmental influences. To detect the moving objects, RGB BM with a new sensitivity parameter was employed to extract moving regions, morphology schemes to eliminate noises, and blob-labeling to group the moving objects. To track the groups of the moving objects, a tracking algorithm was proposed consisting of the prediction of the position of each group, the recognition of the same group, and the identification of newly appearing and disappearing groups.

The procedure of detecting moving objects from the input image consists of the extraction stage based on RGB BM and morphology, and the grouping stage based on blob-labeling. In general, the extraction of moving regions from sequential images is carried out by using BM. This kind of BM involves the loss of image information compared with the color BM using RGB and hue-saturation-intensity (HSI) color space models.

Some work also involves using a stereo camera [3]. A stereo camera is a set of two cameras mounted adjacent to each other, but separated by a small distance. These cameras face in the same direction, but images captured by them differ slightly due to spatial difference between their lenses. These images can be processed together to get 3-D locations of the objects. Such types of cameras allow for good occlusion detection and removal. But, the computational costs are increased as we are required to process two similar images. This approach aims to segment a group of people into individual human object and track them across the video sequence with high accuracy.

3. PROPOSED WORK

The proposed system uses a Convolutional Neural Network for extracting different features from the videos and a Discriminative Deep Belief Network for classifying the recognized actions into normal activities and suspicious activities. The proposed architecture of the system is as shown in Figure 1.

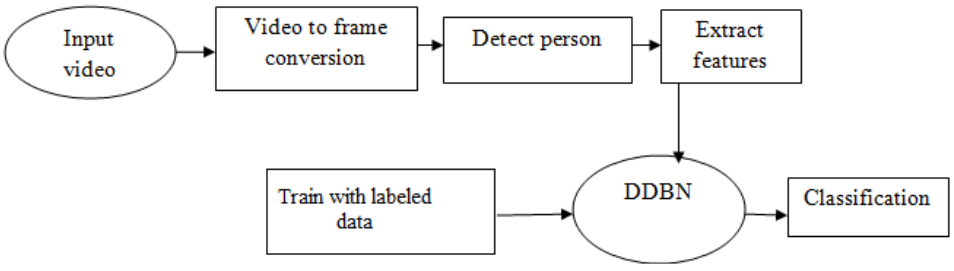

Figure 1: System architecture

In the proposed system, videos are converted to frames. Then humans are detected from these frames by using a Background Subtraction method.

A. Background Subtraction

Background subtraction, also known as Foreground Detection, is a technique in the fields of image processing and computer vision wherein an image's foreground is extracted for further processing (object recognition etc.).

Generally an image's regions of interest are objects (humans, cars, text etc.) in its foreground. After the stage of image preprocessing (which may include image denoising, post processing like morphology etc.) object localisation is required which may make use of this technique. Background subtraction is a widely used approach for detecting moving objects in videos from static cameras. The rationale in the approach is that of detecting the moving objects from the difference between the current frame and a reference frame, often called "background image", or "background model". Background subtraction is mostly done if the image in question is a part of a video stream. Background subtraction provides important cues for numerous applications in computer vision, for example surveillance tracking or human poses estimation. However, background subtraction is generally based on a static background hypothesis which is often not applicable in real environments. With indoor scenes, reflections or animated images on screens lead to background changes. In a same way, due to wind, rain or illumination changes brought by weather, static backgrounds methods have difficulties with outdoor scenes. As a basic, the background image must be a representation of the scene with no moving objects and must be kept regularly updated so as to adapt to the varying luminance conditions and geometry settings. Simple methods such as the running Gaussian average or the median filter offer acceptable accuracy while achieving a high frame rate and having limited memory requirements.

B. Convolutional Neural Network

As the next step in the procedure, features of the frames in which the humans are detected are extracted using a Convolutional Neural Network (CNN). The architecture of the Convolutional Neural Network is as shown in Figure 2 below.

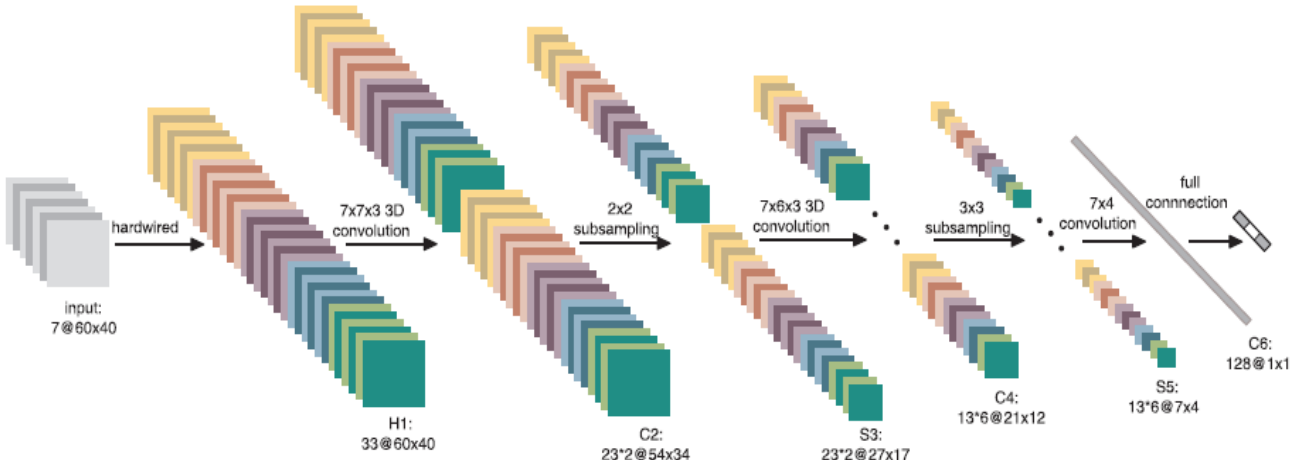

Figure 2: CNN architecture[4]

In this architecture, seven frames of size 60×40 is considered centered on the current frame as inputs to the 3D CNN model. First a set of hardwired kernels is applied to generate multiple channels of information from the input frames. This results in 33 feature maps in the second layer in five different channels denoted by gray, gradient-*x*, gradient-*y*, optflow-*x*, and optflow-*y*. The gray channel contains the gray pixel values of the seven input frames. The feature maps in the gradient-*x* and gradient-*y* channels are obtained by computing gradients along the horizontal and vertical directions, respectively, on each of the seven input frames, and the optflow-*x* and optflow-*y* channels contain the optical flow fields along the horizontal and vertical directions, respectively, computed from adjacent input frames. This hardwired layer is employed to encode the prior knowledge on features, and this scheme usually leads to better performance as compared to the random initialization.

Then 3D convolutions are applied with a kernel size of $7 \times 7 \times 3$ (7×7 in the spatial dimension and 3 in the temporal dimension) on each of the five channels separately. To increase the number of feature maps, two sets of different convolutions are applied at each location, resulting in two sets of feature maps in the C2 layer each consisting of 23 feature maps. In the subsequent sub sampling layer S3, 2×2 sub sampling is applied on each of the feature maps in the C2 layer, which leads to the same number of feature maps with a reduced spatial resolution. The next convolution layer C4 is obtained by applying 3D convolution with a kernel size of $7 \times 6 \times 3$ on each of the five channels in the two sets of feature maps separately. To increase the number of feature maps, three convolutions are applied with different kernels at each location, leading to six distinct sets of feature maps in the C4 layer, each containing 13 feature maps. The next layer S5 is obtained by applying 3×3 sub sampling on each feature map in the C4 layer, which leads to the same number of feature maps with a reduced spatial resolution. At this stage, the size of the temporal dimension is already relatively small (3 for gray, gradient-*x*, gradient-*y*, and 2 for optflow-*x* and optflow-*y*), so we perform convolution only in the spatial dimension at this layer. The size of the convolution kernel used is 7×4 so that the sizes of the output feature maps are reduced to 1×1. The C6 layer consists of 128 feature maps of size 1×1, and each of them is connected to all 78 feature maps in the S5 layer.

After the multiple layers of convolution and sub sampling, the seven input frames have been converted into a 128D feature vector capturing the motion information in the input frames. The output layer consists of the same number of units as the number of actions, and each unit is fully connected to each of the 128 units in the C6 layer.

C. Discriminative Deep Belief Network

The features extracted using the convolutional neural network are compared against the features extracted from labeled sample video of classified actions. That is using a Discriminative Deep Belief Network[5], the system will be trained using either supervised or unsupervised algorithms to classify the recognized actions into user specified suspicious activities. For the training purpose sample labeled video data sets are used.

Classification under insufficient labeled data is a well known hard problem. Unfortunately, this is likely to occur since obtaining the labeled data is often difficult, expensive or time consuming. For example, in content based image retrieval, a user usually poses an example image as a query and asks the system to return similar images. In this case, there are many unlabeled images existing in a database, but there is only one labeled example, i.e. the query image. To address this problem, semi supervised learning, which uses large amount of unlabeled data together with labeled data to build better learners, has attracted more and more attention. Typical semi-supervised methods include: self-training, Expectation-Maximization (EM) algorithm with generative mixture models, transductive support vector machine, graph-based methods, and co-training. Currently, most of semi-supervised techniques use shallow architecture to model the problem, such as kernelized linear model. As argued by several researchers, deep architecture, composed of multiple levels of non-linear operations, is expected to perform well in semi-supervised learning because of its ability to model hard artificial intelligence tasks. Weston simply leveraged shallow semi-supervised algorithms to deep architecture by plugging them into any layer of the architecture as regularizers. And the empirical validation for real classification tasks yielded competitive performance. Inspired by the study of semi- supervised learning and deep architecture, this paper proposes a novel semi-supervised classifier named discriminative deep belief networks (DDBN), based on a representative deep algorithm deep belief networks (DBN). DBN constructs a directed belief nets with many hidden layers based on a greedy, layer-wise unsupervised learning phase. Then, the resulting architecture is refined using a gradient-descent based supervised method. The two-stage construction of DBN makes it natural to semi-supervised learning. Moreover, DBN-rNCA, which combines the DBN and Neighborhood Component

Analysis (NCA) techniques, also demonstrates the good performance for classification task via semi-supervised learning. DDBN proposed in this paper utilizes a new deep architecture for classification and an exponential loss function, aiming to maximize the separability of the classifier. Moreover, we apply DDBN to various image classification tasks successfully. Hence various suspicious activities are detected.

D. Dataset

Many datasets are available among which the datasets used for the project is a real time dataset. The real time datasets are collected from PETS 2007, CAVIAR, and various videos from youtube etc. The real time dataset consists of three suspicious action classes such as unattended luggage, fighting, and loitering.

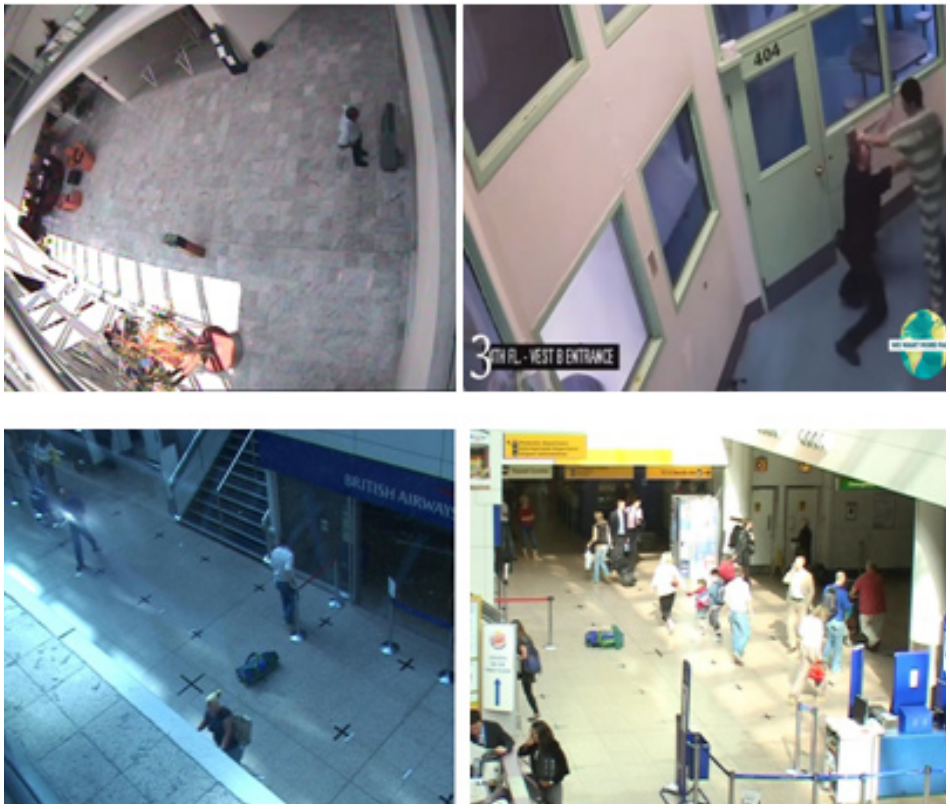

Figure 3: Dataset

4. CONCLUSION

The work proposed a suspicious activity detection from the surveillance video using convolutional neural network for feature extraction and a discriminative deep belief network for action classification. Compared with the previous works, the proposed approach achieves better classification by deep-learning-based model. After humans are detected using a background subtraction method, seven frames are selected in which the area of the bounding boxes calculated for the humans detected are larger among all. From these seven selected frames, 33 feature maps are extracted which are in five different channels defined by, gray channel, gradient-*x*, gradient-*y*, optflow-*x* and optflow-*y* channels. These 33 feature maps are given as input to the CNN which returns a 128D features in a single vector. Then this output is fed to a DDBN which is used to classify the recognized actions into normal and suspicious actions by training the system using semi supervised learning method. The deep learning model ensures more accuracy and lesser false positives.

REFERENCES

[1] Mohannad Elhamod, Member, Martin D. Levine. Automated Real-Time Detection of Potentially Suspicious Behavior in Public Transport Areas. IEEE Trans. Intelligent Transportation Systems. Vol. 14, June 2013.

[2] Jong Sun Kim, Dong Hae Yeom, Young Hoon Joo. Fast and Robust Algorithm of Tracking Multiple Moving Objects for Intelligent Video Surveillance Systems. IEEE Trans. Consumer Electronics, 2011.

[3] Qian Zhang and King Ngi Ngan. "Segmentation and Tracking Multiple Objects Under Occlusion From Multiview Video," IEEE Trans. Processing, Vol. 20, No. 11, November 2011.

[4] Shuiwang Ji, Wei Xu, Ming Yang. 3D Convolutional Neural Network for Human Action Recognition. pattern analysis and machine intelligence, IEEE Trans. Vol. 35, January 2013.

[5] P Shusen Zhou, Qingcai Chen and Xiaolong Wang, "Discriminative Deep Belief Networks for Image Classification" in Proc. IEEE. Image Processing, September 2010.