A Review on Human Action Recognition in Surveillance Videos

Hemangee De, Aanisha Bhattacharyya, Rachna Agarwal, Sudeshna Roy Chowdhury
Institute of Engineering & Management, Kolkata, Salt Lake City, Kolkata, West Bengal-700091

ABSTRACT: Nowadays automated detection and recognition of human beings and their actions in surveillance videos is gaining importance. The paper presented is divided into different segments which include methods of extraction of images of different moving objects from surveillance videos, identification and face and action recognition. The smart surveillance system proposes information processing techniques for CCTV based surveillance systems. It gives information about the advanced surveillance cameras having the property of identifying and providing real time notification. This paper is a study on the methods and applications of the human action recognition through surveillance system. It can be most effectively used in surveillance of crowded public places, finding missing and suspicious people and recognizing their actions, human computer interaction and automated refereeing.

KEYWORDS: Video Surveillance, CCTV, human computer

https://doi.org/10.29294/IJASE.6.S1.2019.60-63 **© 2019 Mahendrapublications.com, All rights reserved**

1. INTRODUCTION

The first technique of computer vision processing for human detection was Convolutional Neural Network. Time delay neural network (TDNN) was the first convolution network introduced [1]. They are fixed-size convolutional networks that can share weights along the time based dimension. It allows speech signals to be processed without varying time, similar to the translation invariance offered by Convolution Neural Network (CNN's). They were introduced in the early 1980s. Similarly, a shift invariant neural network was put forward for recognition of image characters in 1988. In 1991 it was modified and was used for medical image processing and automatic detection of breast cancer in mammograms. Another convolution-based design was proposed in 1988 which was applied to decompose 1D electromyography convolved signals via de-convolution [2]. Another met`hod for this is studying skeleton joints called temporal pyramid or RGB-D dataset. Depth Maps and Postures are used for Human Action Recognition using Deep Convolution Neural Networks. This is a more effective method. It is experimentally proved to result in 91.86% accuracy in action recognition compared to previous method.

To ensure better protection to people this field of surveillance is developing rapidly. It is widely used to ensure the safety of elderly people using the fall detection method, preventing methods of child trafficking using smart surveillance, detecting interpersonal crime and actions of harassment, human gait characterization and face recognition. This work is divided to different segments which include detection and segmentation of object which is in motion, human action classification, human posture recognition, face detection, face-feature extraction and matching with faces of required person. Depth convolution Neural Networks give more accurate results of the same. Considering these benefits this paper presents a review over different papers to put up a complete idea of the same.

The main objective of this paper is to provide a comprehensive study on the different steps involved in the process of human action surveillance. The system can trace location of such identification and provide real time notifications. The advanced surveillance cameras have the properties of 360 degrees field of view, facial recognition, night vision, smart phone integration and all weather radio frequency surveillance. This method involved in moving object extraction from videos. Then the various methods adopted for action detection is explained. Which is followed by the applications of the same in today's world.

2. METHODS OF MOVING OBJECT EXTRACTION FROM VIDEOS

2.1 Convolutional Neural Network (CNN)

Convolutional neural network is a type of feed artificial neural network in which the connectivity pattern between its neurons5 depends on the organization of the animal visual cortex [3]. Individual cortical neurons respond to stimuli only in a limited region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the complete visual field. CNN can be improved on different aspects such as layer design activation function, regularization, optimization and fast computation. These are deep models which are highly memory demand and time consuming and includes manually collecting labelled data set which requires huge amounts of human labours.

2.2 Background Subtraction

Background subtraction is a popular method to

**Corresponding Author: daisy4700@yahoo.com*
Received: 21.05.2019 **Accepted:** 18.06.2019 **Published on**: 20.07.2019

detect an object as a forefront, extracting it from a scene of a surveillance video [4]. It detects moving objects from the difference between the current frame and reference frame using pixel-by-pixel or block-by-block fashion. Other related works are Mixture of Gaussian model method proposed by Stauffer and Grimson which is sensitive to dynamic changes of scenes derived from illumination changes, extraneous events, etc. Temporal differencing includes three important modules: block alarm module, background modelling module and object extraction module.

2.3 Optical Flow Based Object Detection

Optical flow uses vector based techniques to estimate motion in a video by matching points on objects over image frames [4]. Assuming brightness constancy and spatial smoothness, optical flow is also used to describe coherent motion of the points in image frames. The characteristics of flow vectors of moving objects are used for detecting moving region in image sequence. It is useful for crowd analysis and conditions that contain dense motion.

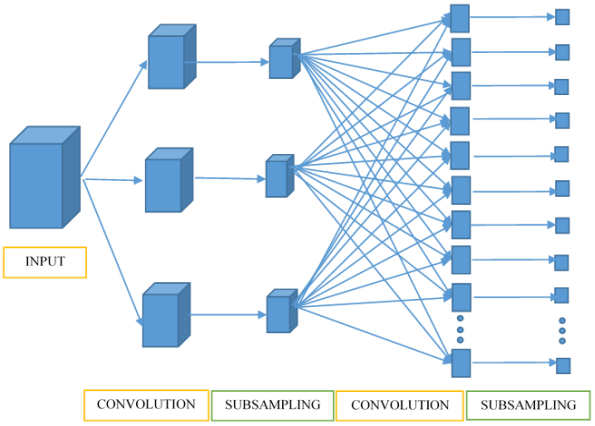

Figure represents the method of object differentiating by pixel-by-pixel or block-by-block method using Convolution Neural Network

2.4 Motion Detection

Motion detection is performed by 3D spatio-temporal data volume [4] covered by moving person in an image sequence. Here they consider motion as a whole to characterize its spatio-temporal distributions. The video sequences are handled using spatial Gaussian and derivatives of Gaussian on the temporal axis. Due to this result the filter shows high responses at region of motion. Such a feature captures the motion and its corresponding spatial information compactly.

Zhang and Parker [5] proposed a technique of 4D local spatio features for improved motion detection. A response function is computed at each pixel using both intensity and depth information within a hyper 4D cuboid. The location of a feature is determined on both the intensity image sequence and the depth image sequence.

2.5 Feature Description

A hyper 4D cuboid is centred at each feature point. To get a descriptor for each 4D hyper cuboid, the intensity and depth gradients are calculated. Thus, in general, the features are intractable. To solve this problem, the principal component analysis (PCA) is applied. PCA is a dimensionality reduction method, which projects each feature vector to a lower dimensional space. A spatio-temporal codeword is then defined to be the centre of a cluster, and the codebook is defined to be the set that contains all the code words. Thus, each extracted feature vector can be assigned to a codeword, and each video sequence can be represented as a bag of code words from the codebook.

In future, developing more sophisticated descriptors with the ability to adjust the size of the 4D hyper cuboid adaptively to deal with the scale variations and be adopted. It will be very effective to let a moving robot gain this ability to recognize human activities.

3. METHODS ADOPTED FOR ACTION DETECTION

3.1 Using Shift Action for Human Detection

Continuously Adaptive Mean Shift (Cam Shift)is based on the mean movement calculation [4]. It uses the Hue channel to follow objects focussed around Hue Saturation Value (HSV) shaded model, objects with distinctive colours might be perceived same like that a human perceives. In case of the colour data, Cam Shift tracks questions and responses faster.

3.2 Modern Methods of Action Detection (Sub Action Descriptor)

Previous works on sub action descriptor [4] provides us complete information about a human action. Like the conventional methods give the action representation of reading for one person who is reading a book while sitting and the same action descriptor for another person who is reading a book while walking. The action information for the two people should be completely different. Firstly difference is posture that is one person is sitting, and the other is standing. Secondly difference is locomotion where one person is stationary, and the other is walking. Thus process involved in a sub action descriptor involves three levels: posture, locomotion, and gesture. Each and every level of sub-action descriptor is one convolutional neural network (CNN)-based classifier which captures different appearance

based temporal features to represent a human sub action.

Using Multi-CNN Action Classifier [6] three appearance-based temporal features are extracted from human action regions for BDI-CNN (binary difference image), MHI-CNN (motion history image), and WAI-CNN (weighted average image), respectively. The first network, BDI-CNN, takes as input the BDI and stores the shape of the actor. The second network is MHI-CNN, operates on the MHI and notes the motion history of the actor. The third network, WAI-CNN, operates on the WAI and stores both the shape and the motion history of the actor. The three CNNs are trained together for getting an accurate result of action classification.

3.3 Action Detection on ICVL Dataset

The ICVL dataset helps overcome the limitations of current action detection capabilities in real-world surveillance environments.[4]Compared to existing surveillance datasets, the ICVL dataset has more single-person actions, which were captured from indoor and outdoor surveillance footages. One important feature of this dataset is that it includes different sub-actions of multiple individuals occurring simultaneously at different space locations in the same scene. It encourages the research on multi human-action detection by proposing this multiple action dataset.

3.4 Rhombus System of Surveillance

Rhombus system is looking to upend the CCTV world with its AI-powered R1 security camera.[7] It has the ability to learn and alert users about the presence of an unidentified person. By combining Artificial Intelligence and Computer Vision, it is capable to provide a video security system unlike any other. The capabilities of this Rhombus System are described below:

1. Rhombus AI and Computer Vision enables advanced facial recognition, people analytics and custom alert.
2. Plug and Play camera can be setup, taken down, and moved in a matter of minute and has unique stability and flexibility.
3. Cloud Management, allows users to easily access their system from any computer or mobile device.

3.5 Depth Maps and Postures Detection using Convolution Neural Networks

This uses two input descriptors for representing actions [8]. The first one is Depth Motion Image (DMI), which can accumulate consecutive depth images of human action detection. The second input descriptor is moving joints descriptor (MJD). This represents how the body joints are moving during action.[2] There are three Convolution Neural Networks of DMI outputs only, both DMI and MJD outputs and the third containing MJD outputs only. These results are merged to get the final result.[9] The final result is experimentally proved to be more accurate. It is 91.86% accurate than the previous systems for human action recognition.

Table 1. Various object detection methods compared in terms of accuracy and computational time

Methods	Type	Accuracy	Computational time	Comments
Background subtraction	Mixture of Gaussian model (MoG)	Moderate	Moderate	It is simple to implement and has a good performance but not so well with dynamic background. It requires parameters defined by the practitioners. It can capture multi-modal scenarios
	Non-parametric background model	Moderate to high	Low to moderate	In dynamic background scenarios, it performs very well as compared to MoG-based algorithm except in occlusion situation. It requires significant post-processing.
	Temporal differencing	High	Low to moderate	Very
	Warping background	High	Moderate to high	Good in outdoor environment with high background motion. It does not handle occlusion well. Some variations are computationally intensive
	Hierarchical background model	High	High Low to moderate	It uses block-based and pixel-based approaches. Is quicker than pixel-based approach but quality could be compromised
Optical flow		Moderate	High	Good for crowd detection but is highly computation intensive
Spatio-Temporal filter		Moderate to high	Low to moderate	Good for low-resolution scenarios but suffers from noise issues

5. APPLICATION OF SURVEILLANCE SYSTEM

4.1 Face Recognition

Facial recognition is a type of pattern recognition where a human's face is stored in a delta base for future endeavours. Software researches in the human face recognition has done a considerable work which have been conscientious with building a robust system that concentrates on both representation and recognition using artificial neural networks. [10]The benefit of face recognition is that it is a technique that can be conducted without participants' interference; this makes it especially suitable for surveillance purposes and is used by Automatic Target Recognition, Human Traffic Census Security and Criminal Identification.

4.1 Fall Detection

Fall Detection is a new technique for healthcare emerging for better care of the elderly people. [11] The video surveillance system provides effective methods to detect personal behavior and uncertain events like falls. Nasution and Emmanuel used projection of histograms of the segmented human body silhouette by classifying the posture vectors and speed of fall to distinguish between an incident of real fall and a simple lying posture. Throme and Mignet used multi-view approach to address occlusion [4]. They created an algorithm to detect very sudden changes for this purpose. Rougier and Meunier used shape comparing technique to identify the person's silhouette along video stills. After classifying the falls are detected from daily activities using Gaussian mixture methods.

6. CONCLUSION

The basic benefits of the human detection by video surveillance system are to track thefts, entry of any suspicious person or object, finding missing people and misplaced objects, used to ensure the safety of elderly people using the fall detection method, preventing methods of child trafficking using smart surveillance, detecting interpersonal crime and actions of harassment, bullying and assaults to ensure safe public places, in crowd analysis and crowd management, human gait characterization and face recognition. Its drawbacks remain due to difficulty in detecting multiple actions through video surveillance and illumination problems due to face detection. Incorporating it with A.I(Artificial Intelligence) it can be used to develop robots to take care of babies. In future, developing more advanced descriptors with the capability to vary the size of the 4D hyper cuboid to deal with the variations in scale.

ACKNOWLEDGEMENT

We are thankful to the entire Physics faculty of our college, Institute Of Engineering and Management. Also we want to acknowledge the organizing committee of SPECTRUM, Prof. Dr. Arun Kumar Bar, Dr. Koyel Ganguly, Dr. Saswati Barman, Prof. Arnab Basu, Prof. Soumyadipta Pal, Dr. Sanhita Paul and Prof. Triparna Datta, to allow us a chance to showcase our ideas in form of this research paper.

REFERENCES

[1]. ConvolutionNeuralNetworks; https://en.wikipedia.org/wiki/Convolution_neural_network

[2]. Aouaidjia Kaml, Bin Sheng, Ping Li, Ruimin Shen; 2018. Deep Convolution Neural Network for Human Action Recognition using Depth Maps and Postures. Page1, 2.

[3]. Jiuxiang Gua;Zhenhua Wang; Tsuhan Chenc; Recent Advances in Convolutional Neural Networks.

[4]. Manoranjan Paul; Shah M EHaque; Subrata Chakraborty; 2013. Human detection in surveillance videos and its applications, in EURASIP Journal on Advances in Signal Processing.2013. Page 2, Page 6 and Page 12.

[5]. Hao Zhang and Lynne E.; 2011.4-Dimensional Local Spatio-Temporal Features for Human Activity Recognition ParkerProc. of IEEE International Conference on Intelligent Robots and Systems, San Francisco, CA.2011. Page 2, Page 3 and Page 5.

[6]. Cheng-Bin Jin; Shengzhe Li; Hakil Kim; Real-Time Action Detection in Video Surveillance using Sub-Action Descriptor with Multi-CNN; Inha University, Incheon, Korea Visionin Inc., Incheon, Korea.

[7]. Rhombus System of Surveillance; mytechdecision.com, New Smart Surveillance.

[8]. EneaCippitelli, EnnioGambi and Susanna Spinsante; 2017. Human Action Recognition with RGB-D Sensors; INTECH. 2011, Neural Network Based Face Recognition Using MatlabShamilaMantrietal in IJCSET, 1(1),6-9.

[9]. Hong-Bo Zhang, Yi-Xiang Zhang, BinengZhong, Qing Lei, Lijie Yang, Ji-Xiang Du and Duan-Sheng Chen; 2019. A Comprehensive Survey of Vision-Based Human Action Recognition Methods

[10]. Rajni Sehgal; Renuka Nagpal; Diya Burman and Shinam Bansal; 2014. Real Time Face-Recognition Based Attendance Generation and Perception Level Extraction. World Applied Sciences Journal 29 (6):805-810.

[11]. Falin Wu, Heng Yang Zhao, Yan Zhao and Haibo Zhang; 2015 Development of a Wearable Sensor Based Fall Detection System.

Selection and/or Peer-review under the responsibility of 2nd International Students' Conference on Innovations in Science and Technology (Spectrum – 2019), Kolkata

All © 2019 are reserved by International Journal of Advanced Science and Engineering. This Journal is licensed under a Creative Commons Attribution-Non Commercial-ShareAlike 3.0 Unported License.