International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 658-660
Published Online April 2020 in IJEAST (http://www.ijeast.com)

ACTION RECOGNITION USING SURVEILLANCE SYSTEM

Rishabh Paunikar, Shubham Thakare, Utkarsh Anuse
Computer Engineering
Bharati Vidyapeeth College of Engineering, Navi Mumbai

Prof. B.W. Balkhande
Computer Engineering
Bharati Vidyapeeth College of Engineering, Navi Mumbai

***Abstract*—Surveillance Systems using CCTV cameras or any other surveillance devices record the footage all the time they are active. Most of the Data recorded is idle data where no activity takes place. When an activity which has occurred is supposed to be viewed the user has to go through all of the footage to check when and what had happened. This system eases this time consuming task. It uses Action Recognition to filter the idle movement data and trims the part where activity has been recorded using movement detection and various actions. It has a prominent scope for utilization in making the process of examining the footages from surveillance systems simple. This project could be implemented at public places and could enable respective concerned departments to investigate the data at an amazingly rapid rate. Thus, in an era of increasing crimes rates, this project could be a breakthrough in untangling the evidence data and eventually assist in increasing the rate of crime resolution.[1]**

Keywords—Surveillance System, Surveillance, Machine Learning, Neural Networks, Action Recognition, Image detection

I. INTRODUCTION

In today's world security is an important concern. And also the space available to store data. Generally surveillance security systems like video cameras store all the video footage recorded. As the qualities of these cameras have been increasing there is a shortage in storage.

Also another main of these security services is that a lot of idle footage exists in these videos when nothing happens. The footage of actual activity is very less compared to the footage where nothing usually happens.

II. RELATED WORK

A. Current Surveillance Systems

The current Surveillance systems used are old and outdated. Most of them run on the same functionality. Storing video for 40 - 45 days and then deleting the old videos. This is due to storage issues. The storage is limited and cannot store so much data. Most of the data however is still footage of nothing happening. This is the biggest drawback. Around 75% of the footage in the regular CCTV systems is useless idle footage with nothing happening in it. If this data could be automatically detected and deleted a lot of storage space can be saved in the database. The storage capacity of the no. of days of video content can be increased up to three times.

B. Surveillance System using Face recognition

Some advanced secure facilities like military research facilities and parts of high secure cities like in China implement a surveillance system using facial recognition. This method requires taking information of a particulonglar person's face and using expensive high quality cameras with codes implemented in them. This system however is very heavy and not light enough to be implemented on a mobile platform or low end computers.

III. PROPOSED SYSTEM

This project aims at creating an easy to use system which can be used in surveillance systems for processing of videos and segregating the video. Our main focus is on segregating video where no activity occurs with the part where any actual activity happens. Further in depth the system should be able to recognize unusual activities which don't usually happen like people fighting, stabbing, holding a gun in hand and other such malicious activities. The front end of the system will be an easy to use interface with simple options for inputting the video and a function for processing it. It will show dialog boxes for the processed videos and any malicious contents and what time they have occurred.

IV. METHODOLOGY OF THE SYSTEM

Our proposed system is based on the idea of using a neural network with different datasets twice for computing an action. For object detection the yoloV3[3] will be used to detect humans and or any other objects and for activity recognition. The same neural network processing Yolo[2] weights will be used again which will be trained on a different dataset with human actions. Instead of building an artificial neural network for special activity recognition we train the same neural network with different weights. From the first dataset it will recognize objects like knives, guns, bricks etc. and if humans and such objects together are

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 658-660
Published Online April 2020 in IJEAST (http://www.ijeast.com)

detected, it will refer from the second dataset which has only human activities recorded so it can calculate what activity is being occurred. The plus point of this is we will have a faster system by using "you look only once" twice only if needed. This system will take less time as the first pass will have already detected where the humans are and it will also down sample the image for easier and faster calculation of what activity is being occurred.

The input for this project would be a simple video and the objects detected in it. It will recognize various activities and will be able to segregate the videos of the activities separately and or output the activities in a log file which can be used as an index to know at what time how many humans were there or at what time a particular activity has occurred.

V. BUILDING THIS PROJECT

This project is being developed in Python due to the availability in libraries and ease in programming. The main part being played is by the Yolo[2] Classifier which does object detection in images. The video processing, Frame extraction is done using Open Computer Vision Library which is a library designed originally by Intel and later supported by Willow Garage.

For Yolo[2] to work the Darknet libraries are required which are originally built in C/C++. We will be using the darkflow library created by "thtrieu" on Github. These libraries will perform a similar function for loading the Yolo[2] weights and using the Classifier. After installing the required components with Yolo [2] and Darkflow for image detection we will start working on training the model to detect various activities in human images.. This training will be separate from the original trained dataset and will be only applicable in the bounding box where a human is detected. We can try down sampling the bounding box where the human is detected so that it is easier for the classifier to calculate the activity. [5] The second trained weights will be only activities done by humans so the detection will be accurate on understanding the activity done. After the activities can be detected by the model with a reliable threshold the two models can be connected to recognize humans and then recognize the activity done by it. If an activity occurs for a particular set of frames the video can classify the activity has occurred in the time span of those particular frames. The threshold for the start of these frames can start small and to be sure if the activity has occurred the threshold in the middle frames (the maximum number of frames) can be kept high. After an activity is recognized it just a matter of presenting it to the user in a user friendly manner so it is understandable. This can be done by creating a separate file where the activity is segregated or generating a log file which will store the timings at which a particular activity has occurred.

VI. WORKING OF CLASSIFIER

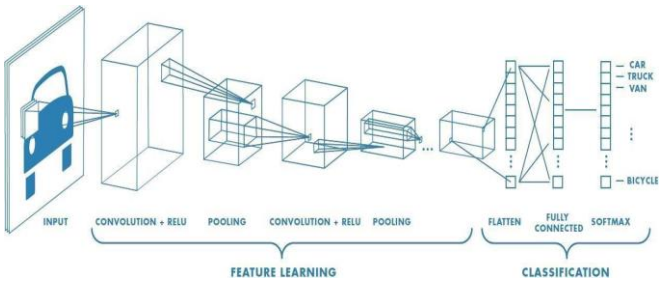

Fig 1. Diagram representing the working of Bounding boxes

Yolo9000[4] is a state of the art, real time object detection system that can detect over 9000 object categories. It is trained on the COCO test-dev. We chose this classifier because it is fast, reliable and accurate for object detection. The way Yolo[2] works is that it produces bounding boxes using dimension clusters as anchor boxes. The network predicts 4 coordinates (tx,ty,tw,th) the top left corner of image is given by (cx,cy) and the bounding box prior has a width and height pw,ph then the predictions are given by the following equations

$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$

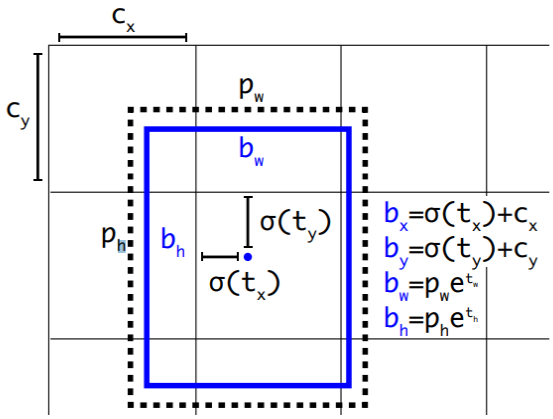

Fig. 2. Diagram representing the working of Bounding boxes

YoloV3[3] predicts object scores. It gives scores to each of the boxes. The higher a score is the more likely the object is accurately detected. When a threshold is set it will only filter boxes which it feels is accurate.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 4, Issue 12, ISSN No. 2455-2143, Pages 658-660
Published Online April 2020 in IJEAST (http://www.ijeast.com)

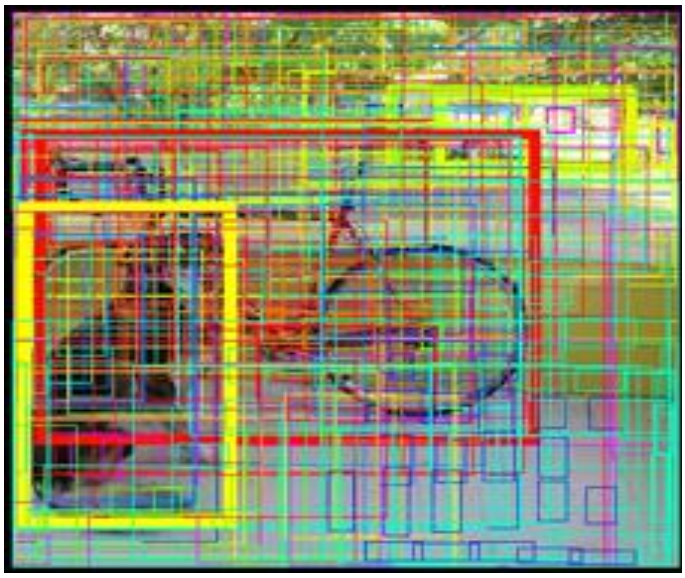

Fig. 3. Bounding boxes on an Image

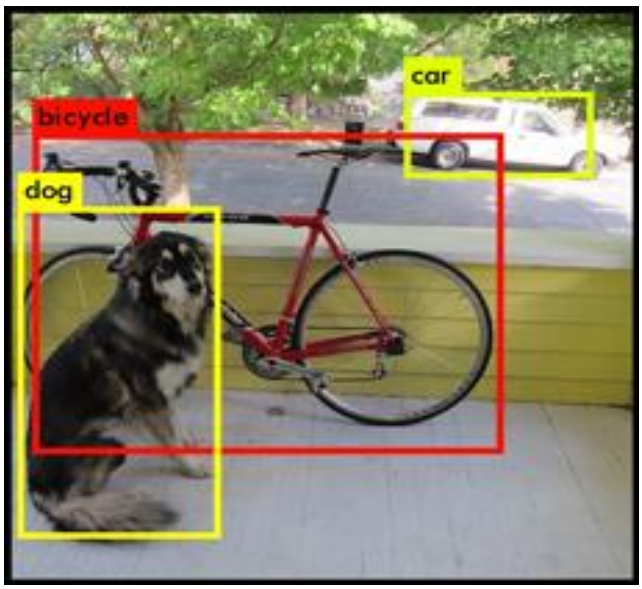

Fig 4. Bounding Boxes when Threshold is set high for same image

As shown above how the bounding boxes are created, the same algorithm can be applied on the individual frames of a video which results in having bounding boxes made dynamically on a video as shown in fig.5

Fig 5. Bounding Boxes on a series of images of a Video

VII. CONCLUSION

This paper demonstrates and evaluates the usage of long-term temporal convolutions (LTC)[1][5] and YOLO[2][3][4] for pattern matching, object detection and action recognition. Using space-time convolutions and YOLO[2][3][4] over a large number of video frames, we obtain bounding boxed that detect the object or human in frame. With consequently larger training dataset, the media output will be much more efficient.

VIII. ACKNOWLEDGEMENT

This paper was supported by our guide Prof. B.W. Balkhande. We would like to thank our Project Coordinator Prof. Kanchan Doke. We would like to thank our Head of Department, Dr. D.R. Ingle.

IX. REFERENCES

[1] Varol G., Laptev I., and Schmid C., "Long-Term Temporal Convolutions for Action" Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence

[2] Redmon J., Divvala S., Girshick R., "You Only Look Once: Unified, Real-Time Object Detection" Ali Farhadi University of Washington , Allen Institute for AI , Facebook AI Research

[3] Redmon J., Farhadi A.,"YOLOv3: An Incremental Improvement"

[4] Redmon J., Farhadi A. "YOLO9000: Better, Faster, Stronger " University of Washington , Allen Institute for AI

[5] Saha S., Rodriguez J., "Convolutional Neural Network" Intel student ambassadors for AI