Video Surveillance System with Auto Informing Feature

Lokesh Chouhan, Ekta Tyagi, Deeksha Rana, Shubham Poddar, Chandranshu Malhotra, Vikas Kumar Sahu and Jayash Verma

December 12, 2019

Video Surveillance System with Auto Informing Feature

Dr. Lokesh Chouhan, Ekta Tyagi, Deeksha, Shubham Poddar, Chandranshu Malhotra, Vikas Kumar Sahu, Jayash Verma
Computer Science and engineering
National Institution of Technology
Hamirpur, HP, India

***Abstract*—The present document represents a thorough study of making of an efficient surveillance system along with a feature of automatically informing the owner about the suspicious movement. In this moving world normally people are suffering with the availability of time, so if any crime had happened at the site, it will take many days of searching for finding actual presence of criminal. And thus a good chance for those burglars to flee away in order to protect themselves. For making the task possible we chose Python as our weapon for this battle and used different efficient techniques like, COCO data set for getting labelled and annotated images, LabelImg for making the annotation set of images, TensorFlow, Object Detection API for object Detection and Faster RCNN for training as faster RCNN has shown the highest accuracy for the COCO data set so far. Owner can be informed in two ways : Either send a message to him via mail or phone or call at the time of suspicious image capturing. Here both of these cases are used : For mail we did it via SMTP and for phone calls we used Twilio which provides us registered phone no. and we can make both outbound and inbound calls. After using all the mentioned things and making the model in a way described above we found that faster RCNN is much accurate than the other conventional methods. The results have come very good as RCNN show 86.7% accuracy and a 100% have come out with the informing module as there simply the mail will be sent to the one whose mail is given in the code and the same is for Twilio calling.**

***Index Terms*—TensorFlow, RCNN, COCO, LabelImg, Twilio, Annotation, SMTP, Preprocessing**

I. Introduction

In this dynamic world which is moving fast along with time and technology, the crimes are also increasing day by day. So the need of CCTV and surveillance systems is also increasing. In general we have a CCTV system and if a robbery or any crime is reported at the site of surveillance, the owner and the checking authority go and check for the suspected movements and till that time they might have lost a lot of goods, precious things and even life too. Also the problem gets more big if the person is deaf or not have dexterity enough to operate his mobile. The traditional systems propose to have an alarm but the problem becomes when person is not near or at home.

This document proposes a system where we are not only doing the surveillance but also informing the owner if anything suspicious is caught by the camera, the person can check his the CCTV recording on his phone at the same time and can take action accordingly. For the same, researchers have proposed many algorithms as in [1] transmittance algorithm and enhancement algorithm for the visual enhancement and visibility range algorithm for pre processing and decomposition algorithm for doing background separation but the system will detect and save the images with it. Document referred in [2] PTZ i.e Pan Tilt Zoom is used which is based upon a fine observation of moving people.

Other algorithms are slow as there we will have to extract the background but in CNN model we will not have that problem, which we are going to make a bounding box. We can have more than one objects in a one frame. So, ultimately we will need to find the object of our interest, which sounds difficult but with the iterative research and advancement in deep learning, we have exceedingly engineered neural networks which makes the task of detecting objects quite simple. To find object position precisely, our algorithm need to be capable more than enough to inspect each and every portion of image.On the informing the owner front, we have different ways in making mails to the owner but out of these all we used SMTP service and for making calls to make the paper more efficient we used Twilio. One can use a separate mail id and calling no. for the informing module as per their requirements.

II. Methodology

This paper aims to get implemented in Python language. The model uses CB Net and CNN. The model work broadly is divided into 3 categories - Data Preparation, Training the Model and Inference. The paper is completed in following steps:

step 1: Collecting training data where we collected the data set for our model on which we are going to train our model.

step 2: Annotation. Here an annotator annotates the images where bounding boxes are made for object detection.

step 3: GPU Preprocessing. Its obvious use case is rendering 3D objects.Inputs are given in TFRecords format. Then the records are processed for further training.

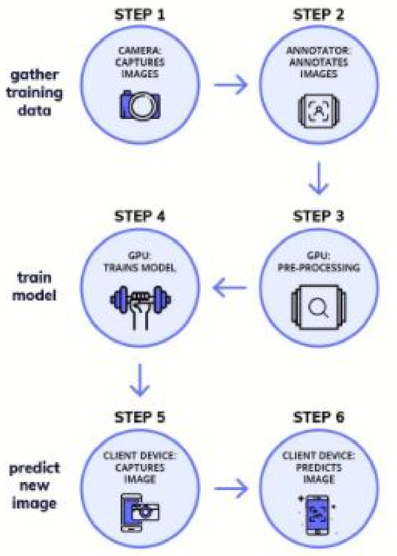


Fig. 1. Flow chart of the procedure

step 4: Training the Model. Tensor Flow Object Detection API provides facility to train model which we can fine tune to our use case. Transfer learning approach is used.

step 5: Prediction of Image. the trained model is then transferred to the client end and then the owner is informed by making a call and mail.

A. Data Preparation

1) Obtain the Data set: Data set is the most important tool in deep learning. And thus it is quite important to have righteous and useful data set for fine precision and recall Most of the time it is very hard to obtain good quality surveillance footage. In that case we need to train our data for efficient detection. For this we can collect our data in two ways — One is the basic where we have to download image and the second option is COCO model.

In the first option we have to download all the sample images enough in no. to train model over the item ob object. E.g for the basic training of the model we will need around 100 images in minimum. Now the images will be having different sizes and formats. So make size of all the images same and make them of same format.

In the second one we just need to have the COCO set which is having set of images of same sizes, good quality and same formats. We have tried both of them in our model.

2) Annotate the Data set: In the world of deep learning we have some household names which are referenced often by researchers and the practitioners also as COCO, Pascal,SUN and Image Net. In this paper we suggest to use COCO data set for the training and the self taken image set also.

COCO i.e Common Object in Context. This data set has images from our everyday scenes hence the name is 'in context'. The data set has sets of images in 91 categories. It provides labelled and annotated image set to us. After this annotation we will get object which we need to detect in bounding-box in output as shown in fig. 2 and later checking will be performed on them.

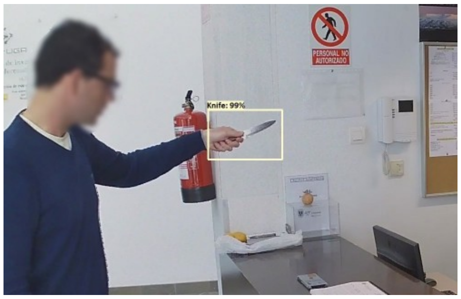


Fig. 2. object in bounding box

If we don't have the object falling in these 91 categories of the available COCO set then we will need to perform annotation on the our own image set. In this case an annotator is to be used. We have an annotator as LabelImg to perform these annotations. This provides a simple platform for making the image annotated. Now the annotated image will be used for further training.

Also, if you want to make annotations of our own we can go for divide & conquer algorithm, as results of this approach also show better and quick results than CRF, linear regression and others.

B. Training the Model

1) Convoluted Neural Networks: One of the growing fields in computer science is image detection and analysis using computer vision. One of the prominent methods to produce efficient results in this field is the use of convoluted neural networks (CNN). CNN is a type of artificial neural network that is profound and feed forward. CNN has many more, deeper layers as compared to customary neural networks. It has biases, weights and yields through nonlinear triggers. The fundamental units of CNN called neurons are arranged in three-dimensional way, i.e., as length, breadth and depth.

- **Traditional Object Detection Method :**
  Until early 2000s, the framework used for object detection was categorized into four steps as shown in the figure below :
  Generating Candidate Regions : The location of the object is detected in this stage. But, the object may be present at any location in the input image. Thus it heavily affects the performance and speed of feature extraction.
  Feature Extraction Feature : The design and performance of the classifier is hugely affected by feature extraction. But, it is very difficult to design a vigorous feature due of diverse external factors.

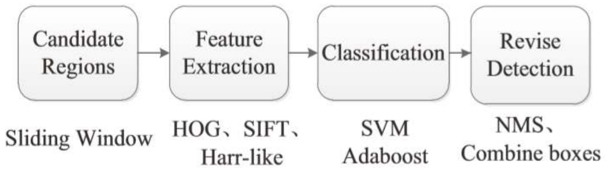


Fig. 3.

Classification: The AdaBoost or SDA classifiers are used to classify the extracted features in this stage.

Revise Detection Results : After classification, there are still a lot of redundant windows , so it is very important to remove those windows. The detection results are optimized by Non-Maximum Suppression (NMS) and overlapped Bounding box.

- **Convoluted Neural Network Architecture :**

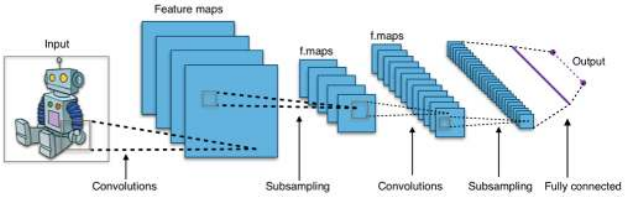


Fig. 4.

As stated earlier, a convoluted neural network has many hidden layers other than an input and output layer. The hidden layers consist of a convolution layer that involve using scalar product or dot product. The convolution layer is latterly followed by a set of pooling layers and fully connected layers. The profundity of the layers increases as we move towards the right but at the same time other dimensions like depth and width decrease in that direction. The last phase of convoluted neural network is the fully connected layer.

Convoluted neural networks take the intensity values of pixels of an image as input. They are in the form of a three-dimensional matrix denoting width, height and depth (RGB), for example [50 X 50 X 3]. The output of the neurons is computed by the convolution layer and these neurons are located in the surrounding areas of the input pixels. The parameters of the layer are consisting of learn able filters, known as kernels, which convoluted across the height and width of the input extending and lastly through its depth, calculating the scalar product between the corresponding values of the filter and the input. In sch a way, an activation map of the filter is created, which is a two-dimensional map and hence the CNN automatically learns some filters that activate when it detects the required types of feature in a particular space in the input. The Rectified Liner unit layer (ReLU) function performs element-to-element activation. The ReLU function is defined as :

$$f(x) = x,\ x > 0$$

$$f(x) = 0,\ x \leq 0$$

The function gives output as zero if input is negative, and shows linear increase in output if the value of input is positive. The volume size is not affected by this. The output of the pooling layer of the CNN. The height and width of the samples are thus reduced. The output of this is a fully connected layer which is also the final layer of the convoluted neural network. The output of this layer is probability distribution, using SoftMax activation, over various output classes.

Convoluted Neural Networks have been much successful over recent years computer vision and related areas. In 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), Hinton along with his student Krizhevsky had used Convoluted Neural networks to image classification and successfully achieved top5 error 15.3%, compared to 26.2% of the traditional image processing methods. This was a major turning point in the field of Computer Vision and Image processing. Thus, CNN was dominantly used in the later computer vision and Image processing tasks due to obvious success. Then in 2013, the R-CNN (Regions with CNN features) method was proposed by a technologist named Ross Girshick, who successfully used it in object detection.

According to the progress in this field in the present years, it is seen that the deep learning methodology can give us higher accuracy and make test-time significantly shorter than traditional method.

2) Object Detection Algorithm: Object Detection algorithm is basically identifying objects like people, animals, objects.Object's are identified on the basis of choices like face , skin, colour, target of interest.

- **Background Modelling :**

  Each pixel has possibly three values maximum Intensity, minimum intensity and maximum intensity difference between two consecutive frames. Algorithms like Extended Kalman Filter don't perform well in presence of nonlinear transformation. Mean shift does not excel because of its sensitivity to window size.

  Best methods were LOTS (Lehigh Omni directional Tracking System) and SGM (Single Gaussian Model).They improve the accuracy of foreground segmentation. Any current frame is compared to previous frame and updating is performed accordingly.
- **Object detection algorithms :**

  It consists of majorly four parts face detection, skin detection, target Detection, colour detection.

  Face detection : Viola Jones algorithm is used to detect aspects of faces like eyes, mouth , nose

Skin Detection : Skin pixels are represented by ‘1’ and non-skin by ‘0’. Methods used for colour spaces are RGB colour space, orthogonal colour space, perpetual colour space and perpetually colour space. Skin detection is performed using YCbCr model. Multiple people skin detection is tougher. It basically involves separation between brightness and colour. Read image and covert RGB to YCbCr using values (ranges) of Cb and Cr.

Target Detection : Image is captured from camera and posterior probabilities are calculated . Probabilities of any pixel belonging to foreground and background are calculated. Background subtraction is used and appropriate RGB euclidean threshold is calculated

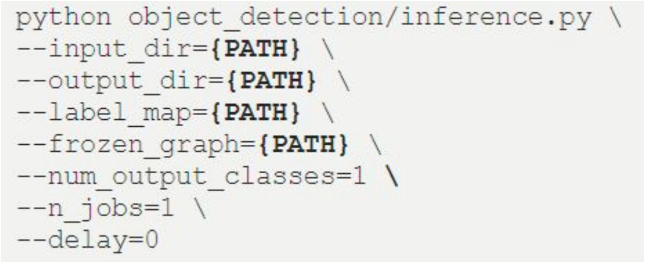
Colour Detection : Illumination is key when performing over colour models. First image pixels are recorded, then thresholding process is used to convert gray scale to binary by using already set threshold values. Now, pixels are assigned ‘0’ and ‘1’ accordingly. It helps in correct classification of object. Shape is also detected using appropriate filter over the image pixels. Different applications are border security, medical image processing, video surveillance, astronomy.

C. Inference

1) Exporting trained Model and prediction of the intuition: After training model on the taken data set we will need to export the model before using it actually. Where it is going to make the decision about the suspected objects in the video frame.

You are going to obtain a file along with a no. of checkpoints.Now extract every individual frame from the video. The task can be done using OpenCV’s method of video capturing. The data extraction code will automatically create a folder with test images.

You can run model on the data set using the following python commands:



```
python object_detection/inference.py \
--input_dir={PATH} \
--output_dir={PATH} \
--label_map={PATH} \
--frozen_graph={PATH} \
--num_output_classes=1 \
--n_jobs=1 \
--delay=0
```

Fig. 5.

Faster RCNN (Region Convolutional Neural Network) is more accurate while RFCN (Region based Full Convolution Networks) and SSD (Single Shot Multi-Box Detector) which is used basically for real-time processing are faster.

A hash function to find the difference between the video that we get as an output along with detection is used and then if a different video is detected it will inform the owner.

2) Informing the Owner: Now comes the last but not least part our paper ”Informing the user”, where we have to inform the registered owner regarding the suspicious activities. The work can be done in two ways : By mailing him or messaging him and by making a phone call. We opted for both as the person will have a written proof along with the time of the activity so that for later interception he can go for that. Now comes the second case like the burglaries mostly take place in night so in that case a person while sleeping will not check the message even if it pops multiple times, So we went for the phone call. The procedure of using both the methods is described below:

- **Informing via Mail:**
  We did the same using SMTP (Simple Mail Transfer Protocol). For sending mails either you can create a different mail id for that or you can do it using your own mail id. For making the code to send mails through email id first you need to do some setting in your mail account. First open myaccount.google.com/lesssecureapps then turn on allow less secure option. Where MUA(Mail User Agent) in which our SMTP(Simple Mail Transfer Protocol) client side being used without having connection of Internet to provide communication between client and Email server. Once the TCP connection is created between the Email client and Email server, Email message is sent to the Email server of sender using SMTP. Using similar protocol, the Email server will send the mail to the another server or directly to receiver’s Email server by Email gateway.
  Import smtp library first and then build a connection between the both using smtp.gmail.com at port 587. Feed your email id and password through which you are going to make mails. Then send mail to the person who is the owner or whomsoever you want to send.
- **Informing via Phone Call**
  Twilio is a platform for cloud communications as a service (CPaaS) company. It provides Twilio Voice which is an API to do and receive voice calls that are completely hosted in cloud. It is used for platform evangelism for acquiring customers.
  Twilio provides a platform doing programmable voice calls. You can actually build a lot different types of applications and saw lot of different types of business problems e.g Capital one has built a custom phone tree on an idea so that when people call in they get the same experience that they wanna give to their customers and modify that quickly and they are able to program what people can say in the beginning and during the call using programmable voice. Companies like amazon provide the facility directly calling through their website, instead of dialing n phone numbers the website acts as the phone so that people can get support for the merchandise they have bought.
  Twilio provides the virtual phone numbers that is programmable phone numbers, that could be local phone

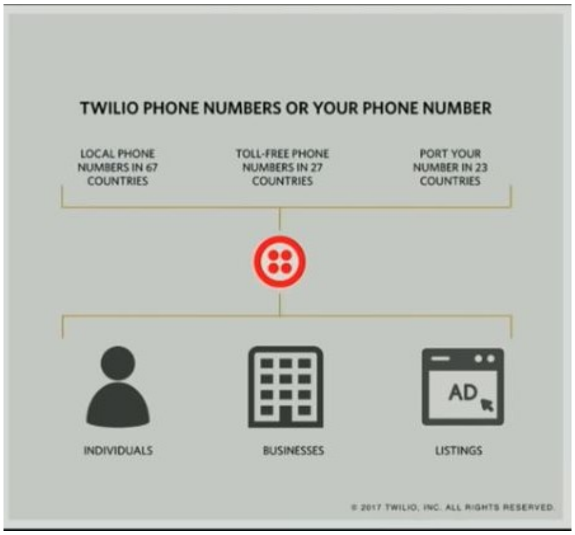

TWILIO PHONE NUMBERS OR YOUR PHONE NUMBER
LOCAL PHONE NUMBERS IN 67 COUNTRIES
TOLL-FREE PHONE NUMBERS IN 27 COUNTRIES
PORT YOUR NUMBER IN 23 COUNTRIES
AD
INDIVIDUALS
BUSINESSES
LISTINGS
© 2017 TWILIO, INC. ALL RIGHTS RESERVED.

Fig. 6. Instantly provision a Phone Number

numbers or any of the numbers from 667 countries in which the countries is operating and toll free numbers also and also provides the facility to port your phone number and make them programmable that you have an individual to make phone calls to others. And yes the very thing about the phone call is that we can control it. Twilio has maintained interconnections with nearly every carrier in the world so one can take call and terminate it into landline or a cell phone. SIP(Session Initiation Protocol is used for terminating that call into a desk phone or PBX infrastructure directly from Twilio.

In making the task possible we took REST API call using HTTP and connected it to server for doing the call through our computer as the deep learning code was running for the validation of if it is going to run as we have to make the phone call only in case of suspicious object caught.

The POST request to REST API must include all the parameters From and To to Twilio to understand where to direct desired outbound call and the information about caller ID.

Phone numbers should have a format with '+' & country code e.g., +16173852212 (E.164 format). SIP addresses need to be formatted like sip:name@example.com. E.g to dial Alice's SIP address at xyz Company, the 'To' parameter will be sip:alice@xyz.com.

Client identifiers need to begin with client:URI procedure. E.g to make a call to a client named alice, the 'To' parameter will be as client:alice.

When we have to initiate an outbound call using REST API of Twilio, Twilio will have to access our instructions for how to handle call. We do this task by making synchronous HTTP request.

If we have specified any URL parameter in our request, Twilio is going to make an HTTP request to that URL for retrieving TwiML to handle outbound call. That request from Twilio is as same as the request Twilio used to send when it is going to receive an inbound call.

When the outbound call gets connected then Twilio will make a request to the VoiceUrl set poresent on our application and this request is similar to the request that Twilio must have sent to Url parameter as described above.

III. Results and Analysis

After performing a fist full of iterations for each module and operations for it, we got a finely performing prototype. In the current module we collected the data set in 2 forms one is COCO data set and collecting the images by ourselves. The first one is best at making the annotation set as it provides us a 100% precision about the same. But the problem with it is that it is that it has annotated set for only a handful of images and to make the set for any different image, we will make the set ourselves and if the images fulfill the constraints then it is possible to make annotated set with a 100% accuracy.

RCNN are a kind of neural network that make progressively much higher level features from groups of pixels generally found in images. But still we can add some amount in improving the performance like: increasing image resolution (progressive resizing); vertical or horizontal shifting :- Randomly flip image vertically or horizontally.

How image scores on those features is now weighted to make classification result. The RCNNs are among the best image classifier algorithms we know about, and they perform specifically well when given a lot of data to work with.

Now comes the last phase that is informing the user :

In mail sending program with an accuracy of 100% it can send mails within seconds.

Also Twilio makes call to the owner instantly.

IV. Conclusion

On the journey of making the paper using a no. of tools and measuring performances according to the paper and our capabilities too, it is concluded that for making a video surveillance system providing auto informing feature, it is better to make the annotation set of your own as you know better that which kind of images you are going to get And on which angles you are going to get. Although it will be aghast as it takes 8-10 hours normally for annotating even 150 images.

For training RCNN is the best in Digital World as it provides around 87% of accuracy and continuously training for more than 8 epochs it will give even better results.

For the mailing and calling Twilio and Smtp are the best as they provide 100% accuracy for making calls and mailing respectively and the service in less than a second.

Acknowledgement

This research was supported by Dr. Lokesh Chouhan. Thanks to provide insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. Thanks for providing us this valuable opportunity to pursue this work.

REFERENCES

[1] Himanshu Kumar ; Saumik Bhattacharya ; Sinnu Susan Thomas ; Sumana Gupta ; K. S. Venkatesh, "Design of Smart Video Surveillance System for Indoor and Outdoor Scenes" Department of Electrical Engineering, IIT Kanpur 2017 22nd International Conference on Digital Signal Processing (DSP).

[2] Zhengya Xu ; Hong Ren Wu on Smart Video Surveillance System in IEEE International Conference on Industrial Technology,20,2010

[3] Yanjun Li ; Ping Guo ; Xin Xin, "A Divide and Conquer Method for Automatic Image Annotation" in 2016, 12th International Conference on Computational Intelligence and Security (CIS).

[4] R. Sureswaran ; Hussein Al Bazar ; O. Abouabdalla ; Ahmad M. Manasrah ; Homam El-Taj"Active e-mail system SMTP protocol monitoring algorithm" 2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology.

[5] Reagan L. Galvez ; Argel A. Bandala ; Elmer P. Dadios ; Ryan Rhay P. Vicerra ; Jose Martin Z. Maningo"Object Detection Using Convolutional Neural Network" 2018 IEEE Region 10 Conference in Jeju Korea.

[6] Apoorva Raghunandan, Mohana, Pakala Raghav and H. V. Ravish Aradhya "Object Detection Algorithms for Video Surveillance Applications " 2018 International Conference on Communication and Signal Processing.

[7] Ajeet Ram Pathaka,*, Manjusha Pandeya, Siddharth Rautaray"Application of Deep Learning for Object Detection " 2018 International Conference on Computational Intelligence and Data Science.

[8] Virgil Claudiu Banu, Ilona Mădălina Costea, Florin Codrut Nemtanu and Iulian Bădescu "Intelligent Video Surveillance System"2017 IEEE23rd International Symposium for Design and Technology in Electronic Packaging (SIITME) .

[9] Stephen J. Krotosky and Mohan Manubhai Trivedi"Person Surveillance Using Visual and Infrared Imagery" AUGUST 2008, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 18, NO. 8.

[10] Jun Zhang, Jianbo Gao, and Weisong Liu"Image Sequence Segmentation Using 3-D Structure Tensor and Curve Evolution" MAY 2001, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 11, NO. 5.