Comparative Study on Object Detection Methods used in Human Action Recognition

Amreen Hasheem [*1] Itisha Paul [*2] Akshaj Jain[*3] Rashmi S R[*4]
[*1,2,3,]Student, [4]Assistant Professor
[*1,2,3,4]Department of CSE, Dayananda Sagar College of Engineering,Bangloor.

Abstract— **The development of Information Communication Technology (ICT) or cyber infrastructure had growth which is very fast in producing a wide range of computer products cause some medium sized organizations are confused and ambiguous as to what should be done to the ICT infrastructure. This resulted in tragedy 'white elephant' where cyber infrastructure is purchased by the organization were not fully utilized or not used at all especially for ICT security infrastructure. While digital transformation has led manufacturers to incorporate sensors and software analytics into their offerings, the same innovation has also brought pressure to offer clients more accommodating appliance deployment options. So, their needs a well plan to implement the cyber infrastructures and equipment. The cyber security play important role to ensure that the ICT components or infrastructures execute well along the organization's business successful. This paper will present a study of security management models to guideline the security maintenance on existing cyber infrastructures. In order to perform security model for the currently existing cyber infrastructures, combination of the some security workforces and security process of extracting the security maintenance in cyber infrastructures. The implemented cyber security maintenance within security management model in a prototype and evaluated it for practical and theoretical scenarios. Furthermore, a framework model is presented which allows the evaluation of configuration changes in the agile and dynamic cyber infrastructure environments with regard to properties like vulnerabilities or expected availability. In case of a security perspective, this evaluation can be used to monitor the security levels of the configuration over its lifetime and to indicate degradations. The focused on the cyber security maintenance within security models in cyber infrastructures and presented a way for the theoretical and practical analysis based on the selected security management models. Then, the proposed model does evaluation for the analysis**

which can be used to obtain insights into the configuration and to specify desired and undesired configurations.

Keywords— ***Human Action Recognition, Object Detection,CNN, RCNN , YOLO.***

I. Introduction

Human action recognition is the recognition of different types of actions performed by humans. Object detection is an important aspect of the human action recognition process. Thus, we went on a quest to compare the different object detection models from Convolutional Neural Networks(CNN) to YOLO(You Only Look Once) Neural Networks.

Object detection is a part of computer vision which deals with identifying different classes of objects and categorizes each object into it's own class. Object detection has seemingly infinite use cases which range from medicinal field to the agricultural field, spanning almost all the fields.

It can be used for security purposes, detecting diseases, traffic management as well crowd control. In order to analyze the existing object detection methods, their application domain and error rates, it is necessary to understand the shortcomings of the earlier methods which formed the basis for the implementation for the present day object detection method. We will also walk through the different standard datasets on which the models are tested upon. Viola-Jones algorithm was the very first object detection technique, which used a minimalistic Support Vector Machine (SVM) classifier to detect faces, but didn't prove to be effective. (Histogram Of Gradients (HOG)s employed the technique of assigning gradients to each pixel based on the relative darkness of the surrounding pixels. However, it proved to be very sensitive to image rotation. As the era of deep learning began, Artificial Neural Networks (ANN) developed which had the ability to learn and model non-linear and complex relationships similar to how it is represented in the human brain. CNN, being less complex, was more efficient in terms of cost and space as it reduced the size of Neural Nets (having billions of neurons) considerably. Region-CNN (R-CNN) created a bounding box around objects before feeding it into the convolution network which reduced the identification time but could not be employed to real time detection. AlexNet was the first algorithm to use deep neural networks in object detection consisting of stacked convolution network which was deeper and had more filters per layer which made training very time consuming. Pose-based (P-CNN)is an optimal method to identify fine gained actions where the variations in actions are very subtle with relatively high speed and increased accuracy. VGG Net is another method popularly used to extract features from the images and has a very uniform architecture. Another algorithm developed which was very close to the accuracy of object identification by humans was GoogleNet (a.k.a Inception V1) which was comprised of 22 layers and proved to be faster and more economical than VGG Net.YOLO took a completely different approach to object identification and classification that out-performed R-CNN and all its variants.This is the reason for YOLO being so fast and its wide application in real time object detection.

II. Index

5.1) MNIST

5.2) MS-COCO

5.3) ImageNet

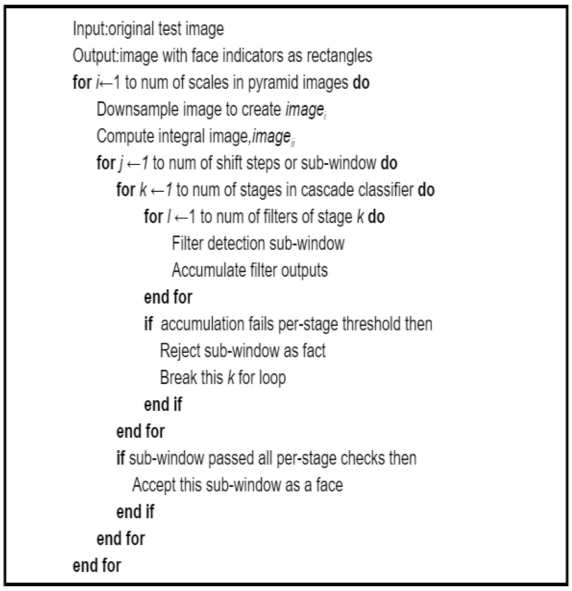
5.4) SVHN

5.5) CIFAR-10.

III. LITERATURESURVEY

A. Viola- Jones Algorithm:

The idea of object detection existed in the early 60s. However, the first ever effective object detection through a web cam was done in 2001 by Paul Viola and Michael Jones. The general idea involved hard-codingfeatures such as eyes, nose, ears and the relationship between these features present in a frame and pass it as an input to a SVM classification model. The SVM is a discriminative classifier that differentiates between a real and fake face. This algorithm proved very efficient in detecting faces and formed the basis to early facial detection methods. The main drawback was that this algorithm fell short in detecting faces in other angles or in configurations that differed from the hand-coded features. Pseudo code of the algorithm is as follows:



```
Input:original test image
Output:image with face indicators as rectangles
for i←1 to num of scales in pyramid images do
   Downsample image to create $image_i$
   Compute integral image,$image_{ii}$
   for j ←1 to num of shift steps or sub-window do
      for k ←1 to num of stages in cascade classifier do
         for l ←1 to num of filters of stage k do
            Filter detection sub-window
            Accumulate filter outputs
         end for
         if  accumulation fails per-stage threshold then
           Reject sub-window as fact
           Break this k for loop
         end if
      end for
      if sub-window passed all per-stage checks then
        Accept this sub-window as a face
      end if
   end for
end for
```

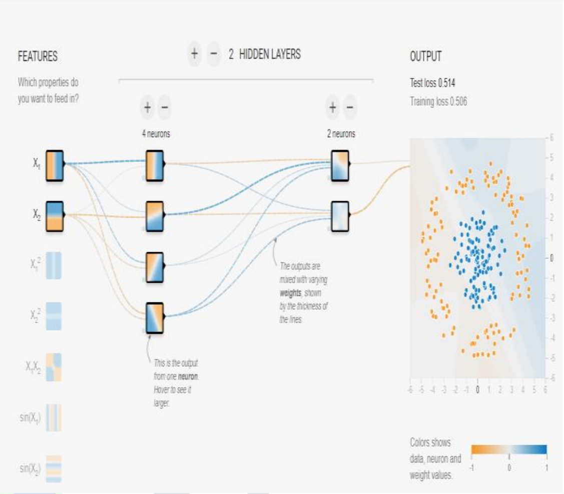
Fig.1. Viola-Jones algorithm

B. HOG(Histograms of Oriented Gradients):

Navneet Dalal and Bill Triggs invented a more efficient technique called HOG in the year 2005, specifically for pedestrian detection. It used the same hand-coded features, however, the feature descriptor (HOG) was better in terms of performance than older algorithms. The principle involved in examining the pixels directly surrounding each pixel in an image and analysing the darkness of the current pixel when compared to the surrounding pixels. Consequently, an arrow, called gradient, was assigned to each pixel in the direction in which the image gets darker. These gradients corresponded to a flow from light to dark across the entire image frame. The entire image was divided into a 16x16 pixels grid and arrows in each cell were replaced by a single arrow in the direction of the most prevalent arrows. The final image is a representation that captures the basic structure of an object. A new object is identified based on how similar it is to an existing feature map based on basic euclidean distances. This feature maps generated using the gradient technique is very similar to the training sets of convolutional neural network (CNN).

C. Artificial Neural Network (ANN):

ANN processes information and represent complex relationship similar to human brain. It exploits non-linear relationship between variables and consists of artificial neurons connected in a way to store information. Composed of mainly three components which are neurons, layers and activation function. ANN consists of three layers which Input layer, Hidden layer and Output layer. Each neuron is connected to several other neurons from the previous layer. Input layer is where data is supplied to the network. Hidden layers does the processing based on activation function. Output layer contains output which are dependent variables.Activation function is the process of manipulating input to obtain desire output, also called transfer function provides mapping from input to hidden layer and hidden to output. The two most common types of ANNs are Convolution Neural Network(CNN) and Recurrent Neural Network(RNN).



The outputs are mixed with varying **weights**, shown by the thickness of the lines.

This is the output from one **neuron**. Hover to see it larger.

Colors shows data, neuron and weight values.

Fig.2. artificial neural network

D. Convolution Neural Network(CNN):

In the year of 2012 the era of deep learning began and convolutional neural networks became very popular for image classification. With this method image classification became simpler and results were impressive. In most of the scenarios we not only classify objects but also detect and classify them. we define boundaries for the objects and its difference from other objects surrounding it. We then proceed with classification. Usually convolutional network classifier like VGG net slides through small window across the image and keeps classifying each window of the image and ends ups with a bunch of classifications and keep only those which the classifier is most certain about and use that to draw a bounding box around the image. This approach is very slow since we need to run the classifier many times and an expensive method.

The most important part of CNN is the convolutional layer where each input image passes through series of convolution layers with filters (kernels). Convolution is the operation of merging two information sets and in CNN it is convolution filter is applied on input data to produce a feature map.

Convolution is the first layer which extracts features from the image given as input, applying different filters with stride on convolution of image performs different operation such as edge detection. In some cases, the filter may not fit perfectly and padding is applied to drop the part of image where filter did not fit and then ReLU is applied to thematrix. Next is the pooling layer which reduces the number of parameters when the images are too large. The output is then flatten and feed into the fully connected layer and finally we have activation function which classifies the output.

E. Regions with Convolution Neural Network Features(R-CNN):

R-CNN was proposed by Ross Girshick et al which uses selective search to creates a bounding box or region proposals in an image before feeding them to the convolutional network. Selective search looks at image from different sized window and for each size it tries to group together the adjacent pixels by texture, colour or intensity to identify objects. R-CNN first generates a set of region proposals for the bounding boxes. The bounding boxes are then run through a CNN to compute features of the bounding box and SVM is applied to classify the image in the bounding box. Linear regression model is run through the box after object Is classified to tighter coordinates for the box. R-CNN cannot be implemented in real time as it takes around 47 seconds for each test image. It is also quiet inefficient as it takes a huge amount of time to train the network. There are many improvements of R-CNN such as fast R-CNN, Mask R-CNN etc.

F. AlexNet:

AlexNet was introduced in 2012 and was the first algorithm to use deep convolution neural network for classification of images and has very similar architecture as LeNet by Yann LeCun et al but is deeper, with more filters per layer and has stacked convolutional layers. AlexNet consists of 5 convolutional layers, dropout layers, 3 pooling layers and 3 fully connected networks. It attached ReLU activation after every convolutional and fully-connected layers. Multiple convolutional kernels extract features in an image with many kernels of same size in a single convolutional layer. The first two convolutional layers are followed by overlapping max pooling and the third, fourth and fifth layers are directly connected where fifth convolutional layer is followed by a max pooling layer whose output goes into a series of two fully connected layers. The seconds fully connected layer feeds into softmax classifier with 1000 class labels. ReLU nonlinearity is applied after all convolution and fully connected layers. It took almost six days to train AlexNet on two GTX 580 3GB GPUs as it has 60 million parameters and 650,000 neurons but todays there are more complex CNNs that can run efficiently on faster GPUs even with large datasets, but back in 2012 it was huge.[1]

G. Pose Based - CNNs (P-CNN):

Methods which recognize actions based on static representation of the motion are very successful for recognizing actions which are not fine or coarse but not optimal in recognizing subtle variation in action. For fine grained recognition of images, one should focus on the structure and spatial alignment at the pre-processing step. P-CNN is based on keeping track of body joints over time, it combines motion and appearance feature for body parts. CNN features are applied separately to each body part in a given frame. For fine grained action recognition PCNN represents person object interaction by joint tracking of hands and objects or by linking object proposal followed by feature pooling in selected regions. PCNN uses position of body joints to define informative image region. it represents body region with motion-based CNN descriptor. Where a descriptor is extracted at each frame and then aggregated over time to form video descriptor. To construct PCNN features we first compute optical flow for consecutive pair of frames, this method has relatively high speed and good accuracy.[2]

H. VGG Net:

VGGNet was introduced in the year 2014 and was developed by Simonyan and Zisserman. VGGNet consists of a very uniform architecture with 16 convolutional layers. It has 3x3 convolutional layers stacked on top of each other in increasing depth, but many filters and is the most preferred method for feature extraction of images. Reducing volume size is handled by max pooling. It has two fully connected layers, each with 4,096 nodes and then followed by a softmax classifier. VGGNet consists of 138 million parameters which makes it difficult to handle. There are two main drawbacks with VGGNet: first it is slow to train and second the network architecture weights in terms of disk/bandwidth are quite large.[1]

I. GoogleNet/Inception:

GoogleNet also called as Inception V1 was developed by Google during the competition ISLRC 2014 in which they conceded. It achieved a top 5 error rate of 6.67% which is a score very close to human level performance. The inception module is specifically a combination of many convolutions in a single image. GoogleNet is a 22 layer model, spreading both linearly as well as laterally. GoogleNet uses image distortions, batch normalization and RMSprop. GoogleNet is faster than the generic VGG Net. The pre-trained GoogleNet is also more cost effective than VGG Net.[1][4]

J. YOLO (You Only Look Once):

YOLO took a completely different approach to object identification and classification that out-performed R-CNN and all its variants. YOLO is not a traditional that is repurposed for object detection like R-CNN, it instead looks at the image frame only once (hence the name) in an intelligent way. YOLO divides the image into a 13x13 grid of cells. Each of these cells is responsible for predicting 5 bounding boxes. The purpose of a bounding box is to detect the object bounded within the rectangle. This means that is not only detect the presence or number of bounding boxes, but also identifies the class the bounding box belongs to, hence classifying the bounding box. YOLO also outputs a score that tells us how certain it is that the predicted bounding box actually encloses some object. In simple terms, higher the score of the bounding box produces a thicker bounding box. YOLO was trained on the PASCAL VOC dataset which can detect over 20 different object classes such as bicycle, dog, car, etc. The confidence score for the bounding box and the class prediction are combined into one final score that tells us the probability that this bounding box contains a specific type of object. Since there are 13x13=169 grid cells and each cell predicts 5 bounding boxes, there are a total of 845 bounding boxes. The main advantage of using YOLO is that even though there are 845 separate predictions, they were all made at the same time – the CNN ran only once. This is the reason for YOLO being so fast and its wide application in real time object detection.[6]

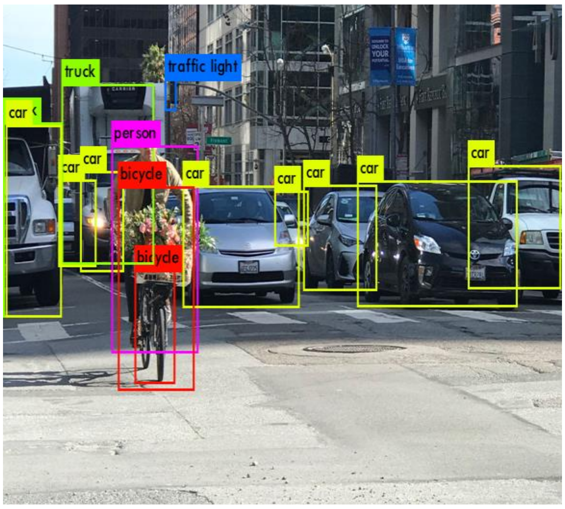


Fig.3. object detection by YOLO

IV. Comparison of Various Object Detection Methods and Most Commonly Used Datasets

This comparison was obtained by applying the most commonly used algorithms described above to the PASCAL VOC [10] dataset. The dataset contains 20 labels having minimum 1000 images which are in jpeg format.[1]

Table 1. Comparison of Object Detection methods

Algorithm	Error Rate
AlexNet	20%
GoogLeNet	15.5%
R-CNN	12%
VGG Net	18.2%
YOLO	6.7%

Conclusion drawn from the table:

From the above table, it is evident that RCNN is more optimized,but at a very basic level with a considerably small error rate of 12%. However, the most efficient of all the object detection methods is YOLO with an error rate of 6.7%. The results whose results are further explained in detail in the preceding conclusions. Evidently, AlexNet has the highest error rate of 20% and is less frequently when compared to the modern object detection algorithms for this very reason.

Accuracy of YOLO Algorithm in Object Detection

PASCAL VOC2007 Dataset

Algorithm	**YOLO**	**YOLOv2**
VOC2007 Accuracy	**63.4**	**78.6**

YOLOv2 is better than YOLO because it has batch normalization, it has a hi-res classifier, it is convolutional, it has dimension priors, it has location prediction, it has pass-through abilities, multi-scale abilities and hi-res detector.[8]

V. Benchmark Datasets for Human Action Recognition:

The most common datasets that are used to test the accuracy and/or error rate of any new or existing object detection models are a selected set of openly available datasets that are used by any machine learning enthusiast.

Image datasets and their usage are as follows:

1.MNIST:

This is the most common deep learning dataset consisting hand written digits (training set of 60,000 examples and test set of 10,000 examples). It is best used for deep recognition patters on real-world data and requires minimum data preprocessing.

2.MS-COCO:

Is a large-scale dataset with several features such as object segmentation, captioning dataset, recognition in context and 330K images.

3.ImageNet:

Is a dataset that is organized into a hierarchy based on the WorldNet hierarchy or images. ImageNet has approximately 1000 images to illustrate each of the 100,000 phrases present in WorldNet.

4.SVHN:

Is a dataset containing real-world images for developing object detection algorithms which require very little preprocessing. The dataset is comprised of the images of house numbers collected by Google Street View.

5.CIFAR-10:

This dataset is very important in image classification and consists of 60,000 images of 10 classes. It is also applied in machine learning and deep learning research.[9]

VI. CONCLUSION

There have been a number of ways which people have used to detect objects in images over the years, out of which the prominent and efficient methods are given below with their advantages and disadvantages.

Method	Application Domain	Advantages
CNN	Image recognition, video analysis, natural language processing, drug discovery, health risk assessment and biomarkers of aging discovery, checkers game	Sparse representation, parameter sharing, equivariance, very good feature extractor
YOLO	object detection (object classification, object localization)	Fast, accurate
VGG Net	Image classification with feature extraction	Good architecture, Simple to explain, pre-trained

		networks available for VGGNet
AlexNet	Non-linear feature extraction, larger scale visual image recognition	Simple, powerful network architecture, uses several GPUs
fast R-CNN	Object detection, object instance segmentation	High detection quality, no disk storage required for feature caching, training is single stage
GoogleNet	Image classification with feature extraction	Faster than VGG, size of pre-trained GoogleNet is comparatively smaller than VGG

REFERENCES

[1] Malay Shah, Prof. Rupal Kapdi "Object Detection Using Deep Neural Networks",ICICCS 2017.

[2] Guilhem Cheron,Ivan Laptev, Cordelia Schmid "P-CNN: Pose-based CNN Features for Action Recognition" , CVF

[3] M.A. Rashidan, Y.M Mustafah, S.B.A Hamid, N.A. Zainuddin, N.N.A Aziz "Detection of Different Classes Moving Object in Public Surveillance using Artificial Neural Network(ANN)", ICCCE 2014

[4] https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5

[5] Paul Viola, Michael Jones"Rapid Object Detection using a Boosted Cascade of SimpleFeatures". Accepted conference on computer vision and pattern recognition 2001

[6] https://arxiv.org/pdf/1506.02640.pdf

[7] https://cdnimages1.medium.com/max/800/1*QOGcvHbrDZiCqTG6THIQ_w.png

[8] https://medium.com/@jonathan_hui/real-time-object-detection-with-yolo-yolov2-28b1b93e2088

[9] https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets

[10] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. Cordelia Schmidand Stefano Soatto and Carlo Tomasi. International Conference on Computer VisionPattern Recognition (CVPR '05), Jun 2005, San Diego, United States. IEEE ComputerSociety, 1, pp.886–893, 2005,.