Comparative analysis of Human Activity Recognition and object detection

Sunil Suthar M., Nithya Darisini P.S*
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
Email:* psnithyadarisini@vit.ac.in

Abstract. Human Activity Recognition (HAR) is the ability to interpret human body gesture or motion and determine human activity or action. Most of the tasks can be done automatically if they can be recognized via HAR system. HAR is an essential part of various scientific research contexts and the daily defence system i.e. surveillance, healthcare and human computer interaction (HCI). In computer vision, images and videos are analysed using different techniques. While recognition of activity in the real time environment the speed of the technique matters because if the system can detect the object in the images and videos at real time then it can be forwarded for the activity recognition process. How correctly the technique detects the object and differentiates the human object and other object and how much time it is taking to give the correct object, the number of frames it processes per second all matters. This paper deals with the comparison of the performance of YOLO (You Only Look Once) and Faster R-CNN in recognizing the human activity and detecting object.

1. Introduction
In Computer vision-based technology, identifying an action or activity done by humans is most trending research area. There are different techniques approached by many researchers to recognise different types of human activities. But recognising an action correctly is a difficult task. To recognise activities perfectly there should be good quality of videos or images recorded by 2-D and 3-D cameras. Activities are of two types: Activity done by an individual and activity done by group of people. Identifying an individual’s activity is simple task and many researchers have worked on recognising individual’s activity. But if there are multiple persons in the image or video then it is difficult to recognise the activity. To understand an activity computer vision technique tracks the object in whole activity.

Image classification and Object detection plays an important role in HAR system. Earlier the CNN was the famous technique in differentiating the object in images [5]. But using it in real time with LIRIS dataset [6] is difficult because there may be different spatial location within the ROI of the object and the output layer size is not constant.

The rest of the paper comprises of the below sections. Section II describes the related Work for the study. Section III describes the different techniques and algorithms of object detection. Section IV about the dataset used. Section V describes the dataset preparation and conversion for different techniques. Section VI, the explanation of activity recognition using YOLO. Section VII describes training process of different techniques. Section VIII shows the comparative analysis and experimental results of used techniques.
2. Related Work
Image classification is needed to correctly classify the activity among different activities. [4] had described all classifiers like SVM uses the support vector for classification, CNN uses the sliding window technique on entire image and extract features, BOW (bag of words), KNN(k-nearest neighbour) and performed the experiments and evaluated the accuracy of different classification method. [1] described the process and uses the centre location of ROI for tracking distance of object in different frames and uses the aspect of ration technique of height and width for activity prediction. [3] proposed method to calculate binary motion of object in all frame and uses a CNN classifier. [2] discussed different way to represent image and ROI in the image, global and local representation, feature extraction and action classification. Faster R-CNN for object detection [7] describes the region base localization of object. [10] the skeleton bases human activity and hand gesture recognition by training in two stages first by CNN and then combination of CNN+LSTM which focuses on the spatial patterns related to the position of skeleton joints and track the joints of the skeleton.

3. Object Detection Methods
In human activity recognition, detecting object and localization of human object in image or video is important. Several object detection techniques are proposed by many researchers, such as:
1 Histogram of oriented gradients (HOG), which describes the features and finds the object by ignoring background and search each pixel in the image by creating feature vector and calculates the gradients.
2 Neural Networks R-CNN, which creates lots of region boxes in images and uses selective search method to search the object in the input image.
3 Faster-R-CNN Object detection instead of feeding all the region to the convolution network, it gives input image to the network and use separate network to predict the region proposal.

4) YOLO Object Detection
YOLO (You Only Look Once) is algorithm in which single convolution network is used for object detection as well as for the classification. This single convolution network predicts bounding box around object and class probabilities of object recognition.
Whenever an image is given as input to the YOLO it divides it into SxS Grid as shown in the figure - 2.

YOLO consider each grid as bounding box and predict the probability for each grid. The higher probability for availability of object. And each grid also contains offset value according to that the bounding box is created after probability of each grid. YOLO Processes 45-48 frame per seconds and also it works for detection and classification and also for localization. The good thing about this is it search entire image for the detecting object where as all other detection algorithm are selective search they find only from the proposed regions of boxes.

4. Dataset

The LIRIS dataset contains the video folder and each folder contains numbered frames, dataset also contains the XML format annotation file. For each video there is one annotation file as in the figure-3 format.

```
<?xml version="1.0" encoding="UTF-8"?>
<dataset>
  <video>
    <videoName>d2/vid0003</videoName>
    <action nr="1" class="6">
      <bbox x="175" y="259" width="57" height="177" framernr="115"/>
      <bbox x="175" y="259" width="57" height="177" framernr="116"/>
    </action>
    <action nr="2" class="3">
      <bbox x="354" y="176" width="122" height="391" framernr="143"/>
      <bbox x="354" y="175" width="122" height="391" framernr="144"/>
    </action>
  </video>
</dataset>
```

Figure 3. XML annotation.

The annotation file contains the action class number for the different group of frames, and box markup tag has the co-ordinates values for detecting the object in the specified frame number in the video.

Class – (class number for corresponding action identity)
X - (x1 co-ordinate of bounding box)
y - (y1 co-ordinate of bounding box)
width - ( width of the bounding box)
height - ( height of the bounding box)

5. Dataset Preparation & Implementation
For YOLO, the annotation should be converted in the following format for each video image and its bounding box and activity:

```
class_number box1_x1_ratio box1_y1_ratio box1_width_ratio box1_height_ratio.
```

For Faster-R-CNN, the dataset should be splitted in the training and testing. The annotation is converted into .csv file and it has the following columns.

```
filename, width, height, class, xmin, ymin, xmax, ymax
```

All the annotation in single file. Again the tensor flow .record file generated from CSV file.

6. **Activity Recognition using YOLO**

You Only Look Once (YOLO) algorithm is used with LIRIS dataset to prepare a model from training.

YOLO is much Faster than all other algorithm in detecting object from the frame. In YOLO convolutional network is used for predicting bounding box and the recognising the class of activity in the frame as process shown in the figure-4.

![Figure 4. Process.](image)

It detects the object from different grids and creates the bounding box by offset value on it and matches the object pose with the trained model and finds the activity with the confidence of activity class. From the identified activity and confidence which is greater than the threshold value is identified as activity.

7. **Training**

For YOLO training we have to create required files Listed below.

- Total number of actions.
- File of all action name.
- File with all path of video frame which will be used for training.
- File for configuration of layer.
- Already trained Weight file.
Steps for training model (Figure 5) is as follows:

- XML label file and coordinates of the video frames.
- Conversion of XML annotation to individual text file for each frame with the action class number and ratio of bounding box.
- Used pre-trained darknet19_448.conv.23.
- Continuous 7 hour training process.

For human activity recognition system, 10 normal day-to-day actions are used, numbered from 0-9 in the Table 1.

### Table 1. Activity Table

| Class Number | Class Name                                                                 |
|--------------|-----------------------------------------------------------------------------|
| 0            | Discussion_between_two_or_more_people                                      |
| 1            | Give_an_object_to_another_person                                           |
| 2            | Put/take_an_object_into/from_a_box/desk                                   |
| 3            | Enter/leave_a_room_(pass_through_a_door)                                  |
| 4            | Try_to_enter_a_room_(unsuccessfully)                                       |
| 5            | Unlock_and_enter_(or_leave)_(a_room)                                      |
| 6            | Luggage_left_unattended                                                    |
| 7            | Handshaking                                                                |
| 8            | Keyboard typing                                                           |
| 9            | Mobile/_Telephone_conversation                                              |

Steps for training Faster R-CNN model (Figure 6) is as follows:

- Conversion from XML annotation to CSV one file for training annotation.
- Generate the Tensor Flow record.
- Configure the number of classes.
- Path to CSV file and tensor flow records.
- Set the fine tuning for model.
8. Performance Analysis of YOLO and Faster R-CNN

|                               | Accuracy (object detection) | Training Time (per 1000 iteration) | Testing object detection time | Average Loss |
|-------------------------------|-----------------------------|-----------------------------------|-------------------------------|--------------|
| YOLO                          | 100%                        | 1 hour                            | 8 ms (milliseconds)           | 0.001000     |
| Faster R-CNN                  | 80%                         | 2.5 hours                         | 6.7 (seconds)                 | 0.2220       |

8.1. Yolo

Figure 7. Testing results YOLO.

The Accuracy depends on the number of training iterations. Accuracy increases as the training iteration increases.
It is suitable for real time processing and action recognition, because it takes very less time for testing.

8.2 Faster-R-CNN object detection:

Figure 8. Testing results Faster-R-CNN.

It is detecting object properly when image clarity is good and object’s size is enough to detect easily.

9. Conclusion

For human activity recognition using LIRIS dataset, the accuracy of YOLO (you only look once) technique is about 97% and for object detection and localization it is giving 100% accuracy whereas Faster R-CNN is detecting object properly based on the object visibility in the image and it is giving accuracy about 80% and taking more processing time as comparison shown in the table-2. To get better accuracy we need to train the model by accurate data and more data for individual activity is needed so that the system can recognise the same activity in different pose and situation.

References
[1] N. Albukhary and Y.M. Mustafah 2017 Real-time Human Activity Recognition 2017 IOP Conf. Ser.: Mater. Sci. Eng. 260 012017.
[2] Ronald Poppe Human Media Interaction Group 2010 A survey on vision-based human action recognition Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente. 6..
[3] Tushar Dhoba,l,ivswan shitole,Gabriel Thomas Girisha Navada 2015 Human Activity Recognition using Binary Motion Image and Deep Learning. December.
[4] Aisté Štulienė and Agnė Paulauskaitė-Tarasevičienė 2017 Research on human activity recognition based on image classification method Faculty of Informatics Kaunas University of Technology. Comput.Sci. 1856. pp.8-13
[5] Shubham Shinde, Ashwin Kotharia, Vikram Gupta 2018 YOLO based Human Action Recognition and Localization Procedia Computer Science 133,831-838
[6] The LIRIS Human Activity dataset https://projet.liris.cnrs.fr/voir/activities-dataset/download.html
[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 06
[8] https://colab.research.google.com/
[9] Juan C. Núñez, Raúl Cabido, JuanJ. Pantrigo, AntonioS. Montemayor, JoséF. Vélez 2018 convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition.