DETECTION OF HUMAN ACTIONS IN LIBRARY USING YOLO V3

BY

MD SHAJJAD HOWLADER
ID: 151-15-4830

REJWANA KARIM RETU
ID: 151-15-5193

AND

MUHAMMAD MAHBUBUR RAHMAN
ID: 151-15-4761

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

December 2018

APPROVAL

This Project titled **"Detection of Human Actions in Library Using YOLOv3**", submitted by MD Shajjad Howlader, ID 151-15-4830, Rejwana Karim Retu, ID 151-15-5193 and Muhammad Mahbubur Rahman, ID 151-15-4761 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on 11 December, 2018.

BOARD OF EXAMINERS

_____

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Prof. Dr. Syed Akhter Hossain, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Dr. Syed Akhter Hossain
Prof. and Head
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

_____
Name: MD Shajjad Howlader
ID: 151-15-4830
Department of CSE
Daffodil International University

_____
Name: Rejwana Karim Retu
ID: 151-15-5193
Department of CSE
Daffodil International University

_____
Name: Muhammad Mahbubur Rahman
ID: 151-15-4761
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty Allah the most merciful, for His divine blessing makes us possible to complete the final year research successfully.

We really grateful and wish our profound our indebtedness to **respectable Supervisor Dr. Syed Akhter Hossain, Professor and Department Head, Department of CSE, Dhaka**. Deep Knowledge & keen interest of our supervisor in the field of "Computer Vision" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We are also thankful to **Dr. Md. Milan Khan, Librarian of DIU library**, and library staffs for giving us permission and associate with us in data collection activity at DIU library.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

We also like to thank all volunteer student who associated with us in data collection procedure at DIU library.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

A person's activity in a library should be monitored to avoid any unwanted problems. In this project, we have investigated a problem of image-based human action detection in a library. It involves making a prediction by analyzing human poses, behavior, and actions with objects from complex images instead of video. Comparing with all approaches, we conclusively decided to use an algorithm YOLOv3 (You Only Look Once) which is latest and more convenient. The algorithm utilizes anchor boxes, bounding boxes and a variant of Darknet. We have created our own dataset collecting images from library and annotated the dataset manually. During the research with this project, we have considered human activities in a library into five section namely studying, phoning, using a computer, taking book and sleeping. The proposed system provides not only multi-tasking knowledge with classification but also localization of human and the equivalent actions instantaneously. Interestingly, the proposed approach achieved a mean average precision (mAP) of 96.3%. In the future, incorporation of real time data analysis will add value to this project.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Introduction

Computer vision and pattern recognition is a vast area in recent research field. From the sector of computer vision we attain knowledge from digital images and videos. In this arena, a vital topic is human action detection. Whatever people do is known as human action. Activity recognition is an essential expertise in extensive computing as it can be applied to many real-life human-centric problems. In the library, students are not only confined to study but also many other actions. Sometimes, they create an unavoidable and noisy atmosphere which is irritating for others as well as authorities. Human behavior and activity in the library often demand to be monitored for preservation reasons and other purposes.

In our project, we focused on their various activities and detect these actions through images instead of videos. We have considered few activities and categorized those into five sections like studying, phoning, taking books, sleeping, using computers. We have tried to outline a solution for recognition of human behavior in the library using YOLO (You Only Look Once) approach for action detection in still image based. Humans glimpse at an image and know instantly what activities are done by the people in the library. The system YOLO trains on full pictures and adjusts action detection directly. We have to use third and latest version of YOLO that is YOLOv3. It's a little conspicuous than the last versions but more accurate and also fast. YOLOv3 calculates an object's grade for each bounding box which uses logistic regression. Each of the box predicts the classes and for the class projections, we have used binary cross-entropy loss while training. For outstanding performance, we used independent logistic classifiers instead of softmax because it is worthless.

For summarization, the main contributions can be represented as:

- Explained how an effective object detection algorithm can be applied in action detection.
- Created a dataset called ‘DIU-LibAct’ containing 600 image data.
- Utilized image data which is computationally less expensive and achieved a satisfactory result.

1.2 Motivation

After spending a whole day in library suddenly I asked my friend, what we have done today or how we spend our time today in library. We realized most of the time we were using the computer. What about our others activities in library like studying, sleeping, phoning, using computer? Is there any way by which we can observe our activities in the library? If there is a daily life action detector, one can easily get the statistics of his/her daily life activity. It will be the same for action detection in library or office or other place. Let's build a library action detector using deep learning. We need a lot of images and there is one way out, we can use web images or real-life library action images for this purpose.

After studying this topic, we found that most of previous works have been done motion/video based datasets. Very few numbers of work have been done based on image dataset. So, we got inspired to work with image. Understanding motion from still images is not an easy task. With the presence of motion, it is far easy to detect action. We have to evaluate the place and pose of the person in still images. Human pose and configuration of body parts are our main concern. Specially we emphasis on hand movement as our detection place is in library. It seems very interesting that an intelligent system is capable to specify that what action is performing a human in library only from capturing image.

1.3 Rational of the study

Human understands things better and fast by visualization. So, to compete with the modern world, computers are also being expert to realize objects from images. Computer vision is concerned with the notion and knowledge for erecting artificial systems that gain information from images.

Our study provides an outline to computer vision including fundamentals of image realization, feature detection and matching, and also classification. Though a lot of works have been introduced in this sector, we focused on action detection which is the recent topmost. What actions human are performing is identifying only from an image instead of a motion-based video. In our daily life, many unethical, unsocial activities are occurring everywhere. Various strict and innovative systems are also being developed for halting them. So, it will be very progressive, if there exists an artificial system where human activities are being recognized automatically from images. That's why we choose this topic.

As a student, we started to implement it from our university library. Here the system will detect human action at library environment like studying, phoning, talking, sleeping, using computer etc. From these the library authority will get a clear conception about the performance of students, stuffs. It is not so far that the human activity detection technique will be surely effective for developing a disciplined and self-controlled world.

1.4 Research Questions

While creating this research paper, we faced some related problems. A lot of questions came to our mind which made us curious that how we will come to an end of these. These questions are the following:

- How will we get data for research?

- How have we processed all these data?
- Which algorithm will be appropriate in deep learning for detect human activities?
- How can we improve accuracy?

1.5 Expected Outcome

From the study in this paper, we expect to achieve efficient solution of:

- Finding the way of using object detection algorithm properly in action detection.
- Detect particular action and multi-action in single image.
- Dataset of common action of general student at library.
- Accuracy of action detection over 90%.
- Can be used in video surveillance in library.
- We have successfully achieved all our expectations. Finally, it has given output result that are shown the Table 4.3.

1.6 Report Layout

The following report consists of five chapters. Chapter 1 is for introduction, which focuses the motivation and goals behind the research. It has six sections. These are:

- Section 1.1(introduction), 1.2(motivation), 1.3(rationale of the study), 1.4(research questions), 1.5(expected output) and 1.6(report layout).

- Chapter 2, titled background, prerequisites information relevant to the research and is distributed into five sections. Section 2.1, 2.2, 2.3, 2.4 and 2.5, identified introduction, related work, research summary, scope of the problem and challenges correspondingly.

- Chapter 3 demonstrates the details of our research experiment spanned over five subsections. Section 3.1, 3.2, 3.3, 3.4 and 3.5 explains each phase of the recognition, i.e. Introduction, research subject and instrumentation, data collection procedure, statistical analysis and implementation requirements.

- In chapter 4 we have declared the experimental results and discussion of that result. It has four sections. Section 4.1, 4.2, 4.3 and 4.4 titled introduction, experimental results, descriptive analysis and summary.

- Finally, chapter 5 have also four sections that titled summary, conclusion, recommendation and implication for future research.

- An appendix, which lists all signs we have used for our experiment, follows the five sections and at the end there is a reference of reading materials we have referred to during our research.

CHAPTER 2

BACKGROUND

2.1 Introduction

In this section, we will review the research work done by researchers in the field of Action recognition and detection in both video and still images.

Human action recognition intends at recognizing human actions in videos or still images, which is an ongoing research topic in computer vision and has a broad range of applications, such as surveillance, action-based image retrieval, Human-computer interaction(HCI). Although a lot of attempts performed in the earlier decades still action recognition resides a pretty challenging job, where the complications occur due to the human pose varieties, cluttered backgrounds, occlusions, and illumination changes. Such challenges are increased for still images as there exist no motion cues, which play a very productive role in revealing human actions in videos.

Nowadays, people used to monitor a whole organization and institution through CCTV. For this purpose, so many man powers are required for proper monitoring. Still monitoring and detecting any unwanted situation as well as taking real-time action is almost impossible for a human being. So we need an automatic surveillance system which can detect human action within a short moment.

A space like a library should also be properly observed in order to restrict using the mobile phone, making noise etc. This is a very unique area of action recognition and detection.

2.2 Related Works

Human action recognition and localization is an ongoing research topic in computer vision. A lot of approaches have been provided from last two decades. But surprisingly most of

them on video or sequential images as well as at the initial stage those were only for action recognition. At first, we will discuss those methods.

2.2.1 Action recognition and detection in video

M. Baccouche [1] applied a 3D CNN to learn spatio-temporal features from video followed by employed an LSTM to classify video sequences. A. Karpathy [2] introduced various ways to fuse temporal information from sequential frames using 2D pre-trained convolutions. Simonyan, Karen & Zisserman [3] presented a clarification to the toughness of deep architectures to acquire motion features. They explicitly created motion features in the form of accumulated optical flow vectors. That's why instead of using a single network for spatial context, this methodology has two separate networks - one for spatial context (pre-trained) and another one for motion context. As temporal dynamics of body parts produces powerful information across time on the exhibiting action. The researcher employed this information for action recognition and localization in numerous works. G. Cheron [4] applied pose information to obtain high-level features from optical flow and appearance. They demonstrate that utilizing pose information for video classification is extremely efficient. C. Wang [5] adopted data mining procedures to achieve a representation for all video followed by applying a bag-of-words model to classify videos. Papadopoulos G.T. [6] which is based on skeleton tracking, in this approach, a video stream is passed into a skeleton-tracking algorithm. Then based on the transition among selected joints and their corresponding angular velocities, action recognition can be performed. Y. Ke [7] introduce a method to event detection in congested videos. Later on, Y. Tian [8] explain a Spatio-temporal Deformable Parts Model to detect actions in videos. M. Jain [9] and K. Soomro [10] apply supervoxel and selective search approach to localize the action boundaries. Kalogeiton [11] use 2D CNN to obtain frame-level features and build action proposals by using the anchor cuboids. Those cuboids are later classified and refined by regression. Hou [12] deploy anchor boxes to form tube proposals, which are joined together and classified. They named it as Tube Convolutional Neural Network (T-CNN). Kevin [13]

proposed a unified network for action detection which can simultaneously produce pixel-wise action segmentation as well as action classification. Gu, Chunhui [14] elaborate the idea of the I3D network for action localization where a region proposal network selects spatio-temporal regions that needed to be classified and refined.

2.2.2 Action recognition and detection in still images

We can broadly classify the existing methods into three categories.

1. Pose-based methods: S. Maji [15] apply part detectors to identify and encode the parts of human bodies to build poselet for action recognition. J. Tompson [16] proposed a CNN design using human pose evaluation for efficient object localization.

2. Context-based methods: B. Yao [17] grant three things, human-object interaction, multi-interactions in an action with human poses and the association between objects. V. Delaitre [18] produce and use discriminative interacted person-object couples for action recognition. G. Gkioxari [19] detect the maximum relevant object to the person in an action by applying learned object detectors.

3. Part-based methods: G. Sharma [20] recommend to use image local patches as parts and learn a DPM like classifier for action recognition. Shugao Ma [21] use web images to train a CNN model for action recognition. A. Prest [22] recognized action images simply using image labels in all images.

And, there exist very few works of action detection in still images.

Fahad Khan [23] use action-specific person proposals to detect action in images. Shi-Yang Yan [24] use the CNN model to detect human action in the office scene.

2.3 Research Summary

Therefore as the early literature review and study exhibits that there has been a fair number of studies in this domain. Before the growth of deep learning, traditional works in this field

of action recognition concentrated on extracting features like densely or a sparse set of interest points from image to distinguish among action classes. These features were combined with high order encodings, e.g., bag of words in short (BoW) or Fisher vector based encodings, to generate a fixed-sized video level description. Then a classifier, like SVM or RF, is trained for action labels prediction. The recent study explained that most of these procedures are not only computationally expensive, but they are also failing in obtaining context and high-level information. After successful implementation of deep learning in image classification, researcher turns the hand-crafted features to learned features, and the learning process becomes end-to-end. CNN learns to utilize high-level information from large-scale video datasets by obtaining localized features as well as context cues. That's why among all the action classification methods the state of the art method involves the use of CNN. Action detection is a more challenging task compared to action recognition as approaches for action detection require models to not only classify actions but also localize them. There exist sound similarities and differences between the methodology of action recognition and detection in video and still images. The major difference due to still images doesn't contain temporal information. That's why it becomes so tough to recognize action in still images.

2.4 Scope of the problem

The librarian can easily keep eye on the activity of a user. Any kind of rules breaking activity can be identified instantly. Librarian can also monitor who are taking books and check whether they note down it before leaving with the book.

Although initially, we work only on still images, the same system can be implemented on video too. Adding more action type will add more variety. Further improvement like taking a snap of rules breaking activity will make it more useful.

2.5 Challenges

Building a system like this required lots of image. That's why the primary challenge is to collect image data. We collect image data with the association of different volunteers. Another challenge is to find out the proper size of the anchor boxes. We deploy a k-means clustering algorithm to face this issue. And, the big challenge is to figure out which techniques works better on this dataset.

CHAPTER 3

Research Methodology

3.1 Introduction

The computer system cannot understand the actions easily so that detection of actions is a challenging task. For this, every researcher tries a different kind of techniques to understand the actions with the computer system. The system we recommended is applied some procedures to detect actions with the computer system. Our proposed method is appropriate for the detection of actions in the library.

In this chapter, we will mainly discuss about the theoretical knowledge of the research work. It will help to understand the concept of work. For this, Research subject and Instrumentation is described shortly. Then we know the heart of machine learning or deep learning which is data. So the data collection and preprocessing are also represented. We closed this chapter by giving clear concept about implementation requirements.

3.2 Research Subject and Instrumentation

Our research topic is "Detection of Human Actions in Library Using YOLOv3 ". It is the field of image processing, Neural Network. We used deep learning algorithms.

Here we will discuss the key tools for our research. We use python for the implementation. Deep learning algorithm of YOLO v3 is used for data training. We also used an annotator software for data annotation. By K-Means clustering, we generate anchors.

3.3 Data Collection Procedure

There exists few benchmark dataset for human action recognition and localization known as detection. Some of those are Stanford 40-action, PASCAL VOC 2012, UCF101, Kinetics. As we are focusing on detecting action in the library we were looking forward to using some action class of those datasets like reading, phoning etc. But all images were from the different scenario. That's why we build a dataset called DIU-LibAct considering of 5 different actions to aid the study of detecting human actions in Library. To build this dataset total 35 volunteers were associated at DIU library. We took images of all of them in the different place of the library. They pose in some predefined actions like studying, phoning, taking books, sleeping, using computers. Figure 3.1 presents a few images of the dataset.

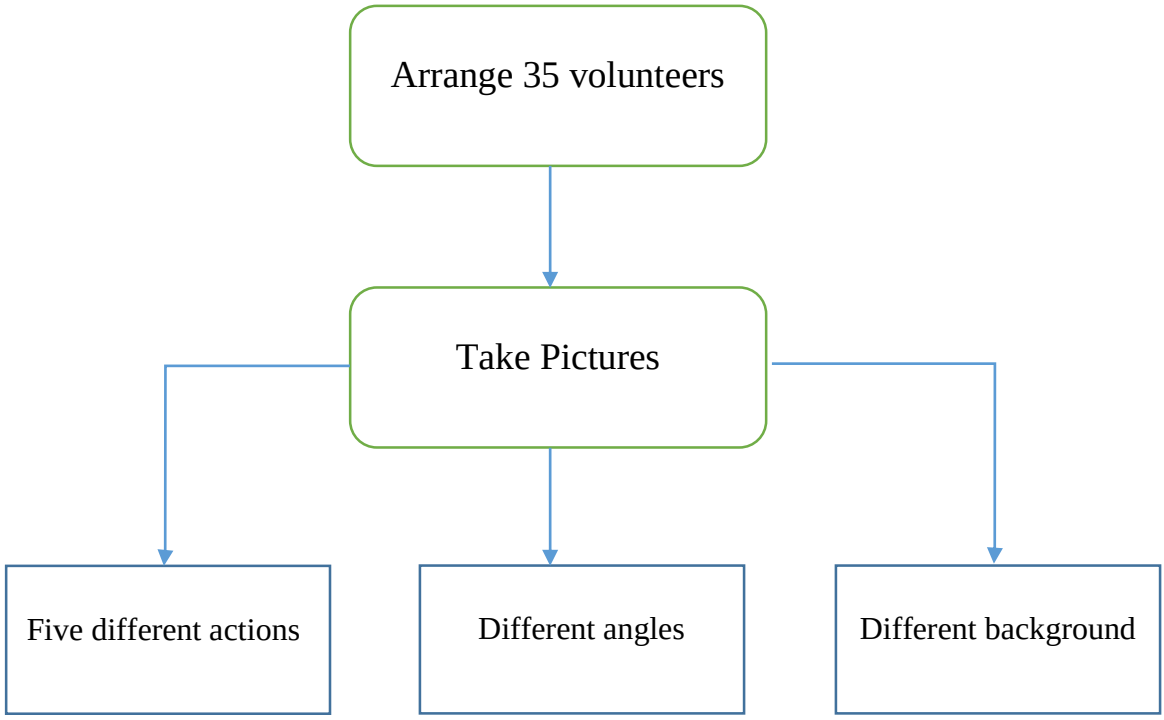


Figure 3.1: Data collection procedure

In the season of data acquisition, we consider a number of factors including the background complexity, crowded background, and the angle of view changes. We also consider

multiple actions in the same image, different distance, and illumination. Thus we collect multiple images of the same action class for a particular human. We gather a sum of 600 images to develop the dataset.

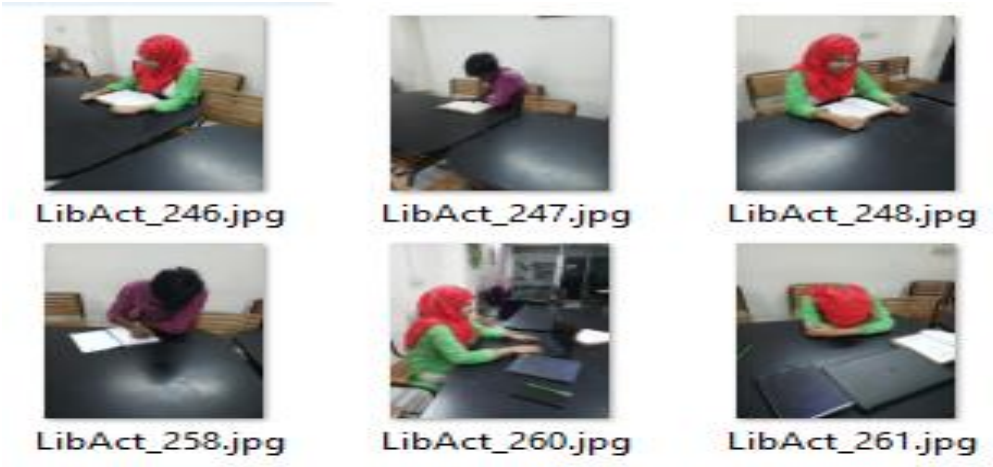


Figure 3.2: Sample Data

3.4 Data Pre Processing

To prepare the training data we mainly provide ground truth action class with a bounding box for each action in the image. For this purpose, we use label Image, which generates a corresponding XML file for each image. The XML file includes the information about the size of the image, its action class, the value of bounding box (xmin, ymin, xmax, ymax). As a singular image may have several actions so there will be more than one bounding box value for those particular images.

Figure 3.3: XML file of label image.

3.5 Methodology

Understanding motion from still images is not an easy task. With the presence of motion, it is far easy to detect action. We need to estimate the place and pose of the person in still images.

Whenever it attains to faster object detection algorithm, we all think about YOLO. Though it has some accuracy arguments. In our work, we adopted YOLO v3 which is better, more accurate, little slower than YOLO v2.

YOLO v3 manages the more complex architecture of Darknet which makes it slower but develops its accuracy. YOLO v3 furnishing us a 106 layer fully convolutional architecture. It applies a variant of Darknet. It makes the detection in three different scales which is the most conspicuous feature of v3. The input image dimensions are 32, 16 and 8 sequentially. YOLO v3 uses 9 anchor boxes. We used K-Means clustering to generate 9 anchors. YOLO v3 predicted more extended bounding boxes than YOLO v2. It can be performed multi-label classification for objects detected in images.

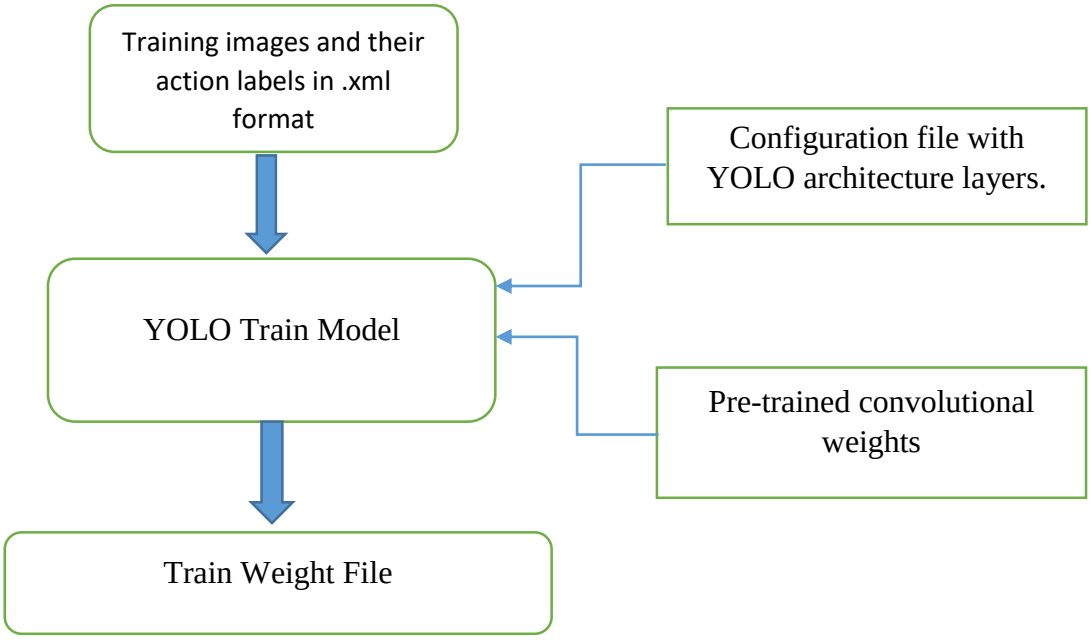


Figure 3.4: YOLO Training Diagram

First, we transformed the "DIU-LibAct" dataset in the form of YOLO supported format. Then we added some file to YOLO training model. In which we specify the number of actions and their names, mention the path where train weight file will be saved, mention the configuration file which contains all layers of YOLO algorithm, pre-trained convolutional weights.

3.5.1 YOLO Architecture:

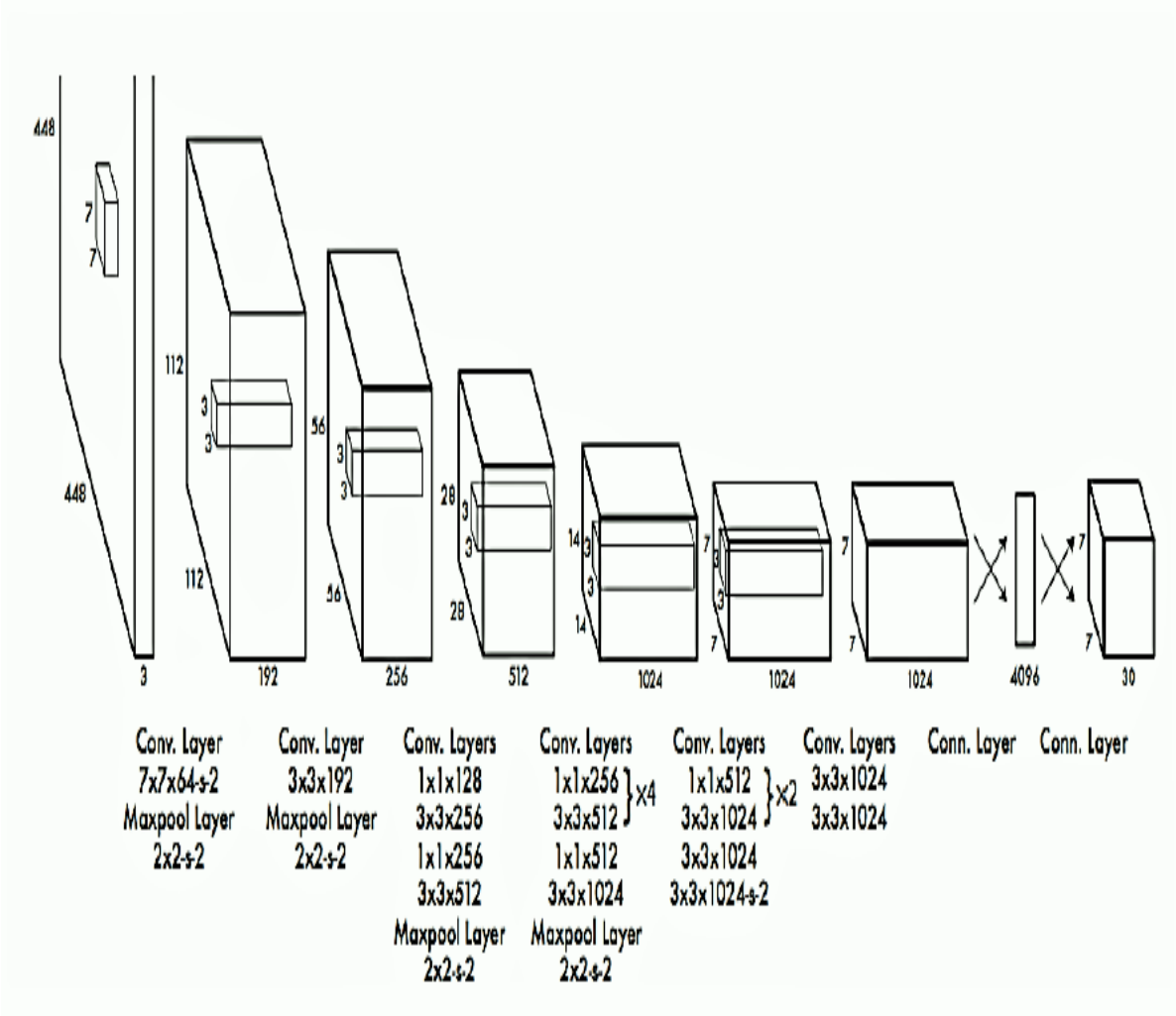


Figure 3.5: YOLO Architecture

This architecture we follow for our action detection task. Dividing each image into S x S regions and within each region, it directly sinks to find B bounding boxes and a score for each of the C classes. It is the key idea of YOLO. For each of the B bounding boxes, there are center x, center y, width, height and confidence of the bounding box. There will only be one set of class scores C for all bounding boxes in that region. The output of the YOLO network will be a vector of S× S × (5B + C) numbers for each image. YOLO was pre-

trained on ImageNet with S = 7, B = 2, and C = 20. In general, the existing YOLO architecture consists of 24 convolutional layers followed by 2 connected layers and a final output layer. Since there are only 5 classes of actions, our last layer requires C = 5.

3.5.2 YOLO Loss Function

The loss function can be divided into five sections, in which sections (i) and (ii) are focusing on the loss of the bounding box coordinates, sections (iii) and (iv) are scolding the differences in the confidence of having an object in the grid and section (v) is scolding for the difference in class probability. The loss function for the bounding box size is based on the square root of the dimensions which is an interesting part to note. The small deviations in longer bounding boxes should provoke less of a penalty than in miniature bounding boxes. The "lamda-coord" hyper-parameter is set to assure “fair” contribution of the bounding box location penalty and the classification penalty to the overall loss function. The "lamda-noobj" is set to scold less for the confidence of identifying an object when there is not one.

$$Loss = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \qquad (i)$$

$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \qquad (ii)$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (C_i - \hat{C}_i)^2 \qquad (iii)$$

$$+ \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} \, (C_i - \hat{C}_i)^2 \qquad (iv)$$

$$+ \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \epsilon classses} (p_i(c) - \hat{p}_i(c))^2 \qquad (v)$$

3.6 Implementation Requirements

So far we have discussed the theoretical concepts and methods. Now a list of requirements has been generated for "Detection of Human Actions in Library Using YOLOv3 ". Some of the probably necessary things are given below-

Hardware/Software Requirements

- Minimum Quad Core Processor
- 8 GB RAM
- GTX 1060 GPU
- Operating system
- Minimum 6 GB free space Hard Disk

Developing Tools

- Python Environment
- Anaconda Prompt
- Annotator Software

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

We do the experiments on the basis of our test data to find out whereby enormously we can detect accurately. To do our experiment, we trained our model at first so that it can absorb important features of our actions from training dataset.

After that, we need to do fine tuning to our model to get better accuracy regarding detection purpose. We get the overall efficiency of 83% after training of 124th iteration. This efficiency further depends on the amount of iteration we perform. In the next section, we will show the accuracy of the training set, the validation set, and the test set.

For the experimental results, we need to calculate the accuracy of the task of detection of human actions in the library. For this, we used the method of Average Precision and Mean Average Precision. Average Precision is used to calculate the accuracy of actions independently. Whereas Mean Average Precision is used to calculate the accuracy of actions in a combination.

4.2 Experimental Results

Most of the time we need to work with different angles of the same action. For this, feature extraction is quite difficult for the detector. Need to consider all the possible features. By the digitalization of the computer, there have been many efficient techniques to perform the detection task.

YOLOv3 is the best algorithm so far to supervise detection correlated difficulty more efficiently. We trained our model with the dataset of "DIU-LibAct" to detect five actions. The first action is "phoning", the second is "sleeping", the third is "studying", fourth is "taking the book" and the last is "using a computer". In our model, there are several stages to extract features from "DIU-LibAct" image dataset and runs many times to get the better result. In the meantime of training, fine-tuning is begin.

When our model parameters will get the "fine-tuning", we will able to detect actions more accurately. When we completed the 33rd iteration, our model stopped training because there is no updating in the last three iterations. We get overall 0.96 mAP which is called 96% accurate. We find out mAP for train, validation and test data. All of are near and over 0.96. The table of mAP values for train, validation and test data are given below.

Table 4.1: Training Set Accuracy (IoU 0.5)

ACTION	AVERAGE PRECISION	MEAN AVERAGE PRECISION
Phoning	0.9574	0.9699
Sleeping	0.9575	
Studying	0.9952	
Taking the book	0.9738	
Using a computer	0.9656	

Table 4.2: Validation Set Accuracy (IoU 0.5)

ACTION	AVERAGE PRECISION	MEAN AVERAGE PRECISION
Phoning	0.9703	0.9640
Sleeping	0.9579	
Studying	0.9946	
Taking the book	0.9500	
Using a computer	0.9473	

Table 4.3: Test Set Accuracy (IoU 0.5)

ACTION	AVERAGE PRECISION	MEAN AVERAGE PRECISION
Phoning	0.9670	0.9633
Sleeping	0.9579	
Studying	0.9946	
Taking the book	0.9500	
Using a computer	0.9473	

4.3 Descriptive Analysis

Before training, we divide our dataset into three parts. They are the train, validation, and test. Train data contains 80% of the "DIU-LibAct" whereas validation and test data contain 10% of the "DIU-LibAct" each. We evaluate our model by test data which is unique from the train and validation set. Our model is already introduced and our model accuracy is 96%.

In the table 4.3, it shows the test set accuracy with the IoU 0.5. Here we see studying has the highest average precision of 0.9946. It indicates that our detector can detect the studying action with the accuracy of 99%. Let us see a detection output of “studying”.

Figure 4.1: Detection of studying.

In the table 4.3, we see that taking the book has the lowest average precision of 0.9500. It indicates that our detector can detect the "taking the book" action with the accuracy of 95%. Let us see a detection output of "taking the book".

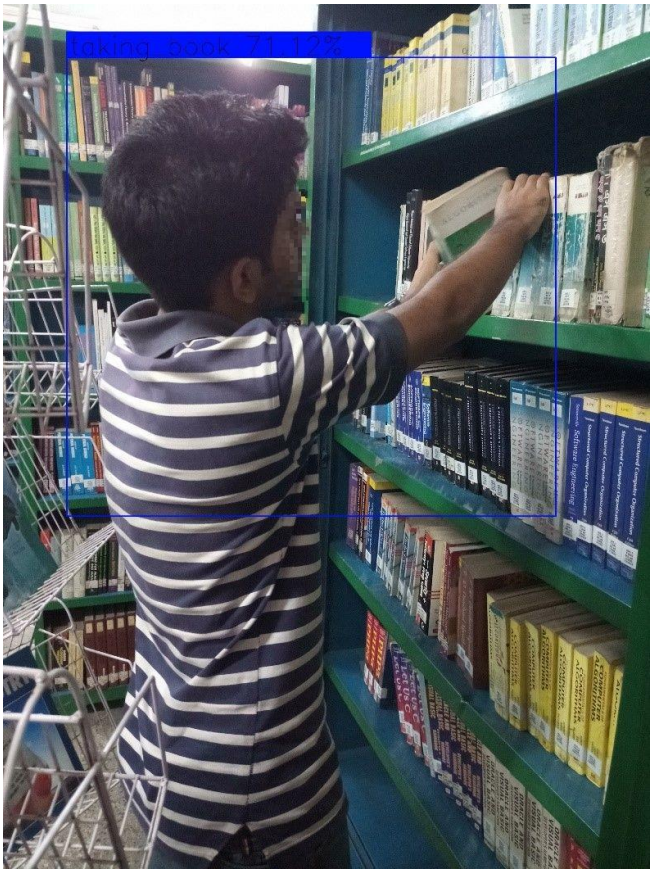


Figure 4.2: Detection of taking book.

We also see that phoning, sleeping and using a computer has the accuracy is 97%, 96%, and 95% respectively. For the test set. Our mean average precision is 0.9633. It means our model overall accuracy is 96%. Let us see some of the detection output of “phoning”, “sleeping” and “using a computer”. In figure 4.3 the person is phoning while studying.

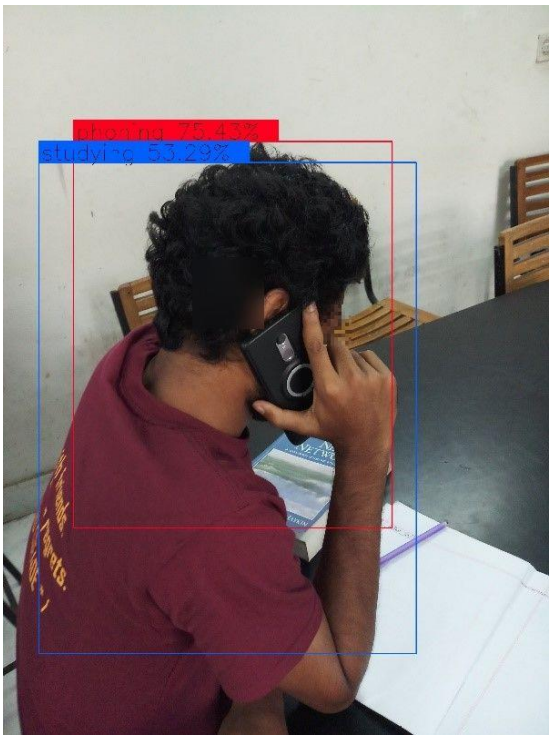


Figure 4.3: Detection of phoning.

Figure 4.4: Detection of sleeping.

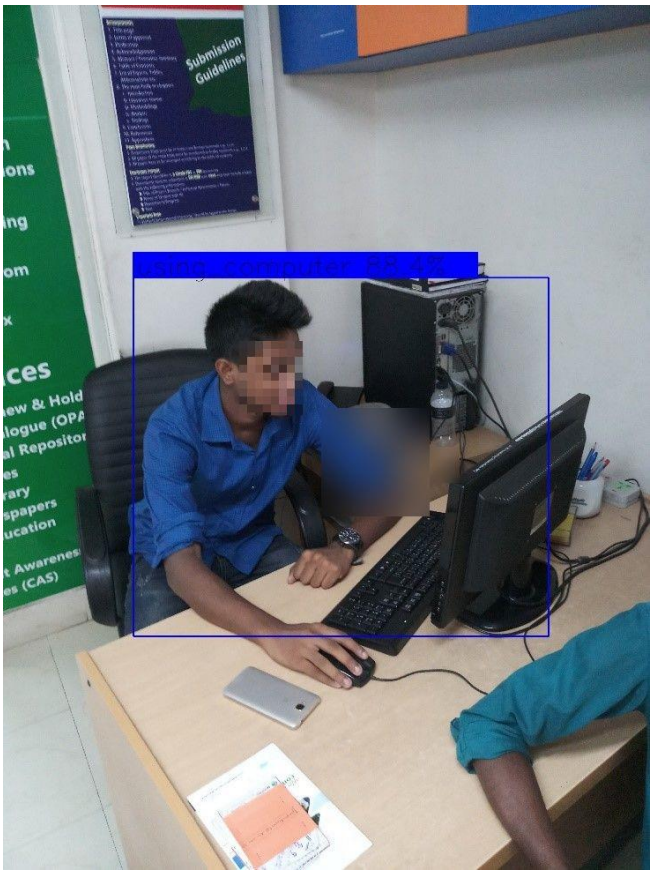


Figure 4.5: Detection of using computer.

As we got high accuracy on every class. Our trained model can detect all action even in a congested image where people are involved in multiple action. The image at figure no 4.6 shows that multiple person's action are detected so accurately even when an action from different class exist in the image.

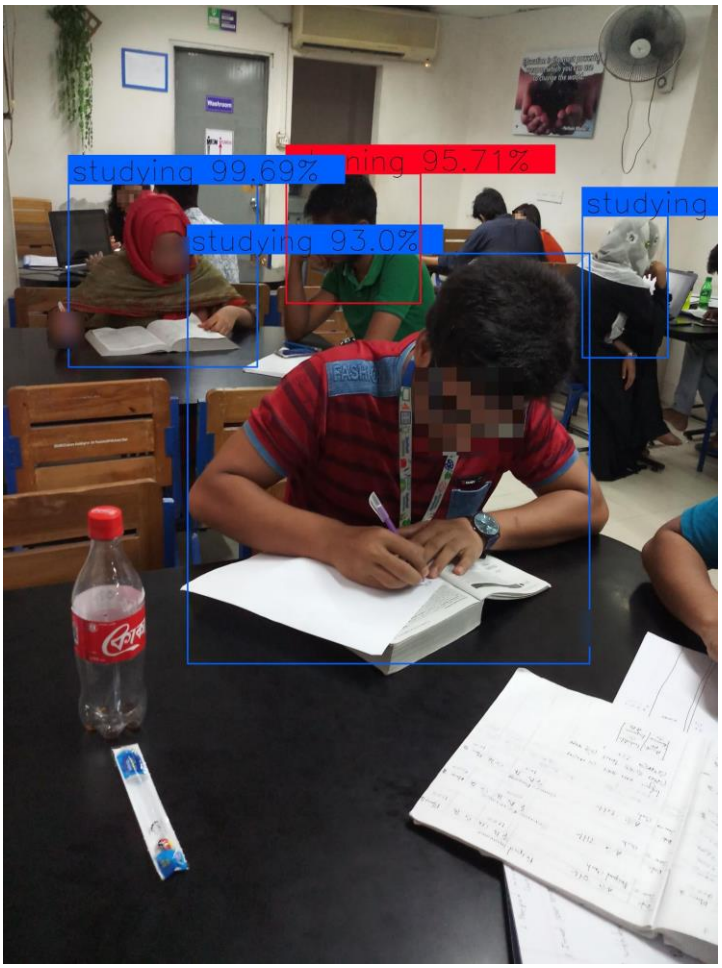


Figure 4.6: Detection of multiple action in a congested image.

4.4 Summary

Though it is tough to detect actions from still images, we get satisfactory results to detect the actions. We get good accuracy to detect the "studying" action. We get less accuracy to detect the "using a computer" action. Our overall mAP is 96% which is a good number. In this time, we cannot differentiate the reading and writing. Cause both are much similar action. Hope we will defeat it in near future.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

We are trying to represent a deep learning technique with YOLO v3 in library action detection problem. The whole research short summary is given below-

Stage I:

- Data collection.
- Data annotation.

Stage II:

- Divided data into train, validation and test set.
- Transfer learning.

Stage III:

- Detect the actions using YOLO v3.

Stage IV:

- Calculate the accuracy to get better result.

Need to do the iterations for fine-tuning the parameters to get better accuracy.

5.2 Conclusions

In this research project, we estimated a real-time approach for human activity detection, localization and image classification based on YOLOv3 from complex scenes. The method is validated with our challenging dataset where are many cluttered and noisy data for checking more accuracy. It can detect more than one person's different activities using more bounding boxes in a single image. Various action detection techniques and some

research topics that are linked to action analysis in still images have been discussed in our paper. Our high-level attainment amounts over the challenging still action set indicates that our method is quite capable of identifying the human actions. Our future work is to further spread the current system for more universal applications like now our work is limited at a library , later it will detect human action at any place. Our target will be also detection for more complex human actions and for more categorization.

5.3 Recommendations

It will be better to growth the amount of training data. The dataset needs data from more different angles and various source of lights. Try other different deep learning algorithms to compare which is better for "DIU-LibAct". This model can be further used for various action detection problems like home, industry action detection problem. A large dataset demand to build to perform various action detection task.

5.4: Implication for further studies

Each and every system has been formed with future advancement opportunities point. In future, this system will be faster and more efficient. Reducing processing time is one of the important issues.

We will be upgraded this for better performance from now. We want to continue the research in this field. We will try to detect the more complicated actions. We will work with videos. We will try to get more accuracy by applying various techniques.

We will make a integrate software by which we can make a report of peoples actions. The automatic alarming system will be developed for unwanted actions.

REFERENCES

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. "Sequential deep learning for human action recognition. In Proceedings of the Second International Conference on Human Behavior Unterstanding" HBU'11, pages 29–39, 2011.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. "Large-scale video classification with convolutional neural networks" In CVPR, 2014.

[3] Simonyan, Karen & Zisserman, Andrew. "Stream Convolutional Networks for Action Recognition in Videos" Advances in Neural Information Processing Systems, 2014.

[4] G. Cheron, I. Laptev, and C. Schmid. "P-CNN: Pose-based CNN Features for Action Recognition" In ICCV, 2015.

[5] C. Wang, Y. Wang, and A. L. Yuille. "An approach to posebased action recognition" In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, pages 915–922, Washington, DC, USA, 2013. IEEE Computer Society.

[6] Papadopoulos, Georgios & Axenopoulos, Apostolos & Daras, Petros "Real-Time Skeleton-Tracking-Based Human Action Recognition Using Kinect Data" 8325. 473-483. 10.1007/978-3-319-04114-8_40, 2014.

[7] Y. Ke, R. Sukthankar, and M. Hebert. "Event detection in crowded videos" In IEEE International Conference on Computer Vision (ICCV), pages 1–8, 2007.

[8] Y. Tian, R. Sukthankar, and M. Shah. "Spatiotemporal deformable part models for action detection" In IEEE Con- ference on Computer Vision and Pattern Recognition, pages 2642–2649, 2013.

[9] M. Jain, J. Van Gemert, H. Jegou, P. Bouthemy, and C. G. "Snoek. Action localization with tubelets from motion" In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 740–747, 2014.

[10] K. Soomro, H. Idrees, and M. Shah. “Action localization in videos through context walk” In IEEE International Conference on Computer Vision (CVPR), pages 3280–3288, 2015.

[11] Kalogeiton, V. Weinzaepfel, P. Ferrari, V. Schmid. “Action Tubelet Detector for SpatioTemporal Action Localization” ICCV (2017)

[12] Hou, Rui & Chen, Chen & Shah, Mubarak. (2017). “Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos” 5823-5832. 10.1109/ICCV.2017.620.

[13] Duarte, Kevin & S Rawat, Yogesh & Shah, Mubarak. (2018). “VideoCapsuleNet: A Simplified Network for Action Detection”

[14] Gu, Chunhui, Sun, Chen, Vijayanarasimhan, Sudheendra, Pantofaru, Caroline, Ross, David A, Toderici, George, Li, Yeqing, Ricco, Susanna, Sukthankar, Rahul, Schmid, Cordelia. “AVA: A video dataset of spatio-temporally localized atomic visual actions” CVPR, 2018.

[15] S. Maji, L. Bourdev, and J. Malik, “Action recognition from a distributed representation of pose and appearance,” in Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition, 2011, pp. 3177–3184.

[16] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, “Efficient object localization using convolutional networks,” in Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition, 2015, pp. 648–656.

[17] B. Yao and L. Fei-Fei, “Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1691–1703, 2012.

[18] V. Delaitre, J. Sivic, and I. Laptev, “Learning person-object interactions for action recognition in still images,” in Proc. Advances in Neural Information Processing Systems, 2011.

[19] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R*CNN,” in Proc. IEEE Int’l Conf. on Computer Vision, 2015, pp. 1080–1088.

[20] G. Sharma, F. Jurie, and C. Schmid, “Expanded parts model for semantic description of humans in still images,” arXiv:1509.04186, 2015.

[21] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, Stan Sclaroff “Do Less and Achieve More: Training CNNs for Action Recognition Utilizing Action Images from the Web”

[22] A. Prest, C. Schmid, and V. Ferrari, “Weakly supervised learning of interactions between humans and objects,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601–614, 2012.

[23] Khan, Fahad & Xu, Jiaolong & Weijer, Joost & D. Bagdanov, Andrew & Anwer, Rao & López, Antonio. (2015). “Recognizing Actions through Action-Specific Person Detection” IEEE transactions on image processing: a publication of the IEEE Signal Processing Society. 24. 10.1109/TIP.2015.2465147.

[24] Yan, Shiyang & An, Yu-Di & Smith, Jeremy & Zhang, Bailing. (2016). “Action detection in office scene based on deep convolutional neural networks” 233-238. 10.1109/ICMLC.2016.7860906.

Appendices

Appendix A: Research Reflection

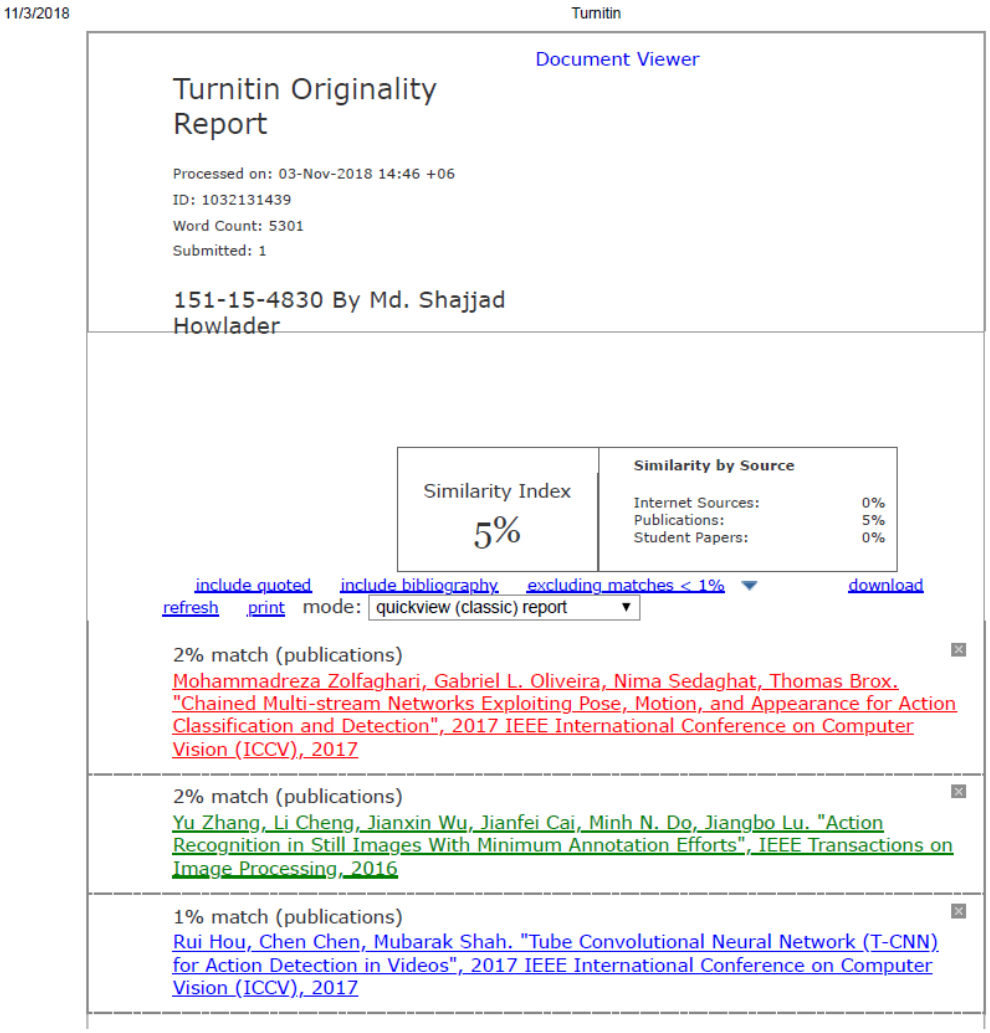
In this appendix, we are going to contribute an introduction to research reflection. To work in a group for this research project was challenging and enjoyable too for many tasks like data collecting and meet new peoples. We do not have much scope to work in a group always. It gives us the chance to work in a group and learn the team culture and how to make a team dynamic.

We gain the experiences that making the plan and take decisions need longer time than an individual decision. To do a good work, need to give the great effort by all the teammates. For this research project, we continue developing and refining each other's idea. We had to go to the library and required permission for collecting data .Our supervisor sir help us lot to manage permissions. It was enjoyable and challenging too. We enjoyed a lot by talking to the students who helped us a lot. This research result would help to evaluate the library actions to students.

Appendix B: Related Issues

During our research, we face many challenges. We had to learn new Machine Learning, Deep Learning techniques. We had to learn Convolutional Neural Networks and different algorithms like YOLO v3. It was difficult to collect image from the library by maintaining silence. We had to go to the library and capture student's action. We had to talk to the students and convinces them to give the poses for the collection of different action data. They were very friendly and cooperative.

We had to annotate the images carefully. Cause wrong annotations will produce wrong training. Overall our journey was good. We learn a lot of things which will help us in our near future.



11/3/2018 Turnitin

Document Viewer

Turnitin Originality Report

Processed on: 03-Nov-2018 14:46 +06
ID: 1032131439
Word Count: 5301
Submitted: 1

151-15-4830 By Md. Shajjad Howlader

Similarity Index 5%

Similarity by Source
Internet Sources: 0%
Publications: 5%
Student Papers: 0%

include quoted include bibliography excluding matches < 1% download
refresh print mode: quickview (classic) report

2% match (publications)
Mohammadreza Zolfaghari, Gabriel L. Oliveira, Nima Sedaghat, Thomas Brox. "Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection", 2017 IEEE International Conference on Computer Vision (ICCV), 2017

2% match (publications)
Yu Zhang, Li Cheng, Jianxin Wu, Jianfei Cai, Minh N. Do, Jiangbo Lu. "Action Recognition in Still Images With Minimum Annotation Efforts", IEEE Transactions on Image Processing, 2016

1% match (publications)
Rui Hou, Chen Chen, Mubarak Shah. "Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos", 2017 IEEE International Conference on Computer Vision (ICCV), 2017