Journal of Positive School Psychology
2022, Vol. 6, No. 3, 8530–8537

Detection and Classification of Weightlifting Form Anomalies using Deep Learning

Mohd Shazwan Sapwan[1], Zaidah Ibrahim[2], Zulaile Mabni[3*] and Noor Latiffah Adam[4]

[1,2,3,4] *Faculty of Computer And Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam,*
Selangor, Malaysia
[1]*shazwan.mohd1997@gmail.com,* [2]*zaidah@tmsk.uitm.edu.my,* [3*]*zulaile@tmsk.uitm.edu.my,*
[4]*latiffah@tmsk.uitm.edu.my*

Abstract

Detection and classification of weightlifting anomalies is important to prevent the risk of inflicting injury and to maximise the effect of the weightlifting exercises. During COVID-19 pandemic, going to the gymnasium is not possible. Thus, for those who have the necessary weightlifting equipment at home but no instructor, having automatic anomalies detection is beneficial. Although, weightlifting form recognition or anomalies detection often require utilisation of external sensors or hardware such as motion and kinetic sensors to produce accurate feedback for the user, not many have access to these external requirements. Thus, the objective of this research is to develop a prototype that is capable of providing feedback on the correctness of weightlifting technique execution using computer vision by implementing deep learning method. One of the popular deep learning methods for detection and recognition is You Only Look Once version (YOLO). Since there is no publicly available dataset for training and testing purposes, videos and images on weightlifting are searched and extracted from the Internet. 387 static images were collected with 219 images for normal forms and 134 images for abnormal forms. A confidence score in the range between 0.8 and 0.9 has been achieved during testing. Even though the performance produced is not high which is mainly due to the size of the training data, it can still serve as a foundation for future implementation for identifying weightlifter's technique execution and help to maximize the exercises.

Keywords: deep learning, weightlifting anomalies detection, You Only Look Once, YOLO

I. INTRODUCTION

Weightlifting is a challenging physical sport as it demands extreme technique and posture movement to properly execute its exercise (Henoch, 2017). This sport requires an explosive force generation to lift the barbell from the floor to its successful overhead and arms locked-out position. The snatch part involves a single movement to bring the barbell to the final position whereas the clean and jerk consists of two-part movement to lift the barbell to its final overhead position. There is no special trick involved in perfecting and mastering the techniques, only constant exposure, and practise in the arts. These exercises are called complex lifts that consist of several consecutive movements attached together to form one uniform action (Nagao, 2019; Mastalerz, 2019).

Training is essential in weightlifting to develop techniques that allow athletes to incrementally lift heavier weights. To incorporate the best possible technique, specialists isolate and correct the errors in techniques that may hinder performance and increase the risk of injury. Weightlifting is considered as one of the most injury prone training methodology due to applying improper techniques of lifting the weights (Yasser, 2019).

The risks prevalent to weightlifting include factors such as unsuitable environment, lack of proper equipment, excessively fatigued training methods, poor technique execution, excessive load, and volume of training methods, incorporating limited rest intervals and recovery, and lack of supervision from qualified person (Woods, 2019). In weight training, a correct exercise execution is critical for maximizing its effectiveness and for eliminating injury risks. However, given the complexity of these movements, it is a challenge for beginners to be aware whether they are performing the exercise the correct way or not (Kowsar, 2016).

A major obstacle of monitoring weightlifting exercises in an automated manner is the complexity and high number of degrees-of-freedom of human movement which implies the huge number of possible mistakes for any given exercise (Kowsar, 2016). There are applications that can capture and evaluate weightlifting performance but requires the aid of external sensors or other hardware requirements to perform their task such as Your Next Personal Trainer (Garcia, 2015). But not many have access to such requirements. Therefore, this research proposes to develop a prototype that is capable of the same task while being fully software-based and only utilising computer vision.

This paper evaluates an individual's weightlifting form and identifies anomalies through video inputs. There are various types of weightlifting exercises, however for this project the exercises are narrowed down to four variations, namely barbell back squat, dead lift, barbell overhead press and barbell row. Since a publicly available dataset for training and testing purposes is not available, this research creates its own dataset that are extracted from the Internet and social media platform, such as YouTube and Instagram. The frames from the videos were extracted at the rate of 1 frame per second (FPS) and utilised as the dataset for training and testing. The total data gathered for training is 387 images, with 219 images for normal forms (164 for training and 55 for testing) while 168 images for abnormal forms (134 for training and 34 for testing). The labels used are correct and incorrect which represent correct and incorrect form, respectively. To test the functionality aspect of YOLOv4 detection and classification performance, six videos of a person performing weightlifting in a gymnasium with varying durations under different views and gender were used.

II. METHODS

Object detection draws a bounding box around the object of interest in an image while object classification assigns the detected object a class label. There are various types of deep learning methods that can be applied for object detection and classification that are Region-based Convolutional Neural Network (R-CNN), Faster R-CNN, and You Only Look Once (YOLO).

R-CNN was first proposed by Ross Girshick and his team back in 2015 to tackle the problem of object localization and scarcity of training data that was present in the previous architecture (Girshick, 2015). R-CNN consists of three modules:

i. The first module generates region proposals that are independent to categorization.
ii. The second module extracts feature vector from proposed regions by using a convolutional network.
iii. The third module is a set of linear Support Vector Machines (SVM).

Figure 1 provides an overview of how R-CNN performs object detection and classification. It proposes region of interest from the input image and extract it to feed the CNN features for classification.

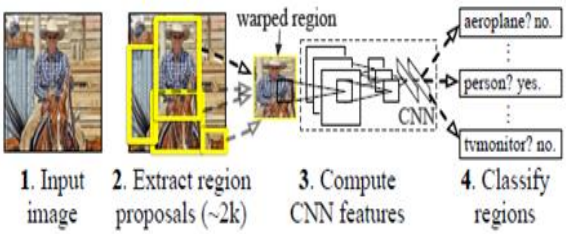

Figure 1 – Overview of R-CNN

Source: (GIRSHICK, 2015).

Although R-CNN achieves excellent object detection accuracy, it also has considerable drawbacks as a trade-off (Bao, 2016) that is as follows:

i. Training is a multi-stage pipeline.
ii. Training is expensive in terms of space and time.
iii. Detection of object is too slow.

To eliminate these deficiencies, Ross Girshick proposed an improved version of R-CNN, named Faster R-CNN (Girshick, 2015). Faster R-CNN introduces Region Proposal Network (RPN) which enables near cost-free regional proposals. RPN is a fully convolutional network that can produce object bounds prediction and object scores at each position concurrently (Ren, 2015). Figure 2 illustrates the architecture of Faster R-CNN. Faster R-CNN does have limitation for detecting small objects in images where large and small objects are intermixed causing low detection performance (Cheol-Roh, 2017).

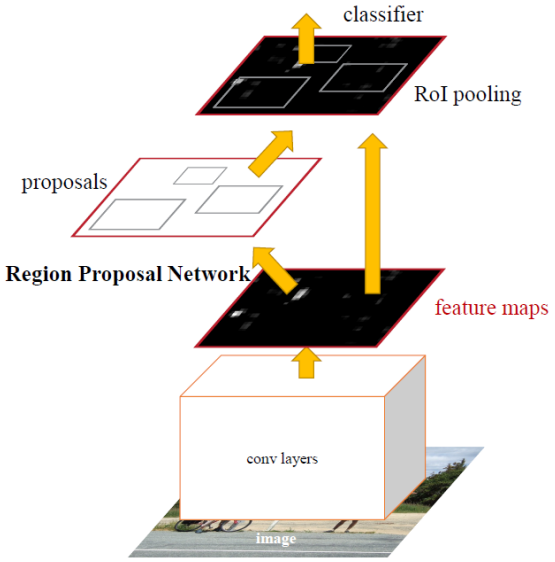

Figure2 – Faster R-CNN architecture

Source: (Redmond, 2016).

YOLO performs object detection by framing object detection as a regression problem to spatially separated bounding boxes and auxiliary class probabilities. In comparison to prevalent state-of-the-art detection systems, YOLO tends to produce more localization-related errors but in return it is less likely to predict false positives on the background (Shinde, 2018). Figure 3 briefly shows how YOLO performs its detection.

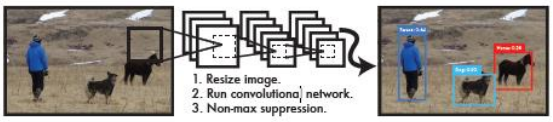

Figure 3 – YOLO detection system

Source: (SHINDE, 2018).

Rather than implementing a two-step method for classification and localization of object like conventional CNN-based detectors, YOLO only applies a single CNN for both classification and localization of the object. YOLO is also capable of processing images at about 40-90 Frames Per Second (FPS) meaning streaming video can be processed in real-time with negligible latency in a few milliseconds (Cao, 2019). Since YOLO is a one-stage algorithm that combines both target location and target recognition into a singular end-to-end detection process altogether, it can afford to take into account both speed and accuracy (Dixit, 2019).

YOLOv2 (also known as YOLO9000) executes at different sizes, giving a trade-off between speed and accuracy. It is a significant upgrade from its predecessor due to the implementation of features such as batch normalization, high resolution classifier and convolutional with anchor boxes. YOLOv2 manages to achieve 78.6 mean Average Precision (mAP) at 40 FPS, significantly outperforming Faster R-CNN. YOLOv2 utilizes Darknet-19 model as its backbone architecture (Shinde, 2018).

YOLOv3 utilizes a new network that combines Darknet-19 and a residual network for feature extraction purposes. The new network is called Darknet-53 because it has 53 convolutional layers (Redmond, 2018). YOLOv3 differs from its predecessor YOLOv2 by achieving object score prediction for each bounding box using logistic regression instead of using dimension clusters as an anchor box. If a prior bounding box is not assigned to an object, it incurs no loss for coordinate or class predictions, jeopardizing only object scores (Bochkovskiy, 2020). Figure 4 shows a sample of bounding box prediction using logistic regression.

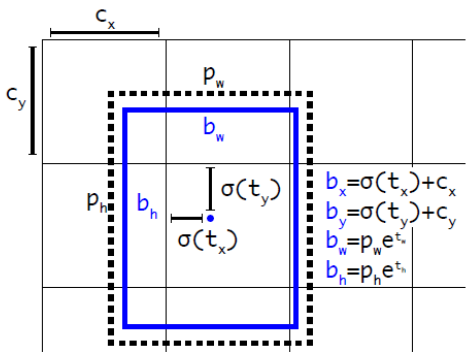

Figure 4 – Bounding boxes with dimension and location prediction for YOLOv3

Source: (Bochkovskiy, 2020).

YOLOv4 was chosen as the approach for this project since detecting weightlifting performance needs to provide near instantaneous feedback as its functional aspect since it is faster than its predecessor (Bochkovskiy, 2020). YOLOv4 consists of three main components that are the backbone which is the CSPDarknet53 (Wang et. al. 2020), the neck which is the Spatial Pyramid Pooling (SPP) (He, 2015) and the head which is YOLOv3. Figure 5 illustrates the components of YOLOv4.SPP works by dividing the feature maps output by the last layer of CSPDarknet53 into several spatial bins with sizes proportional to the image size. The head part performs the detection of the bounding box that consists of two sub-components that are one-stage detector and two-stage detector.

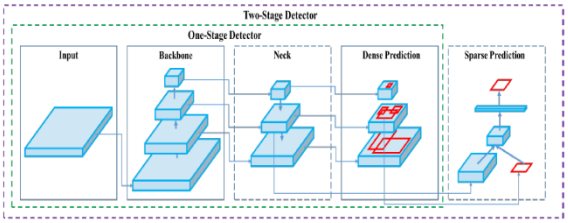

Figure 5 – The main components of YOLOv4

Source: (Bochkovskiy, 2020).

III. RESULTS AND DISCUSSION

In order to train YOLOv4 to detect weightlifting form anomalies, firstly the pre-trained model, CSPDarknet-53, is loaded to initialise the parameters for the feature extraction network. Each epoch will be divided into 64 batches and each batch size was set to 4000 iterations. The initial learning rate for the training is set to 0.01 and the initial network size is set to 416 on both width and height for the input image. These configurations will affect the overall training time along with model accuracy and will be changed accordingly to produce the best model for detection purposes, based on mean average precision (mAP) as the detection evaluation criteria. Once the training process has completed, a video stream is used as input and each frame of the video is converted to an image for detection and classification. Each of the image is passed as parameter for the localisation of bounding boxes and classification method which will plot the bounding boxes on the output image and perform classification of the object within the bounding box.

The dataset was repeatedly trained by fine-tuning different configurations including pre-trained weights, network size and learning rate to identify the best possible model for detection and classification. A total of 8 experiments have been conducted using different values of a few hyper-parameters namely learning rate and network size during training to identify the best possible model for detection using mean average precision. Precision is defined as the following equation:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

where True Positive is the correctly predicted class for the object, False Positive is the incorrectly predicted class for the object. Average Precision is the average precision score across the training. Mean Average Precision (mAP) is the mean of the Average Precision score gathered from the training phase. The formula for mAP is as follows (N represents the number of classes) as follows.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} Average\ Precision_i$$

From these experiments, there are certain trends that can be observed. For example, increasing

the network size gradually from 256x256 to 512x512 while retaining the original learning rate of 0.001 using the original CSPDarknet54 pre-trained weights, the mAP fluctuates. On the other hand, applying learning rate of 0.01 with network size of 512x512 seems to achieve the highest mAP which is 0.57. Henceforth, the testing process utilizes this model for detection and classification. Table 1 shows the experimental results with different network size and learning rate during the training phase.

The detection and classification performance during the testing phase is evaluated using the confidence score computed as follows:

confidence score = box confidence score * conditional class probability

where

box confidence score = $P_r(object)$* IoU

conditional class probability = $P_r(class_i|object)$

$P_r(object)$ is the probability the box contains an object

IoU (Intersection Over Union) between the predicted box and the ground truth

$P_r(class_i)$ is the probability the object belongs to $class_i$

Table 1 – Results from experiments with YOLOv4 during training phase

Experiment	Parameters	Average Loss	mAP
1	cspdarknet54, 256x256, learning rate 0.001	0.301755	0.518881
2	cspdarknet54, 416x416, learning rate 0.001	0.285944	0.541007
3	cspdarknet54, 512x512, learning rate 0.001	0.293728	0.525758
4	cspdarknet54, 256x256, learning rate 0.01	0.224731	0.461186
5	cspdarknet54, 416x416, learning rate 0.01	0.238886	0.434154
6	efficientnetB0, 256x256, learning rate 0.001	0.072017	0.321269
7	efficeinnetB0, 416x416, learning rate 0.001	0.041905	0.337806
8	cspdarknet54, 512x512, learning rate 0.01	0.225372	0.574707

The testing phase involves various scenarios to examine the robustness of YOLOv4 model. The first phase of the test was to test the detection capabilities of YOLOv4 for a single individual (male) from the side view. The 1st test of this phase was conducted by using a video demonstration of an incorrect execution of deadlifting technique. The prototype managed to predict and detect the individual's form correctly as 'incorrect' class with high level of confidence score within the range of 0.8 and 0.9. The time taken for detection feedback for the video is 5.17 seconds.

The 2nd test of this phase was conducted by using a video demonstration of an incorrect execution of barbell row technique. The prototype managed to detect and classify the individual's form correctly as 'incorrect' class most of the time with high level of confidence score within the range of 0.8 and 0.9. The time taken for detection and classification feedback for the video is 5.24 seconds.

The second phase of testing was to test the capabilities of YOLOv4 to detect and recognize female individual(s) from the side view. The 1st test of this phase was conducted by using a video demonstration of two female individuals executing both correct and incorrect execution of barbell overhead press technique. The prototype however only managed to detect and correctly predict the left female's form as 'incorrect' class with varying level of confidence score within 0.8 to 0.9 while failing to detect the second individual. The time taken for detection feedback for the video is 5.05 seconds.

The 2nd test of this phase was conducted by using video demonstration of an incorrect execution of barbell back squat by a female individual. The prototype again predicted both 'correct' and 'incorrect' class for the individual's form, with mostly 'incorrect' class was predicted with confidence score is around 0.8. In this video, the prototype detected and

classifies the second individual and classifies it as 'incorrect'. The time taken for detection feedback for this video is 5.37 seconds.

The third phase of testing involves testing the detection and classification capabilities of YOLOv4 with the individual(s) located in different angles. The 1st test in this phase was conducted by using a video demonstration of an individual executing a correct execution of barbell overhead press technique. In the initial angle, the prototype managed to correctly detect and classify the individual's form as 'correct class' with an acceptable confidence score of 0.67. Once the view angle transitioned to another angle, the confidence score for the 'correct' class can be observed decreasing significantly and at one point incorrectly classified as 'incorrect' class. The prototype manages to detect all the three objects in the video. The time taken for detection feedback for this video is 4.98 seconds.

The 2nd test in this phase is conducted by using a video demonstration of correct barbell back squat technique execution. Initially, the prototype correctly detected and classify the individual's form as 'correct' with confidence score of 0.92, but wrongly classified the individual's form as 'incorrect' in certain frames. Once the angle transitioned to another value, the prototype managed to correctly predict the individual's form as 'correct', but the confidence score can be observed significantly drop from 0.92 percent to only 0.53 percent. The time taken for detection feedback for this video is 5.17 seconds.

Table 2 – Video Testing Results

Phase	Test	Results
1st phase – video of single individual (male) from the side view	1st test - video demonstration of an individual executing a correct execution of barbell overhead press technique	Max confidence score – 0.9 Detection feedback – 5.17 second(s)
	2nd test - video demonstration of an incorrect execution of barbell row technique.	Max confidence score – 0.9 Detection feedback – 5.24 second(s)
2nd phase – video of female individual(s) from the side view.	1st test – video demonstration of two female individuals executing both correct and incorrect execution of barbell overhead press technique.	Max confidence score – 0.9 Detection feedback – 5.05 second(s)
	2nd test - video demonstration of an incorrect execution of barbell back squat by a female individual.	Max confidence score – around 0.8 Detection feedback – 5.37 second(s)
3rd phase – video of individual(s) located in different angles.	1st test - video demonstration of an individual executing a correct execution of barbell overhead press technique	Max confidence score – 0.67 Detection feedback – 4.98 second(s)
	2nd test - video demonstration of correct barbell back squat technique execution.	Max confidence score – 0.92 Detection feedback – 5.17 second(s)

IV. CONCLUSION

The main objective of this project is to investigate the robustness of YOLOv4 to detect and classify weightlifting form anomalies from videos. The experiments indicate that YOLOv4 produces a high detection speed capabilities averaging at five seconds for each of the video used during testing. It manages to detect and classifies single and multiple objects in a video from different views. On the other hand, the low mAP during training and confidence score

during testing is due to the small size of the dataset. But it still has the potential to help inexperienced or novice individuals to identify mistakes in weightlifting form or pose and improve their techniques gradually. This prototype does not require any additional hardware to perform the detection and recognition. Future work includes increase the size of the dataset and investigates other detection and classification method like Single Shot Detector (SSD).

V. ACKNOWLEDGMENT

The authors gratefully acknowledge sponsorship of this project from the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.

BIBLIOGRAPHY

1. HENOCH, Q. (2017). Weightlifting Movement Assessment & Optimization. Catalyst Athletics, Inc.
2. NAGAO, H., KUBO, Y., TSUNO, T., KUROSAKA, S. & MUTO, M. (2019). A Biomechanical Comparison of Successful and Unsuccessful Snatch Attempts among Elite Male Weightlifters. Sports 7(6):151, doi:10.3390/sports7060151.
3. MASTALERZ, A., SZYSZKA, P., GRANTHAM, W. & SADOWSKI, J. (2019). Biomedical analysis of Successful and Unsuccessful Ssnatch Lifts in Elite Female Weightlifters. Journal of Human Kinetics 68(1), 69-79.
4. YASSER, A., TARIQ, D., SAMY, R., & HASSAN, M. A. (2019). Smart Coaching: Enhancing Weightlifting and Preventing Injuries. International Journal of Advanced Computer Science and Applications (IJACSA). Vol. 10, no. 7, 686-691.
5. WOODS, B. (2019). Youth Weightlifting - A Review on The Risks, Benefits, And Long-Term Athlete Development Associated with Weightlifting Amongst Youth Athletes. Journal of Australian Strength and Conditioning. 27(3), 53-68.
6. KOWSAR, Y., MOSHTAGHI, M., VELLOSO, E., KULIK, L., & LECKIE, C. (2016). Detecting Unseen Anomalies in Weight Training Exercises. OzCHI '16: Proceedings of the 28th Australian Conference on Computer-Human Interaction. 517-526, doi.org/10.1145/3010915.3010941.
7. GARCIA, B., KAPLAN, R., & VISWANATHAN, A. (2015). Your Next Personal Trainer: Instant Evaluation of Exercise Form. Retrieved January 4, 2021 from http://cs229.stanford.edu/proj2015/183_poster.pdf.
8. GIRSHICK, R., DONAHUE, J., DARRELL, T., & MALIK, J. (2015). Region-based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, 142-158.
9. BAO, Y., LI, H., FAN, X., RISHENG, L. & JIA, Q. (2016). Region-based CNN for Logo Detection. Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS'16). 319-322.
10. REN, S., HE, K., GIRSHICK, R., & SUN, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137-1149.
11. CHEOL-ROH, M., & LEE, J. Y. (2017). Refining Faster-RCNN for Accurate Object Detection. Fifteenth IAPR International Conference on Machine Vision Applications (MVA). 514-517, doi: 10.23919/MVA.2017.7986913.

12. REDMOND, J., DIYVALA, S., GIRSHICK, R., & FARHADI, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788.
13. SHINDE, S., KOTHARI, A., & GUPTA, V. (2018). YOLO based Human Action Recognition and Localization. Procedia Computer Science, vol. 133, 831-838.
14. CAO, C., JIA-CHUN, Z., YI-QI, H., LIU, H., & CHENG-FU, Y. (2019). Investigation of a Promoted You Only Look Once Algorithm and Its Application in Traffic Flow Monitoring. Applied Sciences 9(17), 3619, https://doi.org/10.3390/app9173619.
15. DIXIT, K. G. S., CHADAGA, M. G., SAVALGIMATH, S. S., RAKSHITH, G. R., & NAVEEN KUMAR M. R. (2019). Evaluation and Evolution of Object Detection Techniques YOLO and R-CNN. International Journal of Recent Technology and Engineering (IJRTE). Vol. 8, Issue 253, 824-829.
16. REDMOND, J. & FARHADI, A. (2018). YOLOv3:An Incremental Improvement. ArXiv, abs/1804.02767.
17. BOCHKOVSKIY, A., WANG, C., & LIAO, H. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934[cs.CV]..