Adapting YOLO Network for Ball and Player Detection

Matija Burić[1][a], Miran Pobar[2][b] and Marina Ivašić-Kos[2][c]
[1]*Hrvatska elektroprivreda d.d., SIT Rijeka, Kumičićeva 13, Rijeka, Croatia*
[2]*Department of Informatics, University of Rijeka, Rijeka, Croatia*

Keywords: Object Detector, Convolutional Neural Networks, YOLO, Sports, Handball.

Abstract: In this paper, we consider the task of detecting the players and sports balls in real-world handball images, as a building block for action recognition. Detecting the ball is still a challenge because it is a very small object that takes only a few pixels in the image but carries a lot of information relevant to the interpretation of scenes. Balls can vary greatly regarding color and appearance due to various distances to the camera and motion blur. Occlusion is also present, especially as handball players carry the ball in their hands during the game and it is understood that the player with the ball is a key player for the current action. Handball players are located at different distances from the camera, often occluded and have a posture that differs from ordinary activities for which most object detectors are commonly learned. We compare the performance of 6 models based on the YOLOv2 object detector, trained on an image dataset of publicly available sports images and images from custom handball recordings. The performance of a person and ball detection is measured on the whole dataset and the custom part regarding mean average precision metric.

1 INTRODUCTION

Object detection belongs to computer vision research field with the goal of classifying certain objects in images and providing their exact position. Many machine learning algorithms have been successfully applied through the last few decades for detection of objects such as human faces (Viola and Jones, 2001) or full human figures (Navneet and Triggs, 2005). Lately, the most widely used techniques for object detection are based on convolutional neural networks (CNNs) such as Mask R-CNN (He et al., 2017) based on the R-CNN family (Girshick, 2015) (Shaoqing and et al, 2015), SSD (Liu and et al., 2016), R-fcn (Dai et al., 2016) etc. However, there is no universal solution for detection for all types of objects, but rather the choice depends on the task which needs to be performed.

Some CNNs perform faster than others, usually at the expense of accuracy, while some provide more information than just the bounding box around the desired object but are more complex and require more resources.

In this paper, we consider the task of detecting the players and sports balls in real-world handball images, which is essential for further research of action recognition in handball sports footages (Pobar and Ivašić-Kos, in Press).

Based on the previous results (Burić et al., 2018) and (Burić et al., in Press), we decided to use variations of the YOLO network (Redmon et al., 2016). The pre-trained YOLO network gave satisfactory results on the person detection task on the test handball dataset.

However, the ball detection proved problematic. In a typical team sport, including handball, the ball is a small and fast-moving object that typically occupies only a very small part of a frame, yet carries a lot of information important for the interpretation of the scene.

The balls themselves can vary greatly regarding color and appearance in the images due to various distances to the camera and motion blur, yet the commonness of their shape makes it easy to confuse

[a] https://orcid.org/0000-0003-3528-7550
[b] https://orcid.org/0000-0001-5604-2128
[c] https://orcid.org/0000-0002-1940-5089

Burić, M., Pobar, M. and Ivašić-Kos, M.
Adapting YOLO Network for Ball and Player Detection.
DOI: 10.5220/0007582008450851
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 845-851
ISBN: 978-989-758-351-3

similarly shaped objects for sports balls, e.g., lamps or player's heads.

Occlusion of already small objects is also present, especially since handball players carry the ball in their hands during the game.

Handball players themselves take positions and posture that are not common to people who walk, sit, or do the usual actions for which most models for detection or classification are commonly learned. Also, when detecting players in the field, there is a problem of occlusion, different distance from the camera, different color of sportswear that sometimes do not differ from the background and a lighting problem.

For these reasons, the experiment in this paper deals with testing the various YOLO based models and network training scenarios using different datasets to improve the sports ball and player detection results on handball images.

In the next section, the Yolo object detector will be described. In Section 3, the experimental setup covering the used dataset, the prepared environment, and the modification of each model is presented. The results are presented in Section 4, followed by the conclusion.

2 YOLO OBJECT DETECTOR

Yolo stands for "You Only Look Once" which describes an approach used by a single-stage network architecture that predicts the class probabilities along with corresponding bounding boxes in a single stage, as shown in Figure 1.

In the original Yolo model, the network architecture has 24 convolutional layers plus two additional fully connected layers.

The convolutional layers perform feature extraction, and the fully connected layers calculate the bounding boxes predictions and probabilities.

The bounding box predictions and class probabilities are associated with grid cells so that if an object occupies more than one cell, the center cell will be designated to be the holder of prediction for a particular object.

In the training phase, when a bounding box prediction holds no object the associated confidence value is zero, and if it holds an object and detection should occur, the confidence represents the intersection-over-union (IOU) score of prediction and ground truth (GT) boxes.

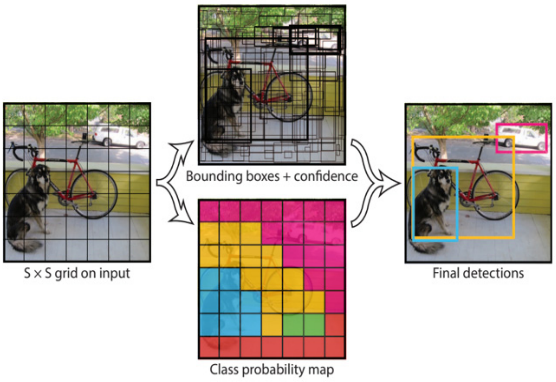

Figure 1: YOLO detection pipeline: the image is divided into S x S grid where the bounding boxes are simultaneously predicted with confidence and class probability for a final decision (Redmon et al., 2016).

There have been few versions of YOLO since it was first announced, the latest release being YoloV3 (Redmon, 2018). However, in order to preserve consistency with the previous experiments (Burić et al., 2017) (Burić et al., in Press), the YoloV2 version (Farhadi and Redmon, 2017) will be used as the starting point here.

YoloV2 differs from the original architecture by five convolution layers which were replaced with max-pooling layers.

Also, the fully connected layers are no longer present in the YoloV2.

To adjust the bounding boxes, YoloV2 uses predefined anchor boxes instead of the proposal box coordinates predicted by each cell in the earlier version.

To define the anchor boxes, YoloV2 uses k-means clustering in a training set of GT bounding boxes where boxes translations are relative to a grid cell.

3 EXPERIMENT

In the experiment, we compare the performance of 6 models based on the YOLOv2 network, trained on an image dataset comprising publicly available sports images and images from custom handball recordings.

The performance of person and sports ball detection is measured on the whole dataset and on the custom part specifically to get an estimate which model would be most appropriate for further research covering action recognition in handball.

The evaluation is based on the mean average precision (mAP) criteria like the one used in the PASCAL VOC 2012 competition (Everingham, et al., 2010).

3.1 Datasets

The datasets used for training and testing the detectors can be divided into two parts.

The first, custom part of the dataset, was acquired from indoor and outdoor footage of handball practice and competition. The recordings were made by the authors of this paper during one week in handball school without additional scene preparation or player instruction to preserve real-world conditions. The subjects in the images are mainly youngsters with accompanying coaches, handling multiple sports ball objects. The recordings were done using GoPro cameras positioned at 1.5 m height at the border of the filed or from the spectator's viewpoint approximately 3.5 m high and 10 m away from the filed limit. Artificial lighting was present during indoor activities with some sunlight through windows. Outdoor scenes were taken during daytime with clear sky or with almost no clouds. From the 751 videos, at 1920x1080 (full HD) resolution and 30 frames per second, 394 training and 27 validating images were selected for training the models. The ball objects came in a variety of colors, and so did the players clothing mainly used for everyday sports activities.

The second, public part of the dataset was used to avoid overfitting and to prepare the model for detection in other sports. It consists of 1445 training and 13 test images of variable sizes from 174 x174 up to 5184 x 3456 with 1 to many ball occurrences on each. This part was gathered in part using an internet search engine and in part from publicly available COCO datasets (Lin et al, 2014). Here, the balls are not exclusive to handball sport and are of different sizes and colors. The persons in the images also take different positions and are dressed differently.

The complete dataset has 1837 images with over 3500 ball objects.

3.2 Models

The description of the tested models is given below. Since the goal of detecting sports balls requires the possibility of discerning small and distant objects in images, the input resolution was increased to 1024x1024 pixels in all models except the first two, where the original input size of 608 x 608 pixels was used. This was done since the ball objects in a large number of source full HD images take up just a few pixels, and the resizing of images to 608x608 resolution can make such objects invisible or almost invisible.

An additional change was made to some classes models need to detect. Since object detection in sports action doesn't require classes like a teddy bear and fire hydrant, models were trained solely on the ball and person classes. All other classes are not considered in this experiment. This affects Mean Average Precision (mAP) which will be used in metrics of how successful the models are.

Our reference model, further marked as Y, is the pre-trained YOLOv2 model with 608 x 608 input image size with weights pre-trained on the COCO dataset and no additional training by the authors of this paper. The pre-trained model contains the person and sports ball among other classes from the COCO dataset.

The model (Y+) is the pre-trained model Y fine-tuned on the 394 images from the custom part of the dataset. The training of this model has proven to be unstable, so only a limited number of epochs was performed. Since the same dataset, consisting of large images (up to 5184 x 3456) with small annotated objects (few pixels high/width), was applied in training without problems to the subsequent models with higher input resolution, it can be concluded that already small objects get too small when resized ending up in error.

The third model (YB) was trained using transfer learning, on both public custom parts of the dataset, 1837 images in total, training only for the ball class. In this and subsequent models in this experiment, the input image resolution was increased to 1024 x 1024 from 608 x 608 of the original model.

The YBP model included both ball and person classes and was trained using transfer learning. It was trained for approximately 80 epochs on both custom and public parts of the dataset.

In the YPB+ model, the custom dataset was doubled using flipped images and flipped annotations. The YPB model was fine-tuned with ten epochs to achieve this result.

For the sixth model (YPBF) it was decided to include all images so far, public, custom and flipped custom dataset. Since flipped custom images were mirrored around Y axes, it was decided to try flipping images around X axes as well which results in unnatural sky-ground and upside-down human position.

All of the models were trained and then tested in the same environment, consisting of a PC equipped with a 12 core E5-2680v3 CPU and one GeForce GTX TITAN X GPU with 12GB of memory, with Debian Linux operating system. Additional programming was done in Python programming language.

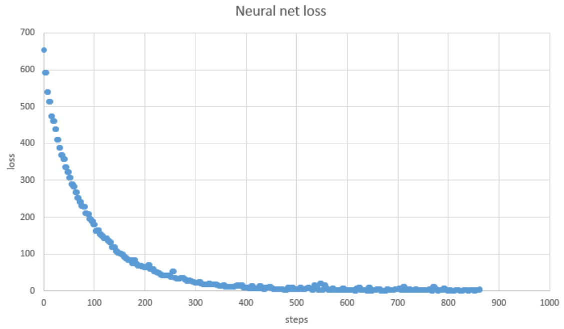

Figure 2: Typical convergence of moving average loss during neural network training.

The models were trained to at least 20 epochs except for the model Y+ which was unstable during training. First, few dozen steps were trained with higher learning rate to reduce the time it takes to converge and after that for more reliable training learning rate was lowered. A typical graph of training loss can be seen in Figure 2.

Transfer learning (Cook et al., 2013) was used to avoid training the models from the beginning. As the foundation, YoloV2 weights pre-trained on COCO dataset were used in all models. Because the COCO dataset incorporates a huge number of classes among which is sports ball, the features fused in pre-trained weights already possess basic information about ball objects.

The detection speed of all models don't differ significantly, and although the environment is suitable for the detection speed test, this aspect was not considered in detail.

4 RESULTS AND DISCUSSION

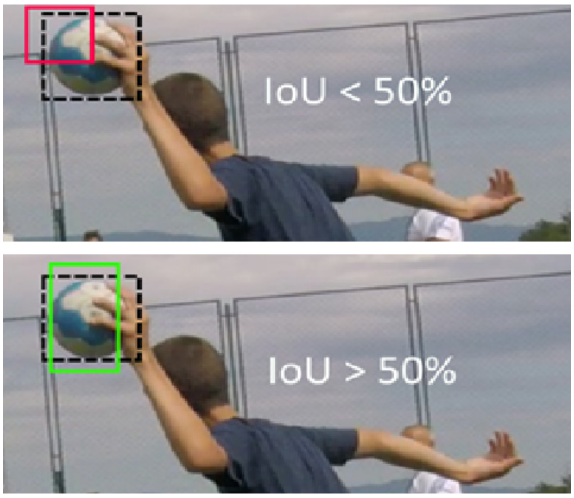

Figure 3: Visual representation of IoU criteria.

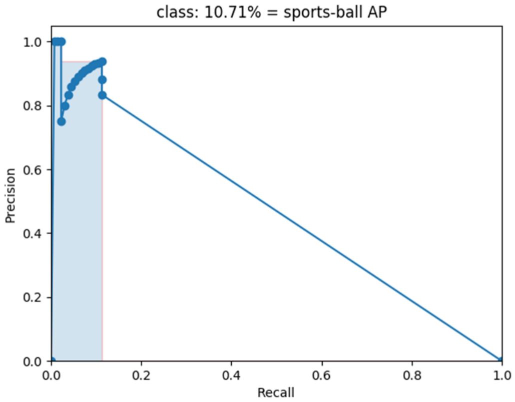

Figure 4: The precision-recall curve for model Y and ball detection. AP is the shaded area under the curve.

Average Precision (AP) will be used to evaluate the performance of the models. To calculate the AP, the result of the detection for each model is compared to the ground truth considering that the IoU should be equal to or greater than 50% (Figure 3). It is measured for each detected object and considered true positive if an object wasn't detected yet to avoid multiple detections of the same object. Based on these criteria, the precision-recall curve is calculated for every class, in this case – person and ball. AP is finally computed as an occupying area underneath the curve.

An example of the case of model Y and ball detection on is shown in Figure 4. To get the mAP value, mean of AP value of all classes is calculated.

Table 1. shows the mAP and class AP scores for all tested models on the whole and the custom (handball) part of the dataset.

Table 1: Class AP and mAP results of tested models (%).

	Whole dataset			Custom dataset		
Model	Ball AP	Person AP	mAP	Ball AP	Person AP	mAP
Y	10.71	48.74	29.73	0.94	43.47	22.20
Y+	10.33	52.99	31.66	0.94	31.45	16.20
YPB	8.78	**62.89**	**35.84**	7.37	**65.89**	36.63
YPB+	10.8	57.9	34.35	**10.61**	64.09	**37.35**
YPBF	**12.25**	54.18	33.22	7.15	59.50	33.32

The reference model Y performed well for near objects, especially persons, but it had difficulties with ball objects, both near and distant. It performed better on the public part of the dataset than on custom images, especially so for the ball class, for which it achieves AP of less than 1% on the custom dataset.

The Y+ model, which was fine-tuned with additional examples of sports ball and person classes, had slightly better detection of the person object, but the majority of distant objects were still undetected.

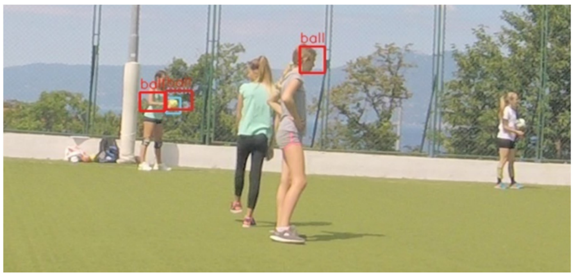

Figure 5: Example of the confused human head for a ball with YB model.

The YB model was trained and tested only for the sports ball class, and it had increased input image size. As was expected, the model performed better on distant objects, however, the detection of ball object on the public part of the dataset was severely degraded with many false positive detections of ball objects on human figures (Figure 5). This suggests that the person class should be included in the training set even when the goal is the only detection of balls.

The YPB model was trained on both person and ball class, and the performance has significantly improved in comparison to the YB+ model.

This model had the best score for person detection among all models tested so far, both considering only the custom part of the dataset and on average for the whole dataset. Ball detection has significantly improved for distant objects and on the custom dataset in comparison with Y and Y+, from 0.94% to 7.34%, even though the average score on the whole dataset dropped slightly. At this point, the results were improved enough in comparison to the base model to be acceptable for continuing research for action recognition in handball.

A simple data augmentation technique of using horizontally flipped images was employed for training the YBP+ model. Since manual annotation of images is a tedious and time-consuming process, requiring hours of human labor, this action has proven to be rewarding because it further improved ball detection. By fine-tuning the YPB model with flipped images, YPB+ model outperformed all other models for ball detection on both parts of the dataset. An example is shown in Figure 6. The final tested model included vertically flipped images in the training set, which further improved ball detection on average, with some degradation for the person class in comparison with the YPB+ model (Figure 8). This outcome was expected since ball objects are round and vertically flipping the image doesn't produce unnatural results, as it does for human figures. The results on the custom dataset have not improved, however.

It can be noticed that there is a huge gap between TP detection on public and custom test dataset for most models.

This gap narrows with YPB+ and YPBF models, where detection on the public dataset didn't degrade, so these models can more easily be applied to other sports besides handball.

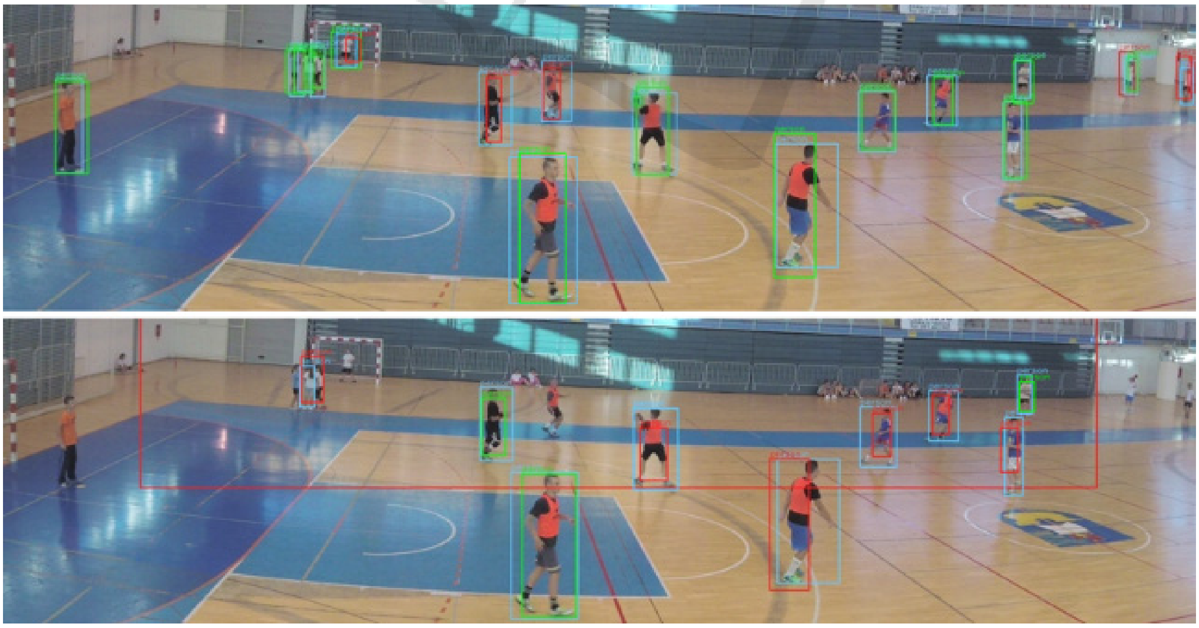

Figure 6: Comparison of YPBF (up) and fine-tuned pretrained Yolo (down) model. Green squares show detection with IoU > 50%. Red squares show FP.

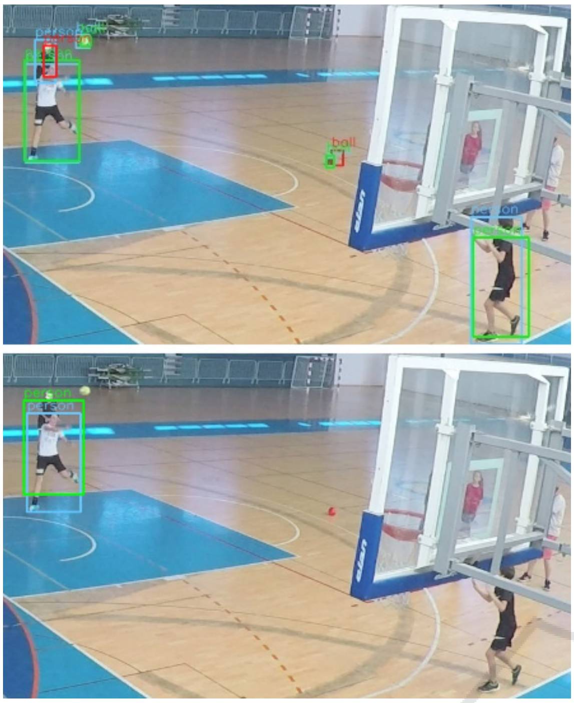

Figure 7: Two ball objects are detected using YPB+ (up) opposite to Y+ model (down), which also has difficulty detecting a person in the bottom right corner.

Occluded objects have proven difficult for all tested models, but YPB+ and YPBF models show better results on near and far occluded objects compared to Y and Y+.

5 CONCLUSION

In this work, we have tested several models derived from the baseline YOLO model for the task of sports ball and person detection.

The best results for person detection regarding mAP were achieved with the YPB model, which was trained on additional examples for both ball and person class and had an increased input image size.

The best score for ball detection was achieved with the model YPBF, which was also trained with flipped images of the same examples.

All models considered in this paper use low-level knowledge inherited from pre-trained publicly available weights. In such a way a time needed for training is drastically reduced in comparison with training the models from scratch, with the additional benefit of shared knowledge about low-level features trained on the large quantity of input data.

Simple modifications of input training data, in this case mirroring of images, has proven to be quite useful and cost-effective. More training images could be generated by performing partial rotation or scaling in the future. The YPB+ and YPBF models tested in this paper provide a better solution than the basic model for use in a sports action recognition framework. To the task of action recognition, ball and person objects need to be detected precisely as possible. Also, some other kind of information such as object tracking can be used in future work to improve both ball and player detection.

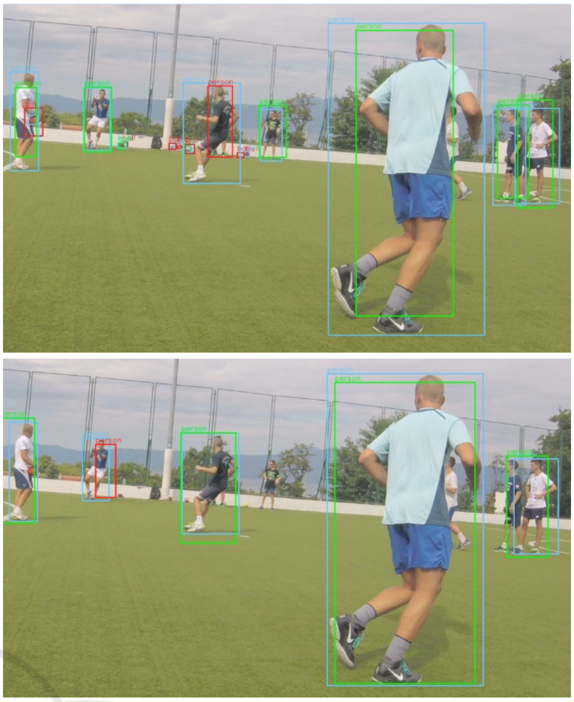

Figure 8: Person detection degraded when images flipped by x-axes are used with YPBF model (down).

ACKNOWLEDGMENTS

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS).

REFERENCES

Burić, M., Ivašić-Kos, M., Pobar, M., 2017. An Overview of action recognition in videos. *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* (pp. 1098-1103). IEEE.

Burić, M., Ivašić-Kos, M., Pobar, M., in Press. Ball detection using Yolo and Mask R-CNN. *5th Annual Conf. on Computational Science and Computational Intelligence (CSCI'18).*

Burić, M., Pobar, M., Ivašić-Kos, M., 2018. Object Detection in Sports Videos. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).*

Cook, D., Feuz, K. D., Krishnan, N. C., 2013. Transfer learning for activity recognition: a survey. *Knowledge and information systems, 36(3),* pp.537-556.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems,* pp. 379-387.

Everingham, M. et al., 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision,* Volume 88(2), pp. 303-338.

Farhadi, J, J., Redmon, A., 2017. YOLO9000: Better, Faster, Stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI,* pp. 6517-6525.

Girshick, R., 2015. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision.*

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV), Venice,* pp. 2980-2988.

Lin, T. et al, 2014. Microsoft COCO: common objects in context. *European conference on computer vision, Springer, Cham,* pp. 740-755.

Liu, W. et al., 2016. Ssd: Single shot multibox detector. *European conference on computer vision. Springer, Cham.*

Navneet, D., Triggs, B., 2005. Histograms of Oriented Gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE,.*

Pobar, M., Ivašić-Kos, M., in Press. Detection of the leading player in handball scenes using Mask R-CNN and STIPS. *ICMV 2018.*

Redmon, J., Farhadi A., 2018. Yolov3: An incremental improvement. *arXiv preprint ,* Volume arXiv:1804.02767.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV,* pp. 779-788.

Shaoqing, R. et al, 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems.*

Viola, P. A., Jones, M. J., 2001. Rapid object detection using a boosted cascade of simple features. *CVPR,* Issue 1, pp. 511-518.