BBM 418 - Computer Vision Laboratory
2022 Spring
Assignment 4

Murat Çelik
21827263
Department of Computer Engineering
Hacettepe University
Ankara, Turkey
b21827263@cs.hacettepe.edu.tr

1 Object Tracking use YOLOv3 and Mean-Shift

In this section, it is to detect the people in the video and estimate their location. This is called tracking by detection. For this, the YOLOv3 algorithm and the Mean Shift algorithm are used. The YOLOv3 algorithm is tested on a few videos. And the data is obtained. Then, the Mean Shift algorithm is used in addition to the algorithm in the parts that YOLOv3 cannot detect and are missing. All these results are analyzed.

2 Experiment Steps

2.1 Detect and Save data with YOLOv3

YOLOv3 (You Only Look Once, Version 3) is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. YOLO uses features learned by a deep convolutional neural network to detect an object. Versions 1-3 of YOLO were created by Joseph Redmon and Ali Farhadi.

2.2 Filter the Main Person in Videos

The YOLOv3 algorithm detects every person in the videos. In this study, analyzes are made for a single person in the video. For this reason, YOLOv3's detections are filtered until only one box remains for each frame. This filtering process is done by calculating the IoU value of the ground truth data and the data of the YOLOv3 algorithm. If the obtained IoU score is above 40%, it is kept, otherwise it is deleted. Thus, the YOLOv3 algorithm produces results only for the object studied in the video.

2.3 Video Creation with Filtered Data

In this section, there are two different data for each frame. One of these data is ground truth data. The other is the data that YOLOv3 predicts. It is known that ground truth data contains one data for each frame. Since YOLOv3's predictions have deficiencies and missed predictions, it cannot be said to contain one data for each frame.

When creating the video, 2 colors are used for the boxes. The blue ones of these boxes show ground truth data. The green one of these boxes shows the data of the YOLOv3 algorithm.

2.4 Video Creation with Filtered Data on Mean Shift Algorithm

In this section, the Mean Shift Algorithm is used. With this algorithm, the previous prediction data is analyzed for the frames where YOLOv3 is missing, and predictions are made. Thus, there is a chance to follow the tracked object for each frame.

When creating the video, 3 colors are used for the boxes. The blue ones of these boxes show ground truth data. The green one of these boxes shows the data of the YOLOv3 algorithm. The red one of these boxes is the data of the Mean-Shift algorithm.

2.4.1 Mean-Shift Algorithm

The Mean Shift algorithm has a basic logic for object tracking. It makes inferences from previous predictions for unpredicted frames. The position of the object in the previous picture is given. High concentration of data points of that object are found. For the next frame, similar data points are scanned around that object and a new estimate is made.

3 Experimental Results

We have 2 types of estimation data. One of them contains only YOLOv3's predictions. The other includes predictions from YOLOv3 + Mean-Shift. These data are separately subjected to ground truth data and IoU (Intersection over Union) calculation. A score is kept for each frame. These scores are saved as files in the drive folder. The score of that frame is given as 0 for data with no conflicts or no predictions. All scores are summed and the last stage is divided by the number of frames.

Model/Label	pedestrian1	iceskater1	gymnastics1
YOLOv3	0.583	0.291	0.318
YOLOv3 + Mean-Shift	0.589	0.352	0.399

3.1 Difference of YOLOv3 + Mean-Shift from YOLOv3

When we look at the scores, we can see that the Mean-Shift algorithm improves the results. It is predicted that the score may increase by making predictions on data sets that have never been predicted. It can be said that the algorithm makes good predictions at points of 2-3 frames, but the size of the error increases as the frame to be estimated increases.

3.2 Best Score

It is observed that the best score is taken from the "pedestrian1" dataset. The reason why this score is the best can be said to be the main reason that the YOLOv3 algorithm makes more predictions in the data. Easy to distinguish the human object inside from the background, not including any zoom-in, zoom-out, convenience has been observed here.

3.3 Average Score

It is observed that the average score is taken from the "gymnastics1" dataset. It can be said that the main reason for this score is the underestimation of the YOLOv3 algorithm in some parts of the data. There are various reasons for this situation. Difficulties have been observed here, as the human object in it cannot be easily distinguished from the background, and it contains zoom-in and zoom-out. There are also incomplete estimations due to the movements of the object in the data. The Mean-Shift algorithm contains incorrect predictions at these points.

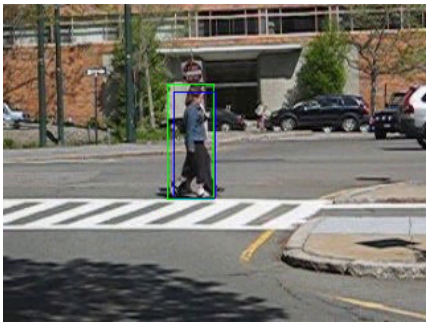

Figure 1: Frame one.

Figure 2: Frame two.

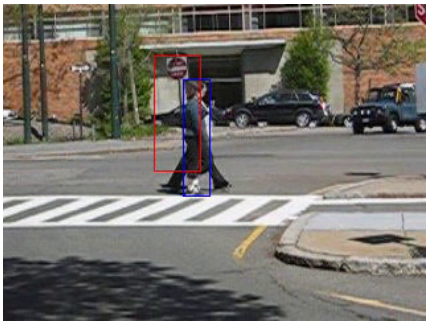

Figure 3: Frame Three.

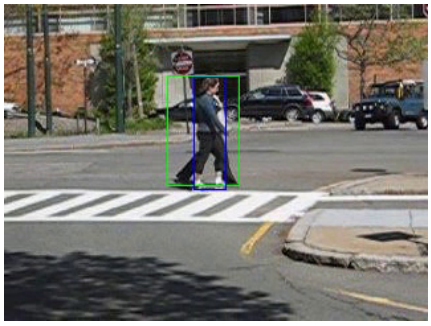

Figure 4: Frame Four.

Figure 5: A successful "pedestrian1" labeled model produced with the YOLOv3 + MeanShift model. Blue boxes are ground truth. Red boxes are Mean-Shift. Green boxes are YOLOv3.

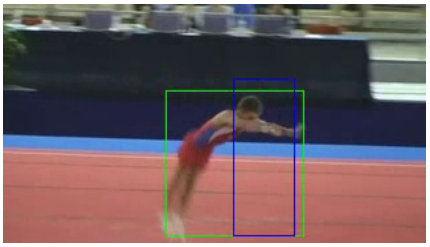

Figure 6: An example where the YOLOv3 estimate performs better than the Ground Truth.

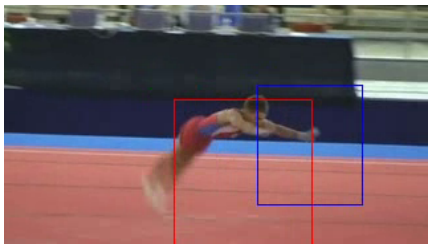

Figure 7: Prediction of Mean-Shift for next frame due to YOLOv3 prediction.

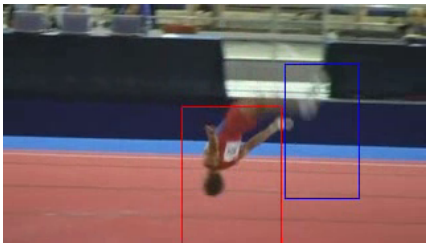

Figure 8: An example where the Mean-Shift estimate performs better than the Ground Truth.

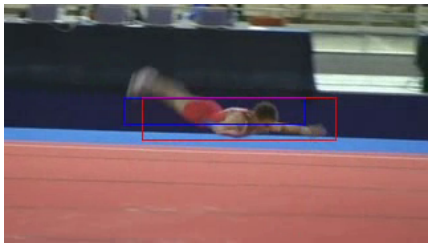

Figure 9: A successful Mean-Shift prediction.

Figure 10: Examples of successful and unsuccessful results. It was chosen according to the reasons affecting the scores. Blue boxes are ground truth. Red boxes are Mean-Shift. Green boxes are YOLOv3.

3.4 Worst Score

It is observed that the worst score is taken from the "iceskater1" dataset. It can be said that the main reason for this score is the underestimation of the YOLOv3 algorithm in most parts of the data. The second main reason is that movements such as the rotation of the object in the video, shrinking and jumping cause difficulties in estimation. Difficulties have been observed here, as the human object in it cannot be easily distinguished from the background, it contains too much zoom-in and zoom-out. There have been points where the Mean-Shift algorithm has been overused, and these points contain erroneous predictions.

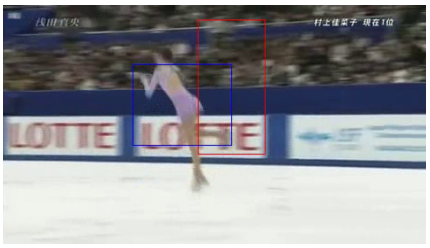

Figure 11: The frame in which the person turns.

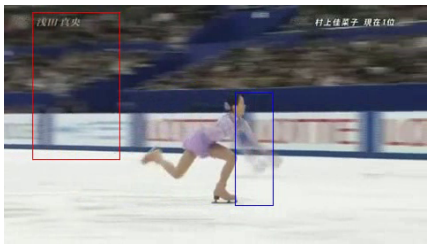

Figure 12: Incorrect prediction by the person's movement.

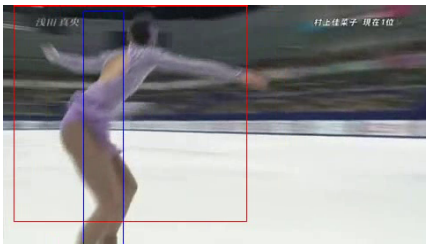

Figure 13: Zoom-in of the camera.

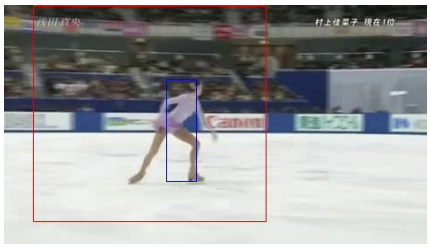

Figure 14: Zoom-out of the camera.

Figure 15: Some wrong prediction models. Red boxes is Mean-Shift.

4 Conclusion

In this part, we tested the object tracking algorithm. Videos were processed using YOLOv3 and Mean-Shift algorithms. All data is saved in the drive folder. Information was obtained about the difficulties of object tracking and the solution of these difficulties. Pytorch and Opencv libraries were used and experience was gained in computer vision.

Note : All data can be found in the drive folder.

5 References

[1] Mean-Shift Algorithm `https://docs.opencv.org/3.4/d7/d00/tutorial_meanshift.html`

[2] YOLO for Object Detection, Architecture Explained! `https://medium.com/analytics-vidhya/understanding-yolo-and-implementing-yolov3-for-object-detection-5f1f748cc63a`

[3] DataSet `https://www.votchallenge.net/vot2017/dataset.html`

[4] YOLOv3: Real-Time Object Detection Algorithm (What's New?) `https://viso.ai/deep-learning/yolov3-overview/`