**About the video**

The short video recorded is aligned with the list of MS COCO class names. The video is 13s long and it is in .mp4 format

There are 6 objects some of them are moving and some are static

* Moving objects: Person, Sports Ball, Hat.
* Static objects: Cup, Book, Teddy Bear.

Due to its shape, size, and color, the hat was detected as a cake in some frames and as a frisbee in others. As this object was not correctly identified, it is not part of the evaluation below.

The video is clear with good lighting. Although in some frames the moving objects are blurry due to the fast motion.

**Ground Truth**

Ground Truth is prepared using a tool called Computer vision Annotation Tool(CVAT). The Output of CVAT is in .xml format and the below code extracts the frame, object ID and the bounding boxes and export into a .csv file for future evaluation.

**Task1 & Task 2:** Object Tracking and Recognition

* The model used for object detection is a pre-trained Mask R-CNN
* Objects are detected frame-by-frame in the video (task1.mp4)
* Association method IoU is used to unique object IDs and track the objects
* Bounding Boxes and labels are drawn on the output video (task2.mp4) showing which object ID corresponds to which ID, with labels and scores.
* A confidence threshold of 0.5 is used to filter out low-confidence predictions.

**Task 3**: The faces, eyes, and smiles of the person were detected, and the face object was tracked.

* *Haar Cascades*: Three different pre-trained Haar Cascade classifiers are loaded :
  + haarcascade\_frontalface\_default.xml for face detection
  + haarcascade\_eye.xml for detecting eyes
  + haarcascade\_smile.xml for detecting smiles.
* *Face Detection*: For each frame, faces are detected using face\_cascade.detectMultiScale(). This function returns the coordinates of the bounding boxes around the detected faces.
* *Eye and Smile Detection*: Inside each detected face, detect eyes and smiles, and draw corresponding bounding boxes inside the face region.
* *Face Tracking:* Detected faces are tracked across frames using IoU-based matching, associating each face with a unique ID.
* *Bounding Box Drawing*: Bounding boxes are drawn around detected faces, eyes, and smiles. The ID of each tracked face is displayed above the face bounding box.
* *CSV Logging*: The face tracking information (frame number, object ID, bounding box coordinates) is logged to a CSV file (face\_detection\_results.csv).
* *Output Video*: The video with overlaid face detection results is saved as task3.mp4.

**Evaluation method:**

* A .csv file (object\_tracking\_results.csv) is also generated to capture the frame, object ID and bounding boxes coordinates for the detected objects.
* The Output of CVAT is in .xml format and the below code extracts the frame, object ID and the bounding boxes and export into a .csv file (groundtruth.csv) for evaluation.
* Comparison of Ground Truth and Detected object/ faces
  + The association method of IoU is used to compute the intersection over union between two bounding boxes.
  + For each object in the ground truth, we check if there is a corresponding detected face in the same frame with an IoU greater than or equal to the threshold (0.5 in this case).
  + If a match is found, it's counted as a true positive.
  + If no match is found, it’s counted as a false positive.
  + Evaluation Metrics:
    - True Positive (TP): Correctly detected faces/objects.
    - False Positive (FP): Detected faces/objects that don’t correspond to any ground truth face/object.
    - False Negative (FN): Faces/objects in the ground truth that were not detected.
    - Precision: Proportion of detected faces/objects that are correct.
    - Recall: Proportion of actual faces/objects that were detected.
    - F1-Score: The harmonic mean of precision and recall, representing the balance between both.

The evaluation results and analysis of the face and object detections and tracking tasks are as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Evaluation Metrics**  **& Analysis** | **Face Detection**  **task** | **Object Detection & Tracking Tasks** | | | | |
| **Person** | **Ball** | **Teddy** | **Book** | **Cup** |
| True positive (TP) | 139 | 290 | 191 | 391 | 272 | 329 |
| False Positives (FP) | 1 | 131 | 2 | 1 | 4 | 72 |
| False Negatives (FN) | 131 | 13 | 19 | 3 | 124 | 66 |
| Precision | 0.993 | 0.689 | 0.990 | 0.997 | 0.986 | 0.820 |
| Recall | 0.515 | 0.957 | .910 | 0.992 | 0.687 | 0.833 |
| F1 Score | 0.678 | 0.801 | .948 | 0.995 | 0.810 | 0.827 |
| **Analysis** | | | | | | |
| 2. Is your model able to detect and track the object in the video? | **Partially**, the model was not able to detect faces in all frames.  The model achieved high precision (99%) but low recall (51%) for face detection and tracking.  This indicates that while the model accurately detected most of the faces it identified, it missed detecting a significant portion of the actual faces present in the video. | **Partially**, the model was not able to detect persons in all frames.  For person detection, the model showed high recall (96%) but moderate precision (69%).  This means the model was able to detect nearly all persons in the video, but it also falsely identified objects as persons more frequently. | **Yes**, the model was able to detect and track the ball in the video with high precision (99%) and good recall (91%). This means it correctly identifies and tracks most of the time, with very few incorrect detections.  However, it missed detecting 9% of the actual ball instances in the video, which shows there is some room for improvement in recall. | **Yes**, the model was able to detect and track the teddy bear with exceptional performance. The precision and recall are both near-perfect, meaning that the model is highly reliable in identifying the teddy and tracking it across frames. | **Partially**, The model was able to detect and track the book with high precision (98.6%), meaning it accurately identifies the books it detects.  However, the recall is lower (68.7%), indicating that the model missed detecting a significant portion of the books in the video (31% were missed). | **Partially**, The model was able to detect and track cups in the video with reasonably good precision and recall.  However, the model misses around 17% of the cups (false negatives) and incorrectly detects objects as cups 18% of the time (false positives). |
| 2.1 Why or why not? | Why not?  there were frames where the face had lateral views and not frontal views. Also, as the face was moving it was blurry in some frames | Why not?  The person was moving and in some frame, it was a bit blurry. Also, the person enters and leaves the frame where the body was partial in some frames. | Why?  The ball had a bright orange color contrasting with the white background wall. Also, the perspective of the ball showing the shape of the ball was consistent throughout. However, in a few frames, the ball was obstructed by the hand leading to a few missed detection. | Why?  Similar to the ball the teddy was bright pink in color contrasting with the white background. Also, it was static throughout the video with some obstructions by the person in a few frames. | Why not?  Although the book was static, it was lying down on a stool and the perspective/view of the book was slanted and not very clear. Also, it was obstructed in a few frames by the person. | Why not?  Although the cup was static, the color of the cup was white almost same as the background wall. This might have led to the missed detection. Also, it was obstructed a few times by the person. |
| 3.What are the challenging issues leading to the initial unsatisfactory results? | Challenges include detecting moving or side-facing faces. Haar classifiers struggle with pose variations, scale, and position changes, leading to false negatives and missed small or distant faces. | A notable number of false positives, likely caused by background objects being mistaken for persons and inconsistent bounding boxes across frames. | The model performs well in detecting and tracking the ball, with a high F1 score of 0.948. The primary challenge lies in improving recall to reduce missed detections, especially under challenging conditions such as fast motion. | Not much of a challenge in this case | The angle of the book in which it was placed did not help much in detection. | The color of the cup was white same as the background.  As the cup was small in size, initially while creating the ground truth, the bounding box was larger which might have impacted the false positives. |
| 4.What have you done to improve the results? | Play around with the Scale factor and minNeighbours in detectMultiScale() function. | * Adjusting the confidence threshold helped minimize false positives while maintaining high recall.   + Higher threshold → Higher precision but Lower recall. * Adjusting the IoU threshold to associate object in different frames.   + Higher IoU threshold → Higher Precision, Lower Recall: More stringent matching improves precision by reducing false positives, but it reduces recall by missing some valid detections. * Adjusted the size of the bounding boxes in the ground truth. | | | | |
| 5.Other Findings or Investigations: | The scaleFactor controls image scaling during detection, with lower values increasing accuracy but slowing down processing, while minNeighbors sets the strictness for valid detections, with lower values being more sensitive but prone to false positives. A good starting point is scaleFactor=1.1 and minNeighbors=3-5 for balancing speed and accuracy. | * Precision vs. Recall Trade-off: Improving recall in object detection may lower precision, requiring a balance to optimize both. * Video Quality Impact: Higher resolution and less noise enhance detection and tracking accuracy. * Tracking Consistency: False positives in tracking could be reduced by using more advanced tracking methods. | | | | |