Report: Crowd Size Estimation on a Bridge

Problem Overview

The objective of this task is to estimate the number of people on a bridge during a specific time window (0:14 to 0:32) of a video using computer vision techniques. Additionally, a method to validate the estimation is required.

Chosen Approach

To estimate the crowd size, a hybrid approach using YOLOv9 for object detection and CSRNet for density-based crowd estimation was selected. This combination provides flexibility in handling varying crowd densities.

1. **YOLOv9 (You Only Look Once Version 9) for People Detection**:
   → YOLOv9 was utilized to detect individuals in moderate-density crowds. The model works efficiently by detecting bounding boxes around people in each frame of the video. A pre-trained YOLOv9 model (yolov9s.pt) was used to identify people (class cls[0]), and a bounding box was drawn around each detected person.

2. **CSRNet (Convolutional Neural Network for Crowd Counting)**:
   → CSRNet was employed for high-density crowd regions, where individual detection might fail. CSRNet predicts the density map, which is summed up to estimate the number of people in the given frame. The model uses a modified VGG16 network to extract features and predict the density maps. Pre-trained weights (CrowdCountingWeights.pth) were loaded to fine-tune the network for this specific task.

Code Explanation

1. **Preprocessing**:
   → The video was split into frames from the specified time window (0:14 to 0:32).
   → Each frame was passed through YOLOv9 for object detection and CSRNet for density estimation.
   → Frames were resized and normalized to fit the input format of the CSRNet model.

2.**People Counting with YOLOv9**:

→ The YOLOv9 model was used to detect bounding boxes for each individual in moderate-density frames.

→ A helper function count_people_with_yolov9() was defined, which iterated over the detected bounding boxes, filtered out non-human objects, and counted the people.

3.**CSRNet for Dense Crowds**:

→ For frames with high crowd density, the CSRNet model generated a density map, which was summed to estimate the crowd size.

→ The density map approach helped overcome challenges where occlusions or dense clusters of people might reduce the effectiveness of YOLO.

4.**Combination Approach**:

→ Both models were applied to different frames based on the crowd density detected in initial frames. For sparse and moderate density, YOLO was the primary model, while CSRNet was used for dense areas.

Alternative Solutions Considered

Alternative 1: YOLOv9 Only:

→ Initially, the idea of using YOLOv9 for the entire video was considered. However, YOLO struggles with dense crowds where occlusions and overlapping people make it hard to detect individual bounding boxes.

Alternative 2: CSRNet:

→ CSRNet alone could have been used, but it lacks the precision required for moderate or sparse crowds, where counting individuals directly is more efficient.

→ In the end, the hybrid approach was chosen to leverage the strengths of both YOLOv9 and CSRNet.

Additional Considerations

During this project, several enhancements were considered but could not be implemented due to time constraints:

1. Dataset Creation with Ground Truth Labels and Fine-tuning CSRNet:
   → I initially planned to create a custom dataset with ground truth labels for the bridge video frames. The goal was to fine-tune CSRNet specifically for this task, which would improve the accuracy and adaptation to the specific scene. However, due to time limitations, this approach was not pursued. Fine-tuning CSRNet with a more tailored dataset could have led to a significant improvement in detecting and counting people, especially in high-density areas.
2. Use of Transformer-based Models:
   → If more time was available, I also considered employing advanced transformer-based models, such as vision transformers, for better segmentation and object detection. These models have shown excellent performance in handling complex scenarios like crowd scenes. However, for this problem, the usage of large models like SAM (Segment Anything Model) was considered unnecessary. The problem at hand did not justify the computational overhead of transformer models, and the hybrid YOLOv9 + CSRNet approach was more suitable given the task's scale.

   → Both approaches would have likely enhanced the overall performance, but considering the scope and complexity of the problem, the current hybrid model was the most practical and efficient choice.

Validation of the Program's Estimate

To validate the results, manual counting was performed on a small subset of frames. The estimated counts from the model were compared with manual ground-truth counts to ensure accuracy. The difference was minimal, indicating that the model was effective.

Conclusion

The hybrid YOLOv9 + CSRNet approach offers a robust solution for estimating crowd sizes in varying density scenarios. The combination ensures precise detection in sparse and moderatedensity regions, while CSRNet effectively handles high-density crowds.