Crowd counting and monitoring for surveillance videos

Problem statement

A Crowd is a gathering of numbers of people at some place. It is not feasible to count/monitor all the people at various places like university, shopping malls, railway stations, airports or at any other place by looking at them. The complexity of monitoring, tracking and counting increases as the size of the crowd increases. We can't monitor the crowd for suspicious behavior as well. However, with the introduction of Closed Circuit Television (CCTV) camera this problem has been solved up to some extent. But still we are not able to track/monitor a large group of people with CCTV. In recent years numerous application of computer vision has come. Researchers are trying to monitor, and count the crowd automatically with the help of machine intelligence. It can be significantly advantageous if we can detect the objects from the videos/cameras. From a technological perspective, detecting, tracking, and analyzing peoples like detecting and tracking a person walking in a university, or identifying the communication between two people, computer vision solutions typically focused on these areas. Generally, there are number of CCTV cameras installed at various places to record the environment. But, the tough task is that it is not feasible for human operator to sit all the time in front of the CCTV and monitor/count the crowd. There is a need of an automated system which can provide us some meaningful information from live or recorded videos.

Detecting objects from CCTV surveillance videos can solve many real-life problems. If we can count/monitor the crowd then we can have the valuable information about the objects within the videos. Knowing that, how many peoples are coming in and going out in the premises can help us to draw valuable insights. One common challenge for any CNN based crowd counting and monitoring is to meet the real-time processing requirements where the Deep Learning model should run on embedded devices with limited processing power and energy. Another challenge in crowd counting is the occlusion, preserving the object across multiple frames when they overlap with each other. In the proposed work the Deep Learning based methods will be applied to recorded video of crowds to count/monitor the human beings.

Background

Despite the challenges, crowd counting and monitoring remains an active research area in computer vision in recent years. Numerous approaches have been proposed over the years. It has an obvious extension to surveillance applications due to the potential for improving safety systems. Many CCTV manufacturers (e.g. Hikevision, ClearView Communications, AllGoVision and Camlytics) now included these feature into their CCTV surveillance systems.

Earlier computer systems were rule based but with the introduction of machine learning, now machine can learn from the data and can act accordingly. Deep learning is the sub fields of machine learning. A lot of research is going on in object detection and object recognition from image and video. Most of the methods for crowd counting can be grouped into three categories: detection-based, global regression and density estimation. The earlier literatures of crowd counting propose the detection-based methods. The detection-based methods perform better in relatively low dense scenes, while, they are limited by the heavy occlusions in dense crowds. To overcome the difficulties of detection-based methods in high dense scenes, global regression based methods were introduced. Global regression based methods only utilize the information of

pedestrian counts, while, the spatial information and body structure information of pedestrians are ignored. Finally, density estimation based methods introduced. These methods demonstrate good performance on crowd counting.

Background subtraction can help to count and monitor peoples from video. Background subtraction classifies the pixels of video streams as either background, where no motion is detected, or foreground, where motion is detected. First detect the blob(s) via background subtraction, then track the blob until it goes off the frame. Once it leaves the frame, the next detected blob must be a new person entering the area and thus can continue counting. Other options could be using HOG (Histograms of Oriented Gradients) Descriptors, or Haar like features. Related work in crowd counting and monitoring can be found in reference section.

Methodology

The architecture of crowd counting and monitoring model is shown in fig 1.

Step 1: Data collection and dataset preparation

This will involve collection of publicly available video dataset from available sources. There are many publicly available datasets like UCSD pedestrian dataset, Grand Central railway station dataset etc.

Step 2: Developing a CNN based Crowd counting and monitoring *model*

To count the people from crowd video, a CNN based model will be developed. The model will have convolution layer, RELU and max pooling layer, fully convolutional layer and Softmax activation function. Firstly, individual frames will be extracted from the video and passed as input to the convolution layer. A convolution of appropriate size (M x N) will be applied on input image to extract the features and create a feature map. The model will be trained by some sample training data and then it will be tested by some test data. The feature extraction will be performed automatically by the model. When training and testing part is over, new input video will be applied and output of the model will be matched with ground truth values. Depending on the accuracy of the result more number of layers will be added or model will be trained against more number of epochs. Popular pretrained CNN feature extraction models such VGG16, ResNet or FCN (Fully Convolutional Network) as shown in fig 2 will be exploited for this task.

Step 3: Training and experimentation on datasets

The Crowd counting and monitoring model will be trained both on the large-scale datasets such as UCSD pedestrian dataset, Grand Central railway station dataset and live recorded videos populated based on CCTV recording camera as part of this project. The part of the dataset will be divided into training data and rest of the data will be used as testing data.

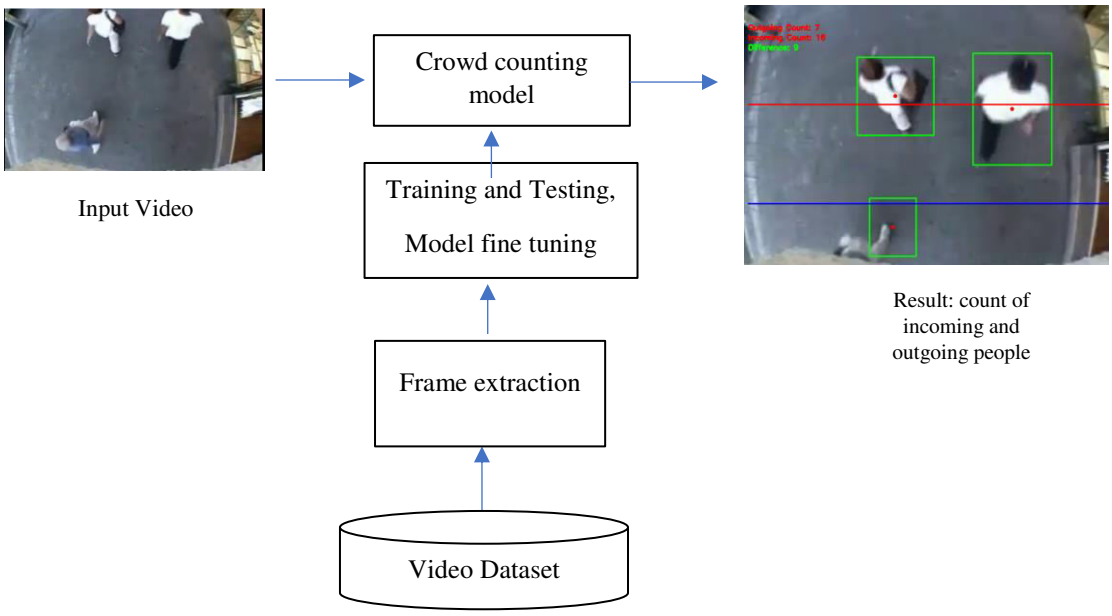

Fig 1. Architecture of crowd counting and monitoring

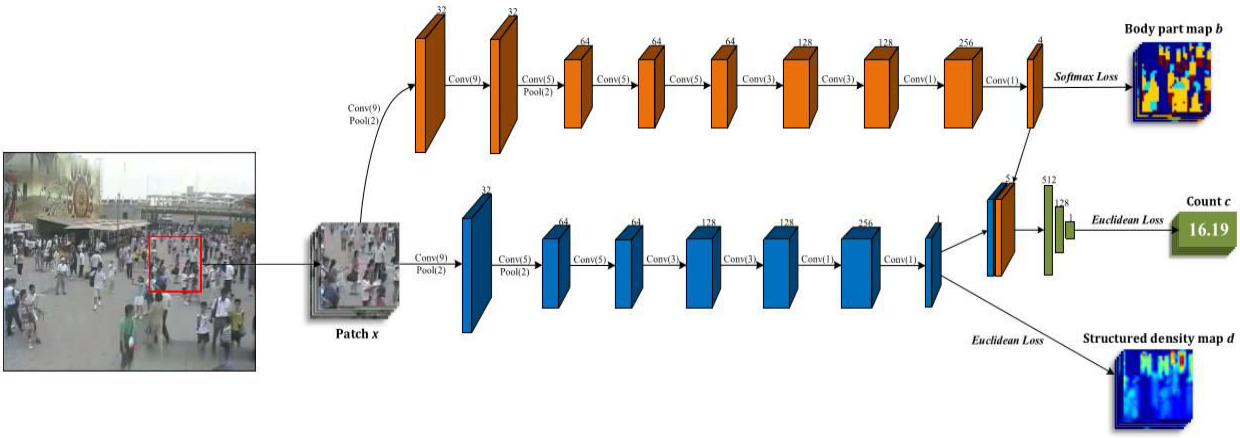

Fig 2. Architecture of FCN Based Crowd counting and monitoring [1]

Experimental Design

Dataset

The UCSD pedestrian dataset contains 2000 frames of a single scene. The video in this dataset is recorded at 10 fps with the frame size of 158×238. The dataset can be downloaded from [3]. Grand Central railway station dataset contains 50010 frames, Frame Rate: 25fps at 720x480. The dataset can be downloaded from [4].

Evaluation Measures

Measures such as accuracy, Mean Absolute Error (MAE) and Mean Squared Error (MSE) will be computed by comparing the counted results and ground truth values from the datasets.

Software and Hardware Requirements

For implementation and experimentation of the project, Python based Computer Vision and Deep Learning libraries will be exploited. Specifically, libraries such as OpenCV, Keras, TensorFlow, YOLO will be used. Training will be conducted on NVIDIA GPUs for training the end-to-end version of CNN based crowd counting model.

References

1. Huang S, Li X, Zhang Z, Wu F, Gao S, Ji R, Han J. Body Structure Aware Deep Crowd Counting. IEEE Transactions on Image Processing. 2018 Mar;27(3):1049-59.
2. Cong Zhang, Hongsheng Li, X. Wang, and Xiaokang Yang. 2015. "Cross-scene crowd counting via deep convolutional neural networks" In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 833–841.
3. http://visal.cs.cityu.edu.hk/downloads/ucsdpeds-vids/.
4. https://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html