Video Multi-Object Tracking

What and Where?

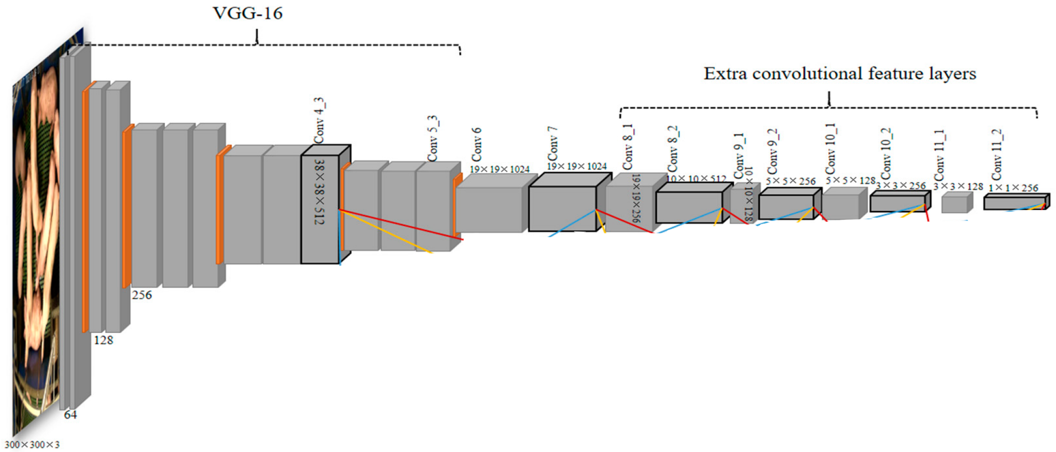
- Multi-object tracking is a computer vision task which can track objects belonging to different categories, such as cars, pedestrians and animals by analyzing the videos.
- Autonomous vehicles, security surveillance, robot navigation, crowd behavior analyses, action recognition are some of the applications which benefits from this high-quality MOT algorithm.
- The output of MOT algorithm is a collection of rectangular boxes associated with a target id to distinguish between the intra-class objects.

Why deep learning?

- The representational power of deep learning models have been exploited in recent years by more and more algorithms in order to learn rich representations and extract complex features from the input.
- Convolutional Neural Network(CNN) and Recurrent Neural Networks(RNN) currently constitute in state-of-the-art in tasks such as image classification or object detection.

Stages in MOT

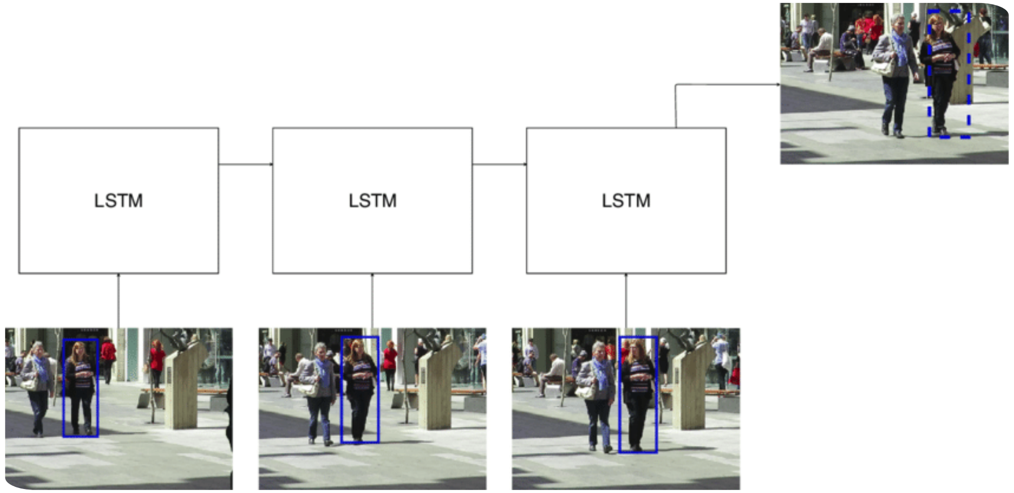
- 1. Detection Stage: The bounding boxes for the objects are obtained in this stage.
- 2. Feature extraction stage: One or more algorithms are used to analyze the detected objects to extract different features.
- 3. Affinity stage: This stage computes the similarity or distance and checks the probability of two objects belonging to the same objects.
- 4. Association stage: The similarity or distance measures are used to associate detection and a numerical ID is assigned to each object.



Detection Stage

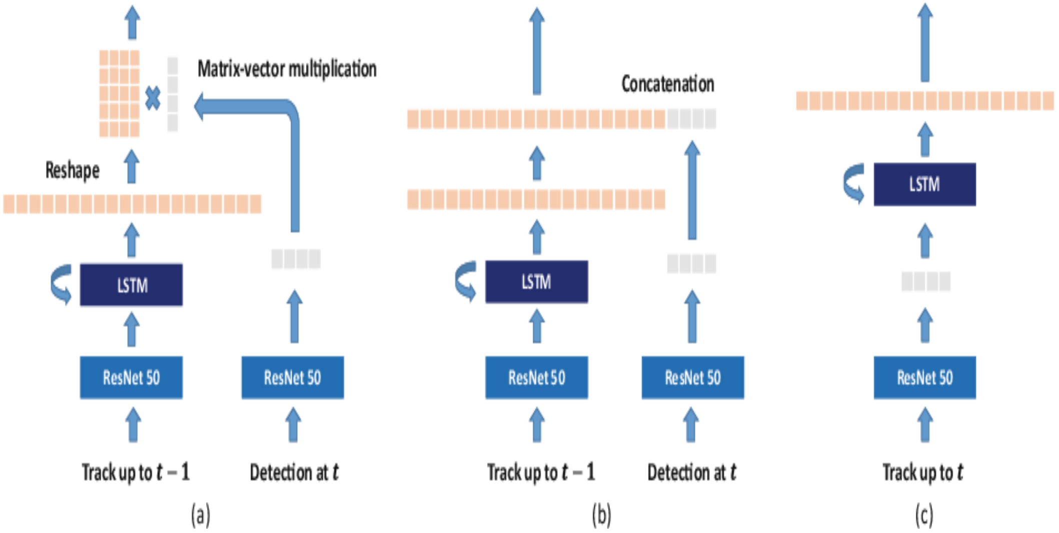
- The overall network consists of two parts: the front truncated backbone network (VGG-16) which is an image classification network and the additional convolutional feature layers which progressively decrease in size. upper from conv7 and lower layers form a single network used for default box generation as well as confidence and location predictions. More specifically, a set of default boxes are tiled with a convolution manner at different scales and aspect ratios for the feature maps.

Feature Extraction

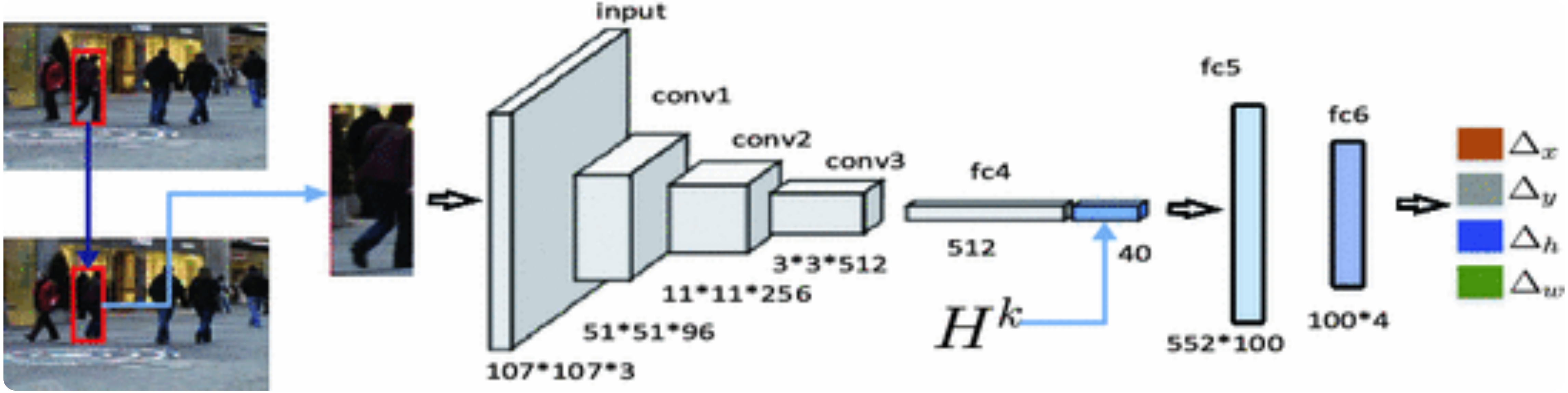


- This model uses three different RNNs to compute various types of features. The input for the first RNN is a visual feature vector extracted by a VGG CNN which is employed to extract appearance features. The second RNN is a LSTM trained to predict the motion model for every tracked object and the output is a velocity vector of each object. The last RNN is trained to learn the interactions between different objects on the scene, since the position of some objects could be influenced by the behavior of surrounding items.

Affinity stage



- The appearance features of a tracklet from the past frames, extracted with ResNet-50 CNN is taken as input. The output of the LSTM is a feature matrix which represents the historical appearance of the tracklet. This matrix is then multiplied by the vector with the appearance features of the detection that is needed to be compared with the tracklet. Fully connected layer on top computes the affinity score. This model was able to store longer-term appearance models than classical LSTMs. A motion modeling classical LSTM was added to compute historical motion features, which is then concatenated to the appearance features before proceeding with the FC layers and the final softmax that output the affinity. The two LSTMs are first trained separately and then fine-tuned jointly.



Association stage

- The algorithm is composed of a prediction network and a decision network. The prediction network is a CNN that predicts the movement of the target in the new frame looking at the new image and the target, also using the recent tracklet trajectory. The decision network is a collaborative system that consists of multiple agents and the environment. Each agent took decisions based on the environment, the information about themselves, and the neighbors. 3 FC layers is used to model the agents on top of feature extraction part of the MDNet.

The intent is to show how some of the deep model algorithms are used for multi-object tracking out of several others.