Video Multi-Object Tracking by Deep Learning

In Single Object Tracking (SOT), the bounding box of the target in the first frame is given to the tracker. The goal of the tracker is then to locate the same target in all the other frames. SOT belongs to the category of detection-free tracking, because one manually gives the first bounding box to the tracker. This means that Single Object Trackers should be able to track whatever object they are given, even an object on which no available classification model was trained.

While in Single Object Tracking (SOT) the appearance of the target is known a priori, in MOT a detection step necessary to identify the targets that can leave or enter the scene. The main difficulty in tracking multiple targets simultaneously stems from the various occlusions and interactions between objects that can sometimes also have similar appearance. Thus, simply applying SOT models directly to solve MOT leads to poor results, often incurring in target drift and numerous ID switch errors, as such models usually struggle in distinguishing between similar looking intra-class objects. A series of algorithms specifically tuned to multi-target tracking have then been developed in recent years to address these issues, together with a number of benchmark datasets and competitions to ease the comparisons between the different methods.

In Multiple Object Tracking(MOT), as its name indicates, there are multiple objects to track. The tracking algorithm is expected first to determine the number of objects in each frame, and second, to keep track of each object's identity from one frame to the next. MOT is a challenging problem: ID switches are hard to avoid especially in crowded videos, and the nature as well as the number of objects in each frame is unknown, so MOT algorithms strongly rely on detection algorithms, which are themselves not perfect.

Multiple Object Tracking (MOT), also called Multi-Target Tracking (MTT), is a computer vision task that aims to analyze videos in order to identify and track objects belonging to one or more categories, such as pedestrians, cars, animals and inanimate objects, without any prior knowledge about the appearance and number of targets. Differently from object detection algorithms, whose output is a collection of rectangular bounding boxes identified by their coordinates, height and width, MOT algorithms also associate a target ID to each box (known as a *detection*), in order to distinguish among intra-class objects. The MOT task plays an important role in computer vision: from video surveillance to autonomous cars, from action recognition crowd behavior analysis, many of these problems would benefit from a high-quality tracking algorithm.

Recently, more and more of such algorithms have started exploiting the representational power of deep learning (DL). The strength of Deep Neural Networks (DNN) resides in their ability to learn rich representations and to extract complex and abstract features from their input. Convolutional neural networks (CNN) currently constitute the state-of-the-art in spatial pattern extraction, and are employed in tasks such as image **classification** or **object detection**, while **recurrent neural networks (RNN)** like the **Long Short-Term Memory (LSTM)** are used to process sequential data, like audio signals, temporal series and text. Since DL methods have been able to reach top performance in many of those tasks, we are now progressively seeing them used in most of the top performing MOT algorithms, aiding to solve some of the subtasks in which the problem is divided.

In this research work, it is proposed to develop the algorithms that make use of the capabilities of deep learning models to perform Multiple Object Tracking, focusing on the different approaches used for the various components of a MOT algorithm and putting them in the context of each of the proposed methods. While the MOT task can be applied to both 2D and 3D data, and to both single-camera and multi-camera scenarios, in this survey we focus on 2D data extracted from videos recorded by a single camera.