The surveillance cameras serve various purposes ranging from security to traffic monitoring and marketing. However, with the increasing quantity of utilized cameras, manual video monitoring has become too laborious. In recent years, a lot of development in artificial intelligence has been focused on processing the video data automatically and then outputting the desired notifications and statistics. This thesis studies the state-of-the-art deep learning models for object detection in a surveillance video and takes an in-depth look at SSD architecture. We aim to enhance the performance of SSD by updating its underlying feature extraction network. We propose to replace the initially used VGG model by a selection of modern ResNet, Xception and NASNet classification networks. The experiments show that the ResNet50 model offers the best trade-off between speed and precision, while significantly outperforming VGG. With a series of modifications, we improved the Xception model to match the ResNet performance. On top of the architecture-based improvements, we analyze the relationship between SSD and a number of detected classes and their selection. We also designed and implemented a new detector with the use of temporal context provided by the video frames. This detector delivers enhanced precision while meeting real-time requirements.