Video retrieval represents a challenging problem with many caveats and sub-problems. This thesis focuses on two of these sub-problems, namely shot transition detection and text-based search. In the case of shot detection, many solutions have been proposed over the last decades. Recently, deep learning-based approaches improved the accuracy of shot transition detection using 3D convolutional architectures and artificially created training data, but one hundred percent accuracy is still an unreachable ideal. In this thesis we present a deep network for shot transition detection TransNet V2 that reaches state-of-the-art performance on respected benchmarks. In the second case of text-based search, deep learning models projecting textual query and video frames into a joint space proved to be effective for text-based video retrieval. We investigate these query representation learning models in a setting of known-item search and propose improvements for the text encoding part of the model.