This research work aims to develop and assess the capabilities of convolution neural networks to identify and localize musical instruments in 360 videos. Using audio and visual cues from 360 video frames along with object detection technologies, sound source separation technologies and sound classification technologies, the research aims to provide a single demonstrable unit that highlights the location of the musical instruments in a video segment, along with their corresponding annotations in the form of bounding boxes. The research is extended to 360 video frames as they are an essential format of a typical virtual reality experience. An input 360 video is split into its constituent audio-visual components, the former being .wav files used as inputs to the sound classifier model, and the latter being the visual image frames being used as input to the object detection framework that outputs the annotation and localization information. An all-inclusive demonstrable unit showcases both the models' functionality and can be used on two-dimensional video frames as well.