The paper introduces a method to integrate contextual information of different elements to understand the scene. It needs to combine three aspects, object detectors, 3D real world scene and camera position/orientation. Human naturally captures an object with the surrounding information, i.e., an object usually appears in particular scene. For example, the car is always on the road. In the scene of road, car might have higher probability to be detected. Using 2D image can infer a 3D real word projection by adding additional dimension depth. From the object height and camera position and orientation, we can estimate the depth. An object candidate is dependent on a background and a bounding box. A scene usually contains multiple objects, which requires computing the likelihood of each bounding box. The assumption is that the local information can determine the likelihood of object, background and bounding box. It also indicates that the object detection can be variant upon the noise. Surface is inferred into three classes which are ground, sky and vertical. A scene is conditioned on multiple image evidence. Viewpoint is learnt by training data. The model achieves average accuracy of 0.423 with small number of training data. The model has potential to perform better to user NN for object detection. Also the object and surface inference between contextual elements could be modeled as HMM or some other probabilistic models.