The paper introduces a approach that enables robot to learn actions from watching unconstrained videos. The idea is to translate the visual video information into semantic commands for robot through learning algorithm. CNNs are used for grasp recognition and object recognition. The inputs for grasp recognition are subsampled by the patches around the hand. Action is learned by a trained language model which predicts the most likely verb with particular objects. However, the location of the objects are not captured in this paper. It brings the risk that the recognized objects are not interacted with the action or the hand is not associated the objects. We define action with associated objects but with associated objects the action might not be recognized. The parsing tree is generated by context free grammar basing on linguistic model. The sub rules in the CFG are equally weighted with the same probability distribution. Action is predicted by selecting the most likely path on the parsing tree. The approach has good accuracy for object and grasp recognition because of the power of CNN model. The commands generation has 68% accuracy comparing to the ground truth. Overall the approach is simple and straightforward. However, the performance is highly dependent on the linguistic model and CFG. Those needs to be handcrafted so the approaches might not be able to scale to larger project. Another observation is that the grasp classes are shrunk into 6 which also makes the problem easier.