The paper presents a hierarchical model to learn activity in egocentric camera. Activity is a concatenation of multiple actions. Actions is defined by hands and objects appeared in the video frame. The classifier is learned from the weakly hand-labeled data. Different feature representations are used for interacting with hand in the model. Performance is evaluated with different action classes. A sequence of actions will define a activity which is learned by Adaboost algorithm. Also the paper shows the actions can be learned by a given activity. As activity is fixed, action labels are assigned to maximize the classification score for a sequence of activities. The objects are learned by a conditional probability model with given action labels. Each action in the training data is associated with some semantic words to represents objects. Objects are collected by bag-of-wards representation. However, in the paper, the correlation between objects are not explored. Even the positive bags occur in the frame it's not necessary that they are bond to the same action. It might be a source of inaccuracy. The results show that concatenating multiple forms of representations outperforms the SIFT or STIP bags. Overall the method is simple and straightforward.