The action can be learned through first person video data. Global and local motion formation can help robot to learn and recognize actions. The global feature is captured by using dense optical flows. It produces a displacement vector for every pixel. The global motion descriptor is computed by clustering optical flow into multiple categories. Bag-of-word is used to represent the global motion descriptor. Local motion descriptor is captured by spatial-temporal cuboid by summarizing gradient values. Both descriptors are clustered by k-mean into histogram representations. Multi-channel kernels are used to compute the distance between two vectors and they are fed into SVM. Combing global and local descriptors produces better accuracy than apply them individually. A video can be parsed into sub sequence which contains particular activities.It can produces the atomic level sub sequence which indicates the duration and type of a action. The hierarchical structure can also be learned. Using this structure matching shows the better performance than others approaches. By using structure matching, we indicate that both the motion information are captured and the sequence of actions are learned, which is a high-level interpretation. However, one challenge is that the complexity of video data can affect the performance the approach. If a lot of objects or actions are interacted in the video data, motion descriptors might not work well. As complexity increase, the possible paring structure can increase exponentially. Using greedy approaches can sacrifice a lot of accuracy in that situation.