The paper deals with initial hole search in a Peg-in-hole task, where only contact force data are available. For this task, authors applied Actor-Critic Fitted Policy Iteration algorithm, where the sample episodes are provided by human demonstrations. A search process is initiated by constructing probability density function of the goal position described with a belief space vector. From these representations, two policies are learned. The first maps belied space to robot linear velocity using Fitted Actor Critic RL while the second maps wrench policy to the angular velocities. Both policies, linear and angular velocities are parametrized with Gaussian Mixture Model framework. Controller additionally modulates velocities when a threshold force is reached in order to react properly on obstacles (such as socket edges).
The approach was validated on searching for the three types of power sockets with KUKA LWR arm equipped with force sensor. Initial socket search demonstrations were provided by blindfolded demonstrators. Velocities and wrenches were captured during the demonstrations. The proposed Q-EM policy was compared to the simplistic GMM policy. Results show, that the proposed approach always outperforms other policies, including human search.
The paper addresses an important area in robotics - learning of a policy from wrench data. The methods applied seems to be adequate and novel enough. Authors provided comprehensive overview of the related work. However, the paper is not easy to follow and requires a lot of previous knowledge in order to understand it.
I suggest rewriting the paper in a sense to initially describe the overall procedure, which will enable to follow the paper also to the readers with less profound knowledge in RL of POMDP.

Minor remarks:
1) authors use the same notation (x and y) for position vector and binary feature vector. The same notation was latter used to denote Cartesian coordinate axis.
2) Figure 2: Likelihood edge contact is not clear to me. Additional explanation would be beneficial.
3) page 4: Authors referee to the Figure 4 bottom, while only Figure 4 top exists.

4) page 6: “Maximization EM step, see Figure 2, is obtained.” Should be Figure 5 instead??
5) page 7: Figure 7: The scale of the value function in the range of 0..100, while in the text is in the range of 0..1.
6) page 8: Eq. (10) is ambiguous, has the same value (\dot{x}) on both sides. Please rewrite it.

7) Experiments considered only position velocity learning. I suggest including also rotational velocity learning in experiments or give an explanation, why this was omitted.