Major comments:

- As a general comment, I have to say that I really liked the idea and approach that the authors describe in this manuscript. The aspect of collecting demonstrations of human search strategy is quite interesting and may be greatly exploited in the field of robot learning. Equally important, the task chosen by the authors is challenging because haptic data is highly noisy and sometimes ambiguous, therefore, when a robot is expected to rely on force/torque (and proprioceptive) data to carry out a task, the problem is significantly harder. The results nicely showed how the proposed learning framework represents an improvement with respect to classical GMM and greedy search algorithms.

- One of my main concerns about this paper is the way it is written. Unfortunately I have to say that the reading flow is not smooth and therefore it is quite hard to follow the details and idea of the paper. The writing is lousy and several mistakes were found (some of them are detailed at the end of this review). This, of course, significantly reduces the quality of this work and makes the material presented in the paper less accessible to a broader audience. So, my first suggestion is to carefully review the writing/notations/concepts of the paper so that the text fulfills the high-standard requirements of the journal to which this paper is submitted to.

- Section Related Work:

POMDP is not introduced in the paper. Despite it is a quite known model, it should be briefly introduced in the paper for completeness.

- Section Experiment methods:

+ What is the motivation of using differentiation entropy to compress the belief PDF? What is the intuition behind this? For example, why not using latent representations if the aim is reducing the dimensionality of the space in which the search is carried out?

+ Authors wrote: "See Figure 4 for the time taken to connect the plug to the socket." However such information is not provided in this Figure. The authors need to add this to the Figure.

- Section Learning Actor & Critic

+ I wonder why two different policies were encoded for the same task. It is true that

+ The authors say that "Both the linear and angular velocity policies are parameterised by a Gaussian Mixture Model (GMM)" However, the way how the GMM parameters are defined later describes means and covariance matrices encoding linear velocities and belief space vectors 'b'. This is really confusing because it is totally unclear if angular velocities are included in \dot{x} (according to the notation this should not be the case), or if the authors made a mistake here and they are encoding only linear velocities and belief space vectors 'b'.

+ I found that Algorithm 1 does not really add something new or different to the description of the Fitter Policy Iteration method, neither facilitates the understanding of this technique. However, I encourage the authors to use this Algorithm to show how the whole learning process is carried out (1. Demonstrations are collected. 2. Belief PDF is approximated by PMF. 2.1. Differentiation entropy is used to compress the belied PDF, etc...)

+ Please explicitly indicates that Q is a function that represents the logarithmic lower bound. This is not indicated in the paper.

- Section Results:

+ Good set of experiments. However, it may be useful to mention how other policy search algorithms may perform here. My point is that greedy search is used as a sort of baseline for comparison purposes. However, as part of the contributions of this paper is about Q-EM, it would be more interesting to see how other policy search methods behave in the same scenarios.

+ I wonder if it is really fair to compare an algorithm that does not include any information about the goal of the task (GMM) while the other has additional information to work with (Q-EM)? In other words, the training algorithm of Q-EM does include information about the task goal, which is not given to the GMM (I might miss something here if this is not the case), therefore, in my opinion, as soon as Q-EM converges, it will provide better results that GMM (even better if the demonstrations are not describing the task goal in a good way).

- Section Discussion & Conclusion:

This looks more like a summary of the paper than a discussion. Nothing important or relevant is added here!

Minor comments:
- Introduction: "...structure of the uncertainty, it uncertainty is not ..." What did you mean here?

- Related work: "force torque" -> "force/torque"?
- Learning Actor & Critic: "It is the expected future reward given the current belief state and policy" Please rephrase this because despite I know you are referring to the value function defined in Eq. (1), the way how it is currently written is completely ambiguous for an inexperienced reader.

- Learning Actor & Critic: " To learn the value function we take Fitted RL [14] approach is taken." -> "To learn the value function, the Fitted RL [14] approach is chosen."

- Learning Actor & Critic: "...Maximisation EM step, see Figure 2, is obtained" -> "...Maximisation EM step, see Figure 5, is obtained"

- Learning Actor & Critic: "When Q-EM is applied the Gaussian functions of the GMM will favour these locations" Please rewrite this sentence.

- Reference 26 and 27 are the same!