The main contributions of this paper are:

1) the introduction in the learning by imitation framework of a method to track complex belief state of the performer

2) the Q-Em algorithm used to fit the policy in the belief space

I think that the contribution is good and valuable but needs refinement in the presentation.

main issues:

What does it mean “Both groups A and B took 9 ± 10s to find the socket’s edge, regardless of the socket type. This is to be expected since the sockets are at the same location.” what is the uncertainty considered/tackled by the humans then?

How are the two policies coordinated? “Two policies are learned: one to map belief space to linear velocity πθ1 : bt → x^{.}_t and other to map sensed wrench to an- gular velocity, πθ2 : φt → ωt.”
(in section 3.1 the discussion is only about x^{.}_t)

secondary issues:

Autonomy of the approach is limited because the definition of the environment and of the task is given more than to autonomy I would just relate the main contribution to imitation in partially observable domain or in search
in most of the imitation literature in robotics the state space is given and there is no partial observability. representing the uncertainty of the performer is one of the main teacher of the paper compared to other classical approaches to POMDP and mapping in robotics the agent needs just to focus on a specific part of the environment (the socket)... thus using an ad hoc representation strongly simplifies the learning task (see for a similar search task acquiring a representation strongly different from the real one:

Ecological Active Vision: Four Bio-Inspired Principles to Integrate Bottom-Up and Adaptive Top-Down Attention Tested With a Simple Camera-Arm Robot
D Ognibene, G Baldassarre
Autonomous Mental Development, IEEE Transactions on
2015)

in the abstract there is no need to go into the detail of how usually people solve this problems with pomdp and dynamic programming

Reorganise the description of the model in the abstract as following:

"A group of human teachers demonstrate the PiH- search task whilst blindfolded. The position uncertainty, represented by a Point Mass Filter (PMF), is recorded and compressed to the most likely state and entropy. A belief space value function is learned offline in a Fitted RL framework and it is used to update the parameters of a Gaussian Mixture Model (GMM) policy.”
->
“The critic learns offline in a Fitted RL framework A belief space value function using demonstration from a group of blindfolded human teachers. The demonstrations are compressed as sequences of the most likely state and entropy extracted by a Point Mass Filter (PMF). The critic is then used to train the actor policy represented as Gaussian Mixture Model (GMM)”

why Q-EM ?

“Greedy and non-optimised GMM policy for the distance taken to localise the socket and then establish a connection” this is not clear… connection? too much detail for an abstract

I would just write in a more standard and compact form, after explaning the name of the algorithm more clearly:
"The GMM Actor-Critic policy, called Q-EM, is compared with both a Greedy and non-optimised GMM policy for the distance taken to localise the socket and then establish a connection. The ability of the learned models to generalise to different socket types and locations is evaluated. The results show that the Actor-Critic policy is always more performant in terms of distance travelled to localise the socket. For the same task, the Q-EM, GMM and Greedy policies were tested on the KUKA LWR robot for three different power sockets. The results show that when the socket has no distinctive features both data driven policies perform better than the Greedy.”
->
“Tests performed on the KUKA LWR robot showed that the proposed algorithm outperforms other (standard/state of the art) approaches in terms of distance travelled to localise the socket and that together with other data driven (?) methdos it perform better than alternatives when the socket has no distinctive features.”

page 1
“Depending on the task and structure of the uncertainty, it” -> if
maybe need to specify state uncertainty..

"It is assumed that a good mixture of explorative-exploitative behaviour is present” + “in the dataset”

after “The second is that PiH is a very important component in manufacturing processes and we seek to demonstrate that this task can be accomplished without the need of a costly vision system.”
add
“It is also the simplest setting where search is different from extensive exploration, e.g. in slam style, of the environment and where not strictly task related features are not relevant, providing little information, and sparse. So the applied methodologies can be likely transferred to other similar tasks like seek and rescues or cibarsi with limited effort”

citation 38 is not complete

it is not clear how the following related work tackles the partial observability issue, what would happen if the hole position is changed and is unknown? that’s the main contribution of the paper, if this papers don’t do that it should be clearly stated:
"Another approach consists of learning task space policies and gradually adapting the parameters based on a reference Force/Torque profile. In [38] the authors learned a time- dependent Dynamic Movement Primitive (DMP) [33] Cartesian end-effector policy for the Cranfield benchmark object from hu- man teleoperated demonstrations. Similarly in [25, 1], a F/T profile is encoded separately by a regressor function along the DMP policy. Successive refinements of the DMP policy are achieved through using force feedback to adapt the parameters of an admittance controller such so as to reproduce the same F/T (encoded by a separate regressor).

Reproducing exactly the same force torque profile for the complete trajectory could be unnecessary as the force torque profile is used predominantly during the final stage of the PiH task, to avoid jamming during insertion [20, Chap. 5].
Reinforcement learning has also been applied to PiH. In [19] a DMP policy is initialised with kinesthetic demonstrations of picking up a pen. The recorded Cartesian trajectories are en- coded in a parameterised DMP policy and augmented with a F/T profile. After 110 trials the policy was found to be a 100% successful. In [17] a 18 dimensional input and a 6 dimensional output (linear

and angular velocity) neural network is learned and successful after a 100 episodes. This work has a similar approach, however instead of considering autonomous rollouts common in RL, relies solely on the data provided by human teachers.”

This point can be clarified introducing earlier the problem of diverging value function. Also note that online RL wihtout batch updates can easily lead to suboptimal behaviours, while genetic/evolutionary/otpimisation approaches don’t. Batch RL can be shown to fall in the same class I think. But actually once provide the belief space the partial observability disappears (e.g.https://www.jair.org/media/613/live-613-1809-jair.pdf)

“By remembering all the state transition pairs and by applying multiple synchronous Dynamic Programming (DP) and function approximation updates, the problem of diverging value function approximators is resolved."

“The human’s location INITIAL belief is represented by a probability density function (pdf) which is assumed to be uniformly dis- tributed in the orange area.”

end of page 3

“differentiation entropy”->differential entropy

“maximum a posteriori”-> mean or median or maximum peak

incomplete statement, connect to previous, break paragraph after

"It is the expected future reward given the current belief state and policy.”

not clear:

We define the actual measurement to be a function of the raw wrench, y t̃ = h(φt), which is a binary feature vector

I would move section 3.2 before 3.1 to keep all definitions and math together so maybe unify chapter 3 and 4

this looks like a strong assumption and it is not clear if the belief is on the position of the arm or on the position of the target and how the two would be related, becuase the rsulting belief state would have a complex shape with holes for the points visited by the arm :

"All subsequent beliefs can be in- ferred from the measured velocity and measurements provided by the ATI and OptiTrack sensors.”

The belief management in this paper is not trivial given the complexity of the areas where the robot did not explore and how this should be matched with the training samples. A future work would be to compare this filter with other belief state management methods.

The use of fited iterated value function is not exactly the same of the iterated q function, for which the convergence proof is available in the refence 14. Still the convergence should be trivial considering a one action mdp.

page 4 Policy improvement section would strongly benefit from the motivation of the design choice. what are the properties of the modified algorithm? convergence? has it been presented in other places? this looks like the main novelty of the paper but I would follow the usual onion style presentation (motivation-choice-more detailed motivation-more details on system… fig6 looks very interesting but it may need additional considerations: what are the necessary conditions for this approach the selected policy is far from the demonstration. maybe the predcition would not be valid, e.g if there is another obstacle..

My suggestion would be to add a small section in the introduction, after fitted policy iteration part, on the Monte-Carlo EM-based policy search approach and to highlight the difference and advantage in this section. I would add the main equations of the original methods too instead of only presenting an extensive literature review of quite unrelate works (e.g. DFQ and atari). I would also decrease the details on the participants and similar, this is not a medicine paper.

for fiugre 7 change Middle-right to Middle and right

remove space from “is conditioned on the state space“