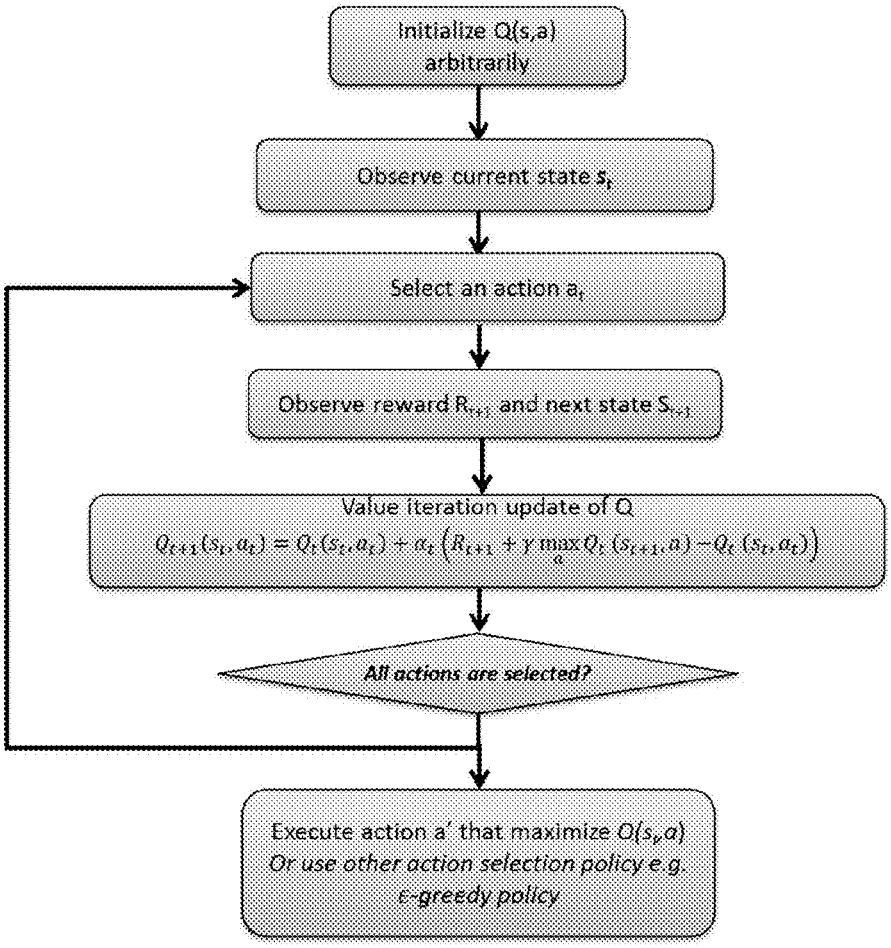
Initialize Q(s,a) arbitrarily
Observe current state s_t
Select an action a_t
Observe reward R_{t+1} and next state S_{t+1}
Value iteration update of Q
$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t \left(R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)\right)$
All actions are selected?
Execute action a' that maximize $Q(s_t, a)$
Or use other action selection policy e.g. ε-greedy policy

FIG. 6