PIR : Mixed-Initiative Mission

Thomas Vagneron

Summary

- Context
- Theory
- Experiment
- Conclusion

Context and Goal

Human-machine systems

Increasing use of automated systems

aircrafts, cars or even games player (A fresh example is AlphaGo which is a Go playing machine).

Context and Goal

Human-machine systems

Increasing use of automated systems

aircrafts, cars or even games player (A fresh example is AlphaGo which is a Go playing machine).

- **Increasingly autonomous robots:** technical advances in AI, machine learning.

Context and Goal

Human-machine systems

Increasing use of automated systems

aircrafts, cars or even games player (A fresh example is AlphaGo which is a Go playing machine).

- **Increasingly autonomous robots:**
  technical advances in AI, machine learning.
- **Human operator still vital:**
  - judiciary responsible
  - able of creativity or improvisation.

Context and Goal

Human-machine systems

Increasing use of automated systems

aircrafts, cars or even games player (A fresh example is AlphaGo which is a Go playing machine).

- **Increasingly autonomous robots:**
  technical advances in AI, machine learning.
- **Human operator still vital:**
  - judiciary responsible
  - able of creativity or improvisation.

Human factors involved in 80% of AAVs accidents! [Wil]

Context and Goal

Human operator weaknesses

Potential effects of a mission on human operators:

- stress (danger, pressure)
- workload (multi-task, hard tasks)
- fatigue, boredom (long mission)

Context and Goal

Human operator weaknesses

Potential effects of a mission on human operators:

- stress (danger, pressure)
- workload (multi-task, hard tasks)
- fatigue, boredom (long mission)

Consequences:

- mental confusion
- attentional tunneling
- mind wandering
- lower vigilance
- ...

increase in accident risk resulting in mission fails

Context and Goal

use of human state feedbacks!

beamer-onera-head

Human operators equipped with sensors

data from the human operator state can

refine supervision of human-robot team

Context and Goal

use of human state feedbacks!

Human operators equipped with sensors

data from the human operator state can
refine supervision of human-robot team

- [SCD15] target identification task (ground robot)
  - **devices:** eye tracking + electrocardiography
  - **human state:** *cognitive availability* estimation
  - **superv. validation:** simulations of the system (including human behavior)

Context and Goal

use of human state feedbacks!

Human operators equipped with sensors

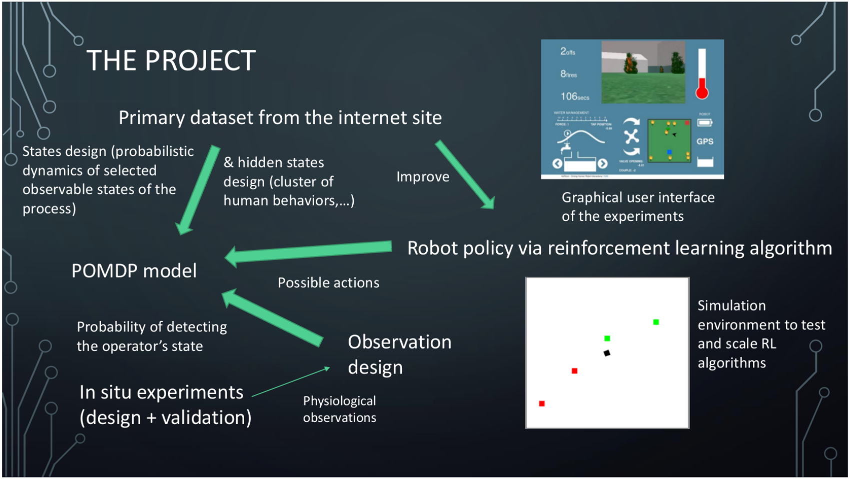
data from the human operator state can
refine supervision of human-robot team

- [SCD15] target identification task (ground robot)
  - **devices:** eye tracking + electrocardiography
  - **human state:** *cognitive availability* estimation
  - **superv. validation:** simulations of the system (including human behavior)

- search and rescue task (AAVs) [GCLD16]
  - **device:** eye tracking
  - **human state:** *current human task* = human gaze
  - **superv. validation:** tested on 10 volunteers

Context and Goal

The Project



Theory

History

Domain's birth : late 70's by the encounter of

- experimental psychology
- computational neuroscience
- dynamic programming

Theory
History

Domain's birth : late 70's by the encounter of

- experimental psychology
- computational neuroscience
- dynamic programming

An approach to experimental psychology : Law of effects (Thorndike, 1911)

" The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond".

Theory

Reinforcement Learning

What is Reinforcement Learning?

It is an approach, here computational, to decision-making, understanding and goal-directed learning.

Theory

Reinforcement Learning

What is Reinforcement Learning?

It is an approach, here computational, to decision-making, understanding and goal-directed learning.

What does it needs? The definition of the interactions between a learning agent and its environment.

- the states
- the actions
- the rewards

Theory

Reinforcement Learning

What is Reinforcement Learning?

It is an approach, here computational, to decision-making, understanding and goal-directed learning.

What does it needs? The definition of the interactions between a learning agent and its environment.

- the states
- the actions
- the rewards

RL objectives :

the automatized acquisition of skills for decision making in a complex and uncertain environment and learning by "experience" a behavioral strategy (named policy) reflecting the failures and success (reinforcements or rewards).

Theory

Markov Property and Markov Chain

Markov property

all the useful information for the future prediction is in the present state.

Theory

Markov Property and Markov Chain

Markov property

all the useful information for the future prediction is in the present state.

Markov Chain

a dynamic system with discrete time $(x_t)_{t \in \mathbb{N}} \in X$, where X is space of states such as :

$$\mathbb{P}(x_{t+1} = x | x_t, x_{t-1}, \dots, x_0) = \mathbb{P}(x_{t+1} = x | x_t).$$

Markov Decision Process (MDP)

defined by (X, A, p, r), where :

- X space of states
- A space of actions
- $p(y|x, a)$: probability of transition from a state $x \in X$ to $y \in X$ when the action $a \in A$ is chosen:

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y|x_t = x, a_t = a),$$

- $r(x, a, y)$: reward when passing from x to y using the action a.

Theory

Value function and Bellman operator

Value function

The value function are defined over the state space :
for $x \in X$, $V(x)$ is the mean of the sum of the rewards over time for a process starting in x.

$$V^{\pi}(x) = \mathbb{E}[\sum_{t \geqslant 0} r(x_t, \pi(x_t))]$$

Theory

Value function and Bellman operator

Value function

The value function are defined over the state space :
for $x \in X$, $V(x)$ is the mean of the sum of the rewards over time for a process starting in x.

$$V^{\pi}(x) = \mathbb{E}[\sum_{t \geqslant 0} r(x_t, \pi(x_t))]$$

Bellman operator

Written T^{π}, it is defined by :

$$T^{\pi} = V^{\pi}(x) = r(x, \pi(x)) + \sum_{x'} p(x'|x, a) V(x')$$

Theory

Some solving methods

- **Direct solving** of the linear system $(I - \gamma P^{\pi})V^{\pi} = r^{\pi}$. That's the Gauss elimination method, but it have a complexity of $O(N^3)$

Theory

Some solving methods

- **Direct solving** of the linear system $(I - \gamma P^{\pi})V^{\pi} = r^{\pi}$. That's the Gauss elimination method, but it have a complexity of $O(N^3)$
- **Iteration on the values for a permanent policy** : we iterate on the operator T^{π} (Bellman operator). Considering a given value function V_0, $V_{n+1} = T^{\pi} V_n$. We have then convergence of V_n toward V^{π}. The problem is that the convergence is asymptotical, but the advantage is that it as a lower complexity than the direct solving method ($O(N^2 \frac{log(1/\epsilon)}{log(1/\gamma)})$ for an approximation of ϵ (interesting if γ is not too close from 1)).

- **Monte-Carlo** : we simulate n trajectories $((x_t^i)_{t \geqslant 0})_{1 \leqslant i \leqslant n}$, starting from x and following the policy π : $x_{t+1}^i \; p(.|x_t^i, \pi(x_t^i))$, so :

$$V^{\pi}(x) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{t \geqslant 0} \gamma^t r(x_t^i, \pi(x_t^i)).$$

this method is interesting if we want to evaluate a unique state. It has an approximation error of order $O(1/\sqrt{n})$

- **Monte-Carlo** : we simulate n trajectories $((x_t^i)_{t\geqslant 0})_{1\leqslant i\leqslant n}$, starting from x and following the policy π : $x_{t+1}^i \ p(.|x_t^i, \pi(x_t^i))$, so :

$$V^\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geqslant 0}\gamma^t r(x_t^i, \pi(x_t^i)).$$

  this method is interesting if we want to evaluate a unique state. It has an approximation error of order $O(1/\sqrt{n})$
- **Temporal differences TD(λ)** : This is a smart method for using the trajectories to evaluate all the states crossed by those trajectories, by evaluating the value of a state x, by the sum of the observed temporal differences $r_t + \gamma V(x_{t+1}) - V(x_t)$ at the future instants t, weighted by a "trace" λ.

The impact of the temporal differences of future transitions on the estimation of the current state value is controled by λ.
TD(λ) is a compromise between :

- **TD(0)** (to estimate the fixed point of the operator of Bellman T^π):

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) d_k.$$

- **TD(1)** (to estimate the mean):

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l \geqslant k} d_l.$$

The impact of the temporal differences of future transitions on the estimation of the current state value is controled by λ.
TD(λ) is a compromise between :

- **TD(0)** (to estimate the fixed point of the operator of Bellman T^{π}):

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) d_k.$$

- **TD(1)** (to estimate the mean):

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l \geqslant k} d_l.$$

the choice of λ :

- $\lambda < 1$ allows to reduce the variance of the estimators compared with $\lambda = 1$.
- $\lambda > 0$ allows to propagate faster the rewards compared with $\lambda = 0$.

Theory

TD(λ)

the algorithm TD(λ) After the observation of a trajectory $(x_0, x_1, \ldots, x_K = 0)$, we update V_n to the states $(x_k)_0 \leqslant k < K$ following :

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{l=k}^{K-1} \lambda^{l-k} d_l,$$

$$\text{where } d_l = r^\pi(x_l) + V_n(x_{l+1}) - V_n(x_l).$$

with η_n the learning step, typically $\frac{1}{n}$.

the algorithm TD(λ) in the actuated case : We can define the actuated value function when, at each time step, the process is stopped with probability $1-\gamma$ with $0<\gamma<1$:

$$V^{\pi}(x) = \mathbb{E}[\sum_{t \geqslant 0} \gamma^t r(x_t, \pi(x_t))].$$

Algorithm TD(λ) becomes

$$V_{n+1}(x_k) = V_n(x_k) + \eta_n(x_k) \sum_{t \geqslant k} (\gamma\lambda)^{l-k} d_l.$$

Experiment

The mission

A search and fight type mission : the firefighter

Experiment

The mission

A search and fight type mission : the firefighter

The objectives : extinguish the maximum number of fire

Experiment

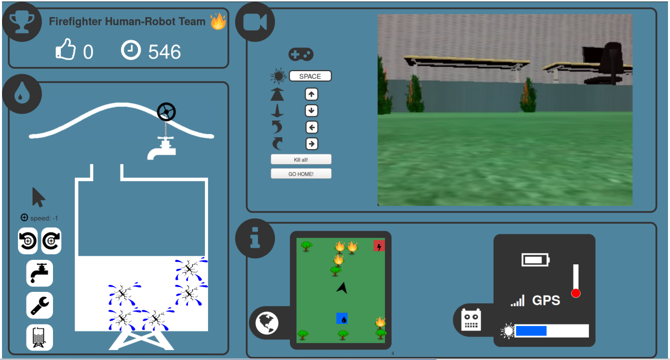
The mission

A search and fight type mission : the firefighter

The objectives : extinguish the maximum number of fire

The drawbacks :

- the battery
- the water level
- the temperature



Experiment
The environments

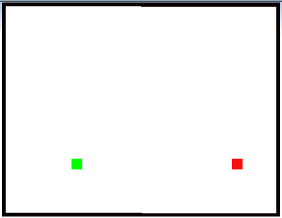


Figure : View of the simulation of the simple environment

Experiment

The environments

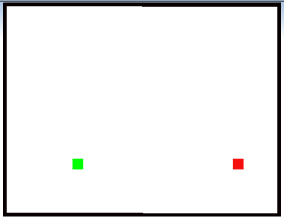


Figure : View of the simulation of the simple environment

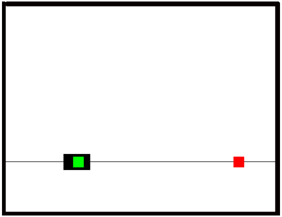


Figure : View of the simulation of the 1D environment

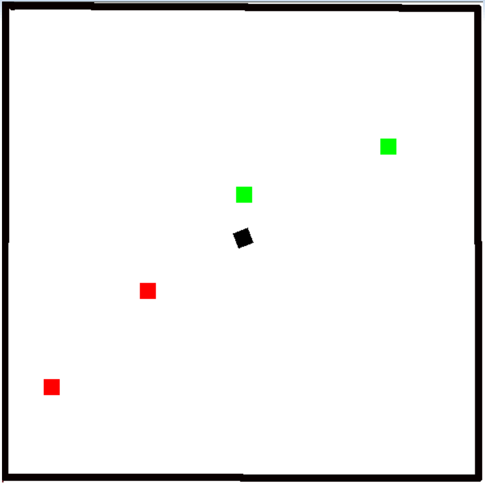


Figure : View of the simulation of the 2D environment

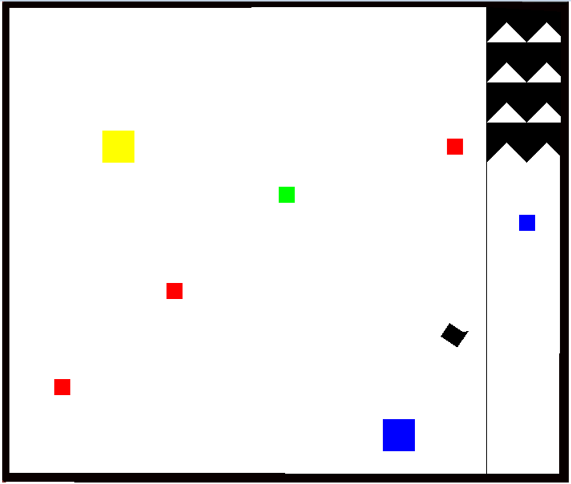


Figure : View of the simulation of the complex 2D environment

Experiment

Results

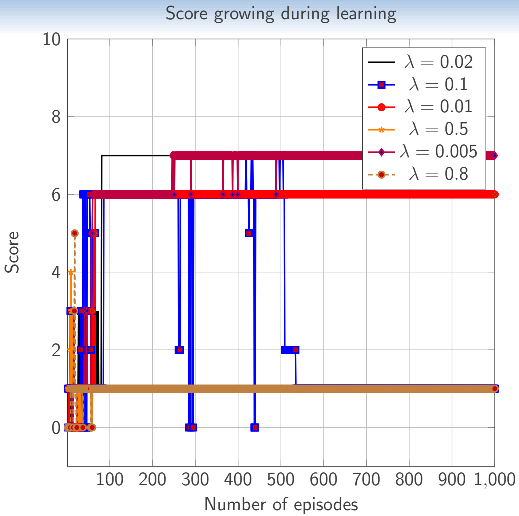


Figure : Score of the deterministic 1D environment for λ

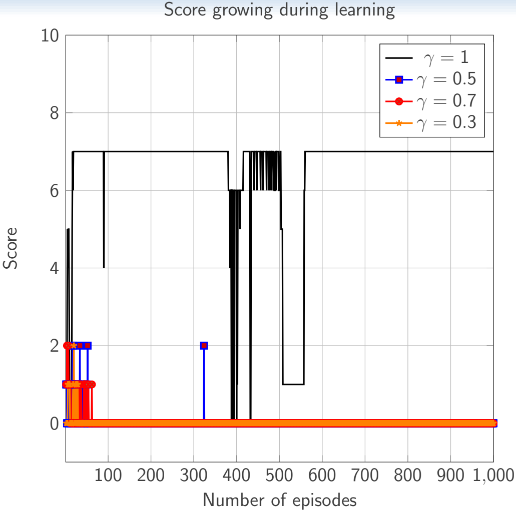


Figure : Score of the deterministic 1D environment for γ

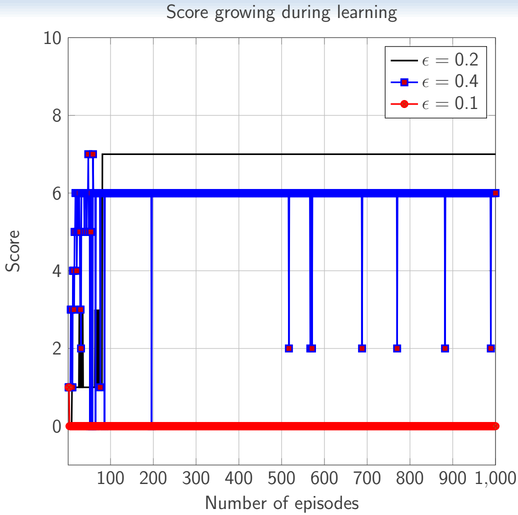


Figure : Score of the deterministic 1D environment for ϵ

Conclusion

Thibault Gateau, Caroline Ponzoni Carvalho Chanel, Mai-Huy Le, and Frédéric Dehais.
Considering human's non-deterministic behavior and his availability state when designing a collaborative human-robots system.
In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS '16, 2016.

Paulo Eduardo Ubaldino de Souza, Caroline Ponzoni Carvalho Chanel, and Frederic Dehais.
Momdp-based target search mission taking into account the human operator's cognitive state.
In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '15, 2015.

Kevin W. Williams.
A summary of unmanned aircraft accident/incident data: Human factors implications.
U.S. Department of Transportation, Federal Aviation Administration, Civil Aerospace Medical Institute.

Thank you!