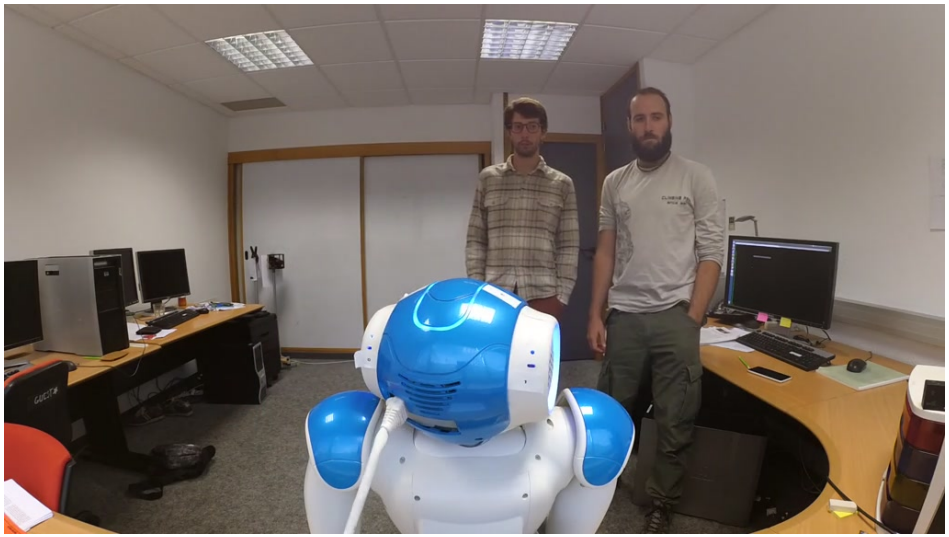
Neural Network based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction

Stéphane Lathuilière, Benoit Massé, Pablo Mesejo and Radu Horaud

Perception team, Inria Grenoble

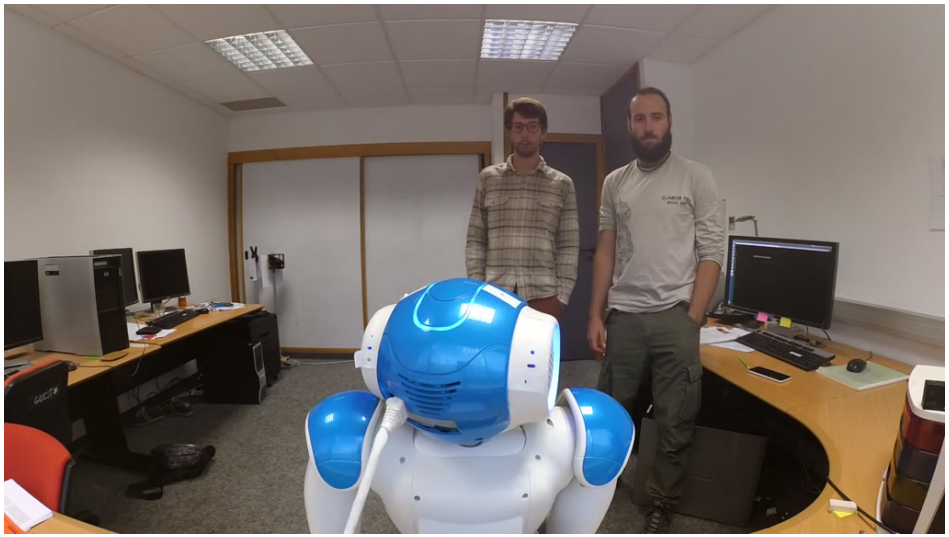
March 30th 2018

Gaze control for Human-Robot Interaction



- Where to look at?
- What are the good motor commands inorder to obtain a rich environment representation?

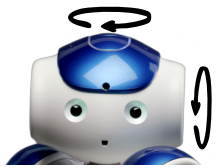
Gaze control for Human-Robot Interaction



- Where to look at?
- What are the good motor commands inorder to obtain a rich environment representation?

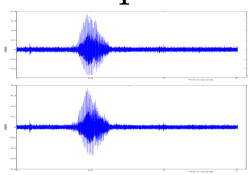
Sensors:

Motor Positions



Camera

Microphones



Gaze control: related works

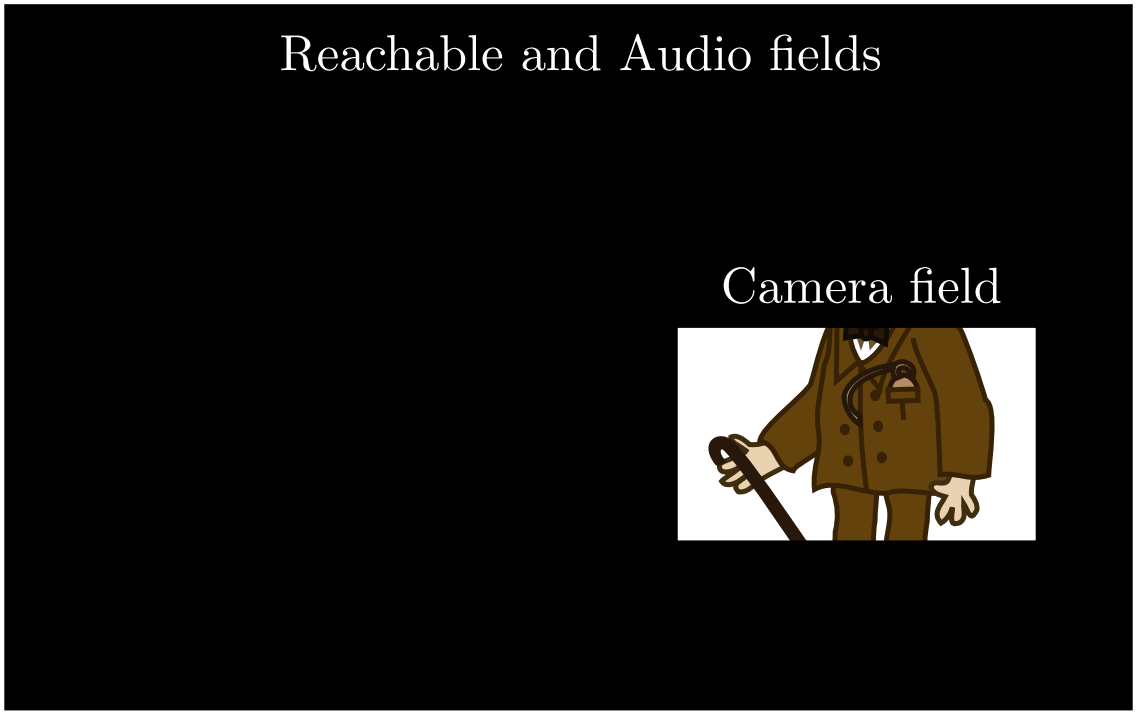
- Vast literature on visual servoing [1]: How to align an object (eg. a face) with a target image location (eg. image center)?
  - Not designed for multi-person scenarios
  - It does not use audio information
- Few works on multi-person scenarios [2,3] :
  - Handcrafted rules
  - Without properly combining audio and visual data.

[1]: Cretual and Chaumette, Application of motion-based visual servoing to target tracking, 2001

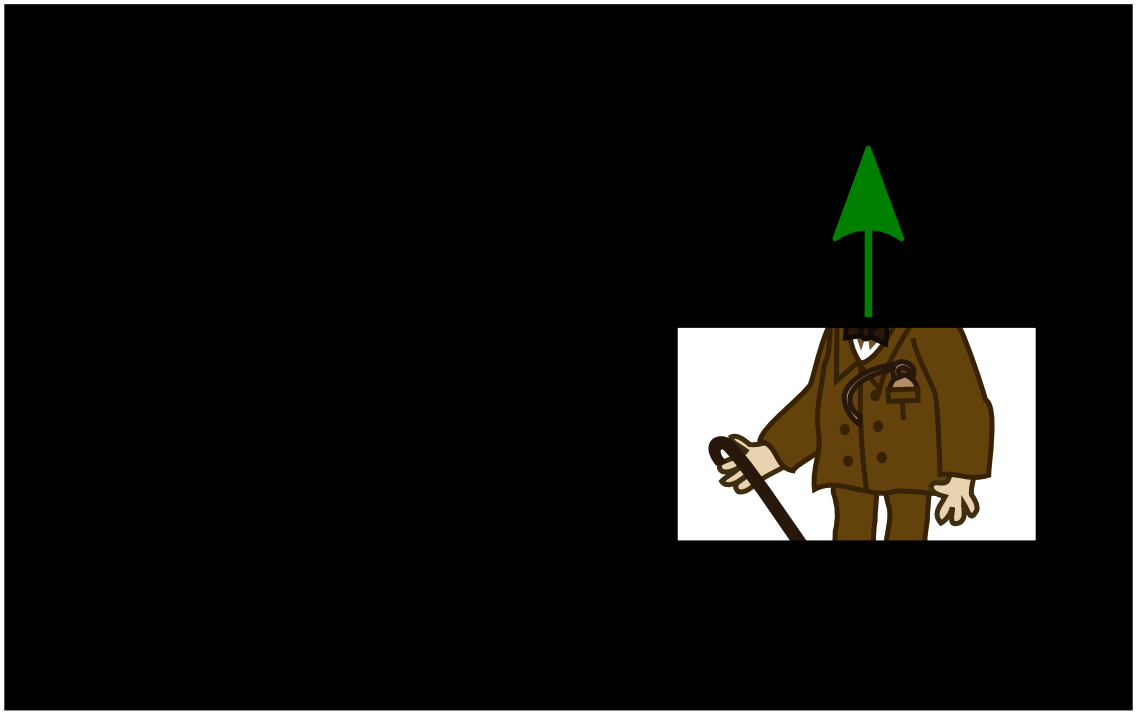
[2]: Bennewitz et al., Towards a humanoid museum guide robot that interacts with multiple persons, 2005

[3]: Ban et al., Tracking a Varying Number of People with a Visually-Controlled Robotic Head, 2017

Gaze control: a difficult problem?

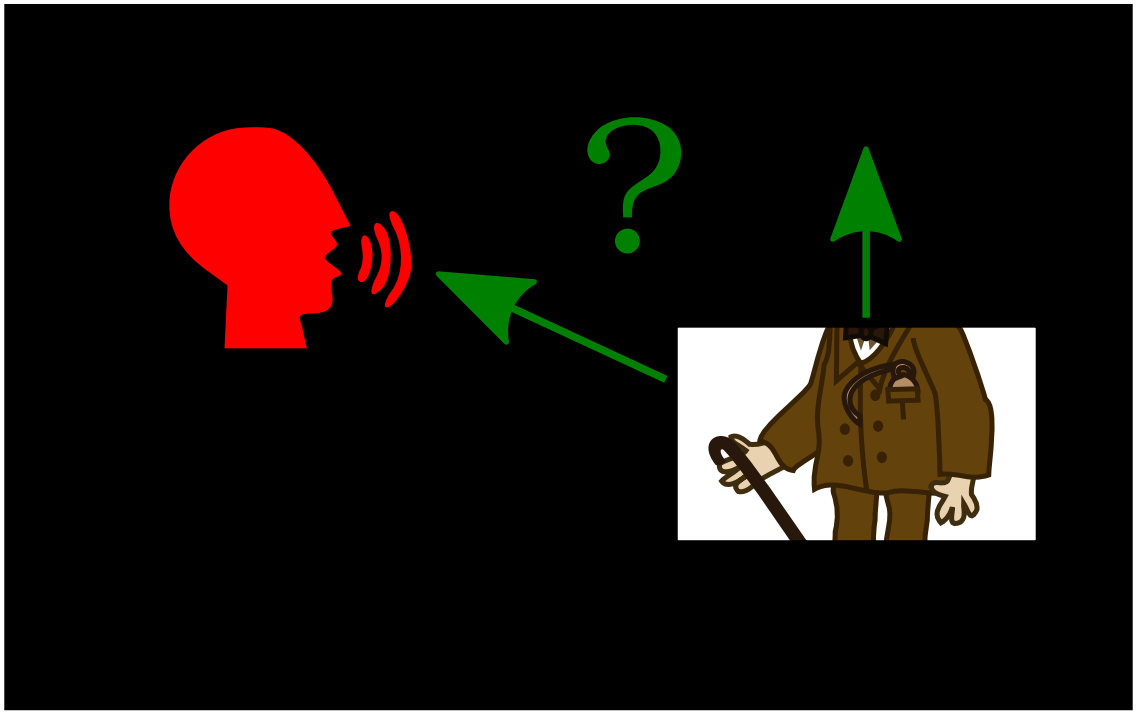


Gaze control: a difficult problem?



We need to include the body geometry in the decision process.

Gaze control: a difficult problem?



The decision depends on:

- Behavior we expect.
- The speech localization accuracy
- Motion properties of the robot.

Gaze control: a difficult problem?

Handcrafted rules: limitations

- The problem is too complex to handle all possible cases.

We need Learning!

Gaze control: a difficult problem?

Handcrafted rules: limitations

- The problem is too complex to handle all possible cases.

We need Learning!

Problem: We cannot have a training set containing observations with their associated optimal head commands. A discriminative approach would be difficult.

Approach: We evaluate each head movement by counting the number of detected faces after moving the head.

We need Reinforcement Learning (RL)!

Gaze control: our contributions

- RL formulation for the Gaze control problem.
- Audio-visual fusion framework.
- Deep RL to model the action-value function (DQN [4]) and propose architectures specifically for our task.
- Simulated environment to train our model.

[4]: Mnihn et al. Human-level control through deep reinforcement learning, 2015

Gaze control in the RL framework

Proposed reward:

$$R_t = F_{t+1} + \alpha \Sigma_{t+1},$$

with:

- $F_{t+1} \in \mathbb{N}$: number of detected faces
- $\Sigma_{t+1} \in \{0, 1\}$: $\Sigma_{t+1} = 1$ if someone is speaking within the camera field
- $\alpha \geq 0$: adjustment parameter.

Gaze control in the RL framework

Proposed reward:

$$R_t = F_{t+1} + \alpha \Sigma_{t+1},$$

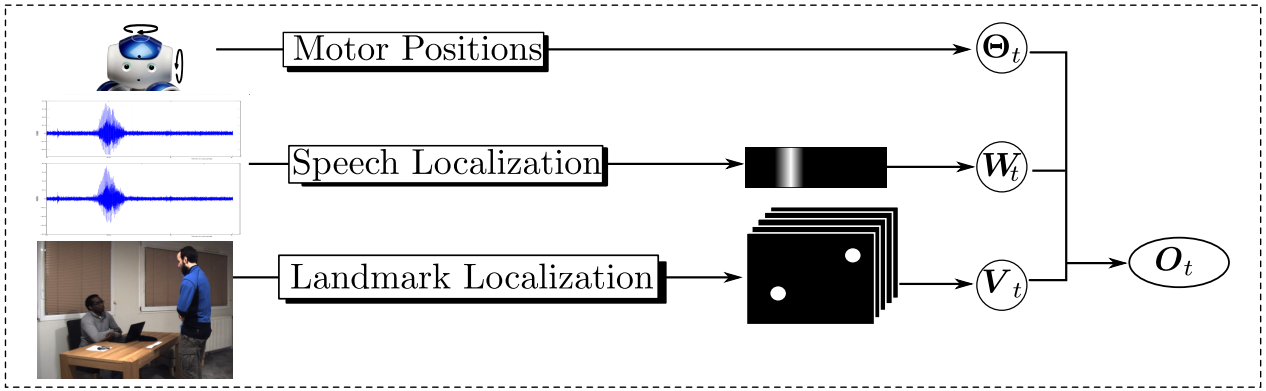
with:

- $F_{t+1} \in \mathbb{N}$: number of detected faces
- $\Sigma_{t+1} \in \{0, 1\}$: $\Sigma_{t+1} = 1$ if someone is speaking within the camera field
- $\alpha \geq 0$: adjustment parameter.

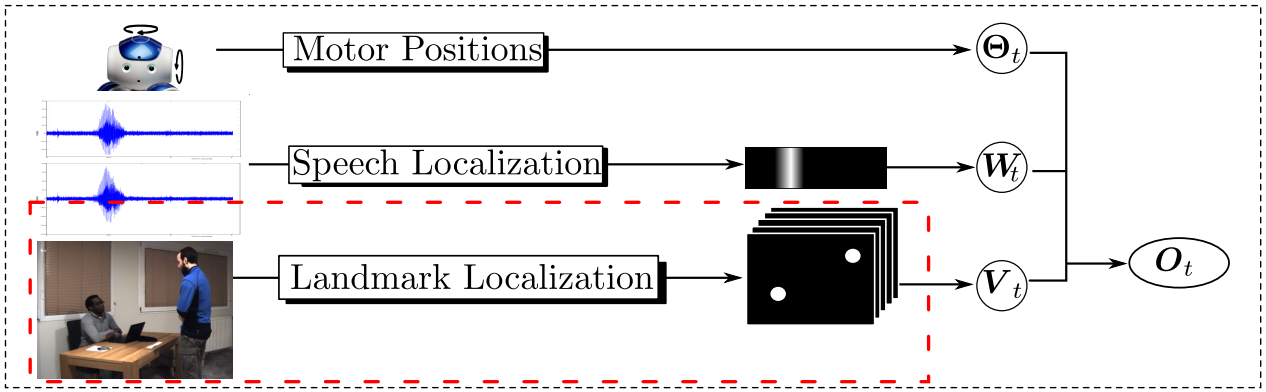
Set of actions:

$$\mathcal{A} = \{\varnothing, \leftarrow, \uparrow, \rightarrow, \downarrow\}$$

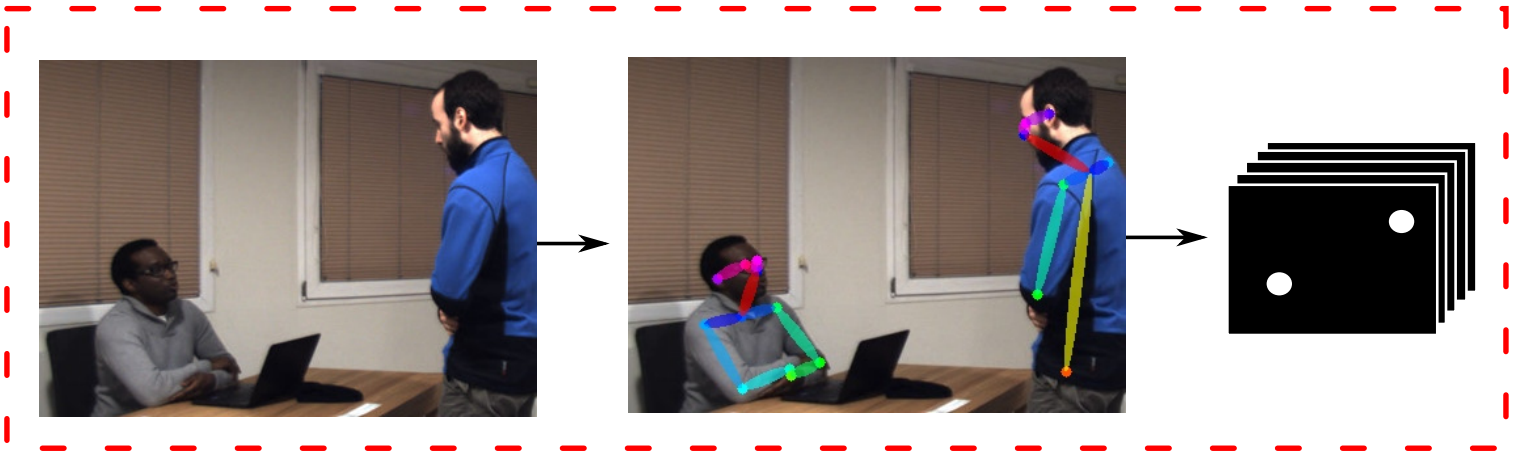
Proposed pipeline



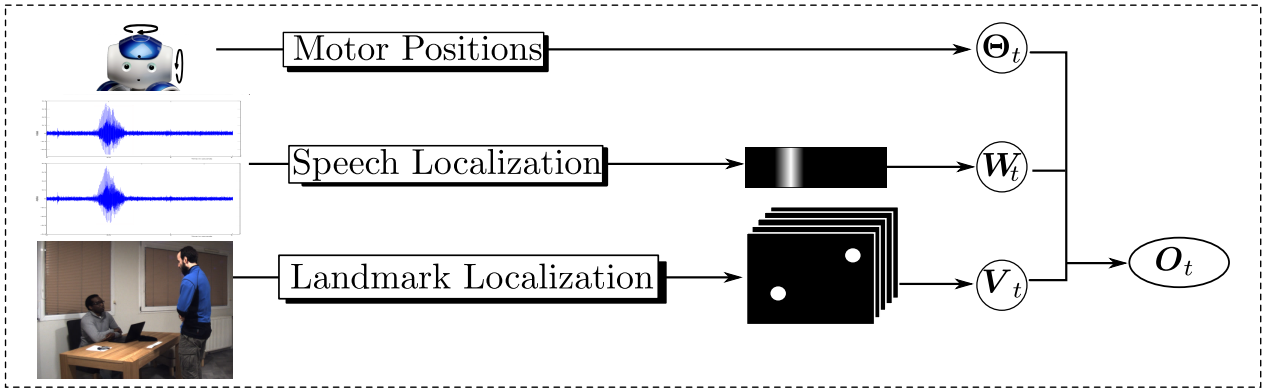
Proposed pipeline



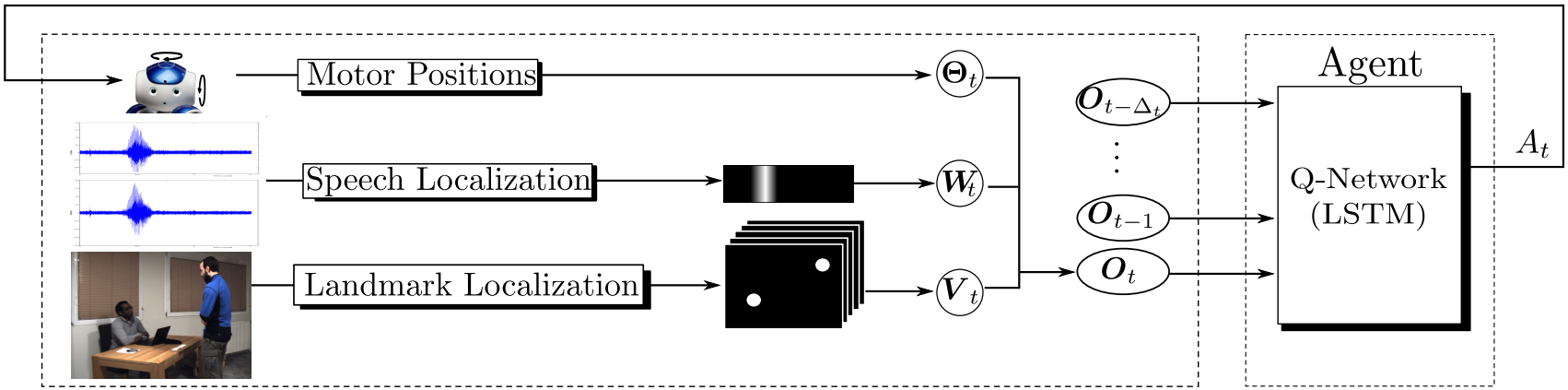
Zoom on Landmarks Localization:



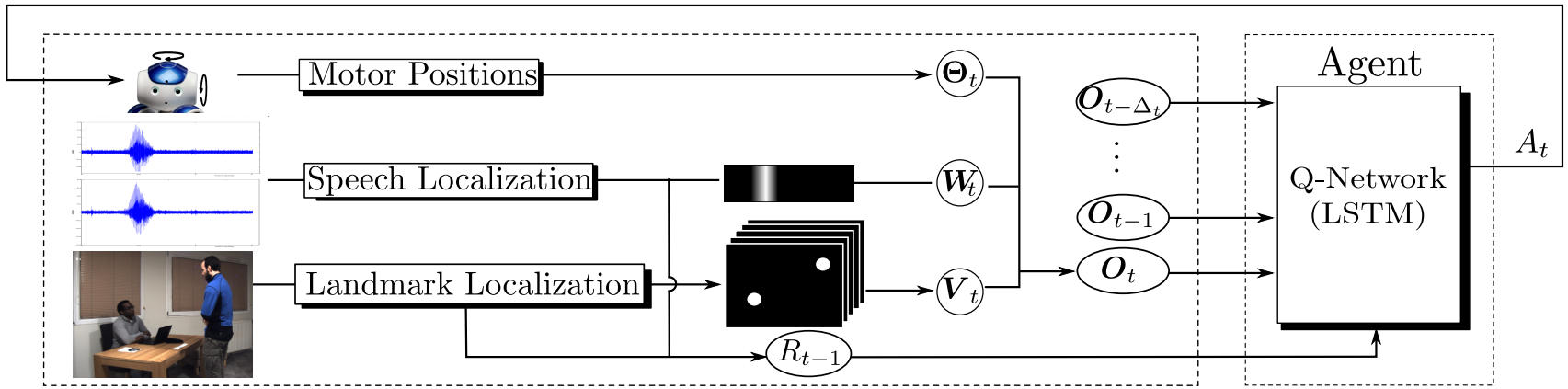
Proposed pipeline



Proposed pipeline

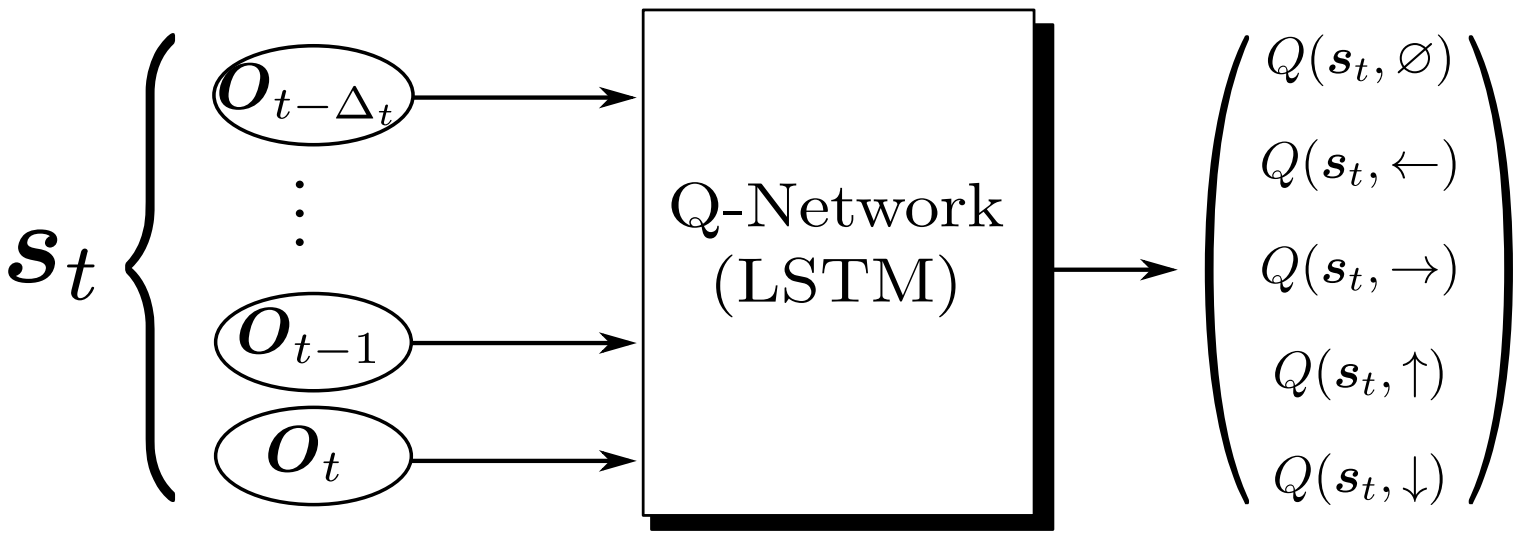


Proposed pipeline



Deep Q-Learning[4]:

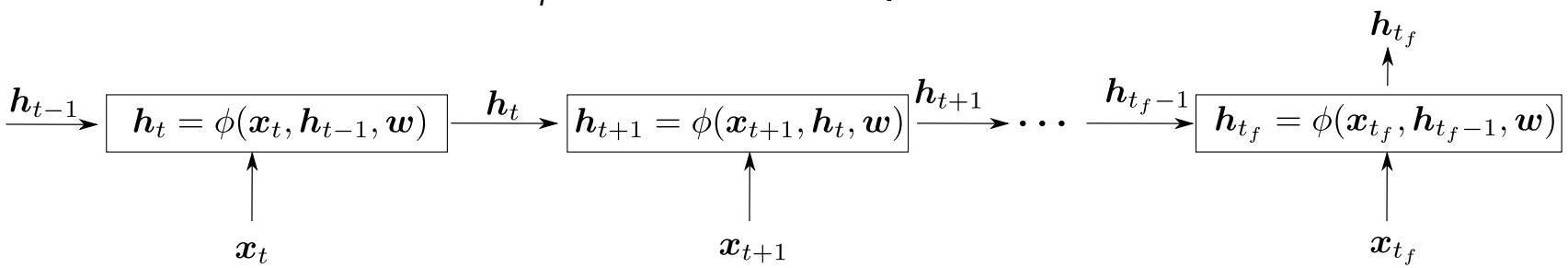
- Discounted future return: $\bar{R}_t = \sum_{\tau=0}^{T-t} \gamma^\tau R_{\tau+t}$.
- Q-function: $Q_\pi(\boldsymbol{s}_t, a_t) = \mathbb{E}[\bar{R}_t | \boldsymbol{S}_t = \boldsymbol{s}_t, A_t = a_t]$.



[4]: Mnihn et al. Human-level control through deep reinforcement learning, 2015

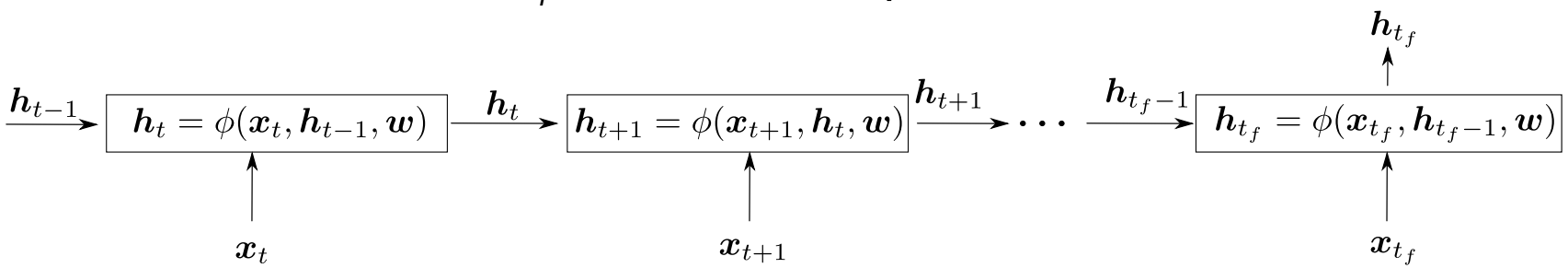
Recurrent Neural Network

Given $x_1 \ldots x_t \ldots x_{t_f} \in \mathbb{R}^{D \times f}$ a sequence of observations:

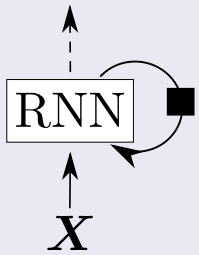


Recurrent Neural Network

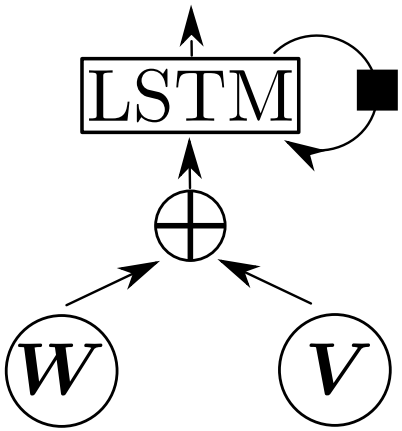
Given $x_1 \ldots x_t \ldots x_{t_f} \in \mathbb{R}^{D \times f}$ a sequence of observations:



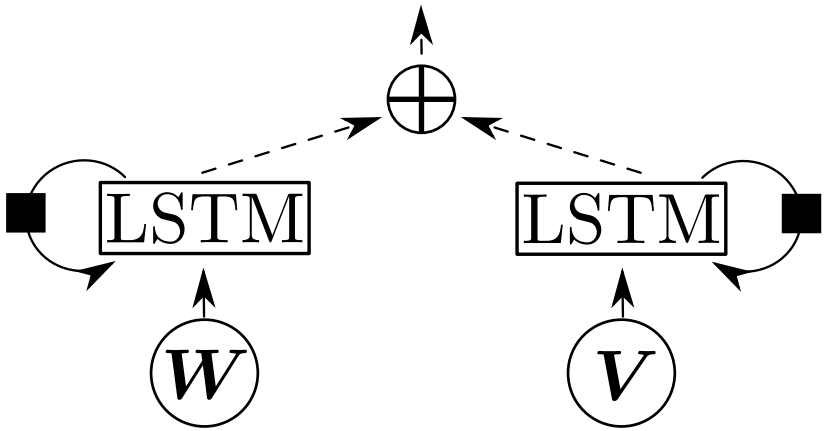
Compact Representation



Gaze control in the RL framework



(a) *Early Fusion Net*



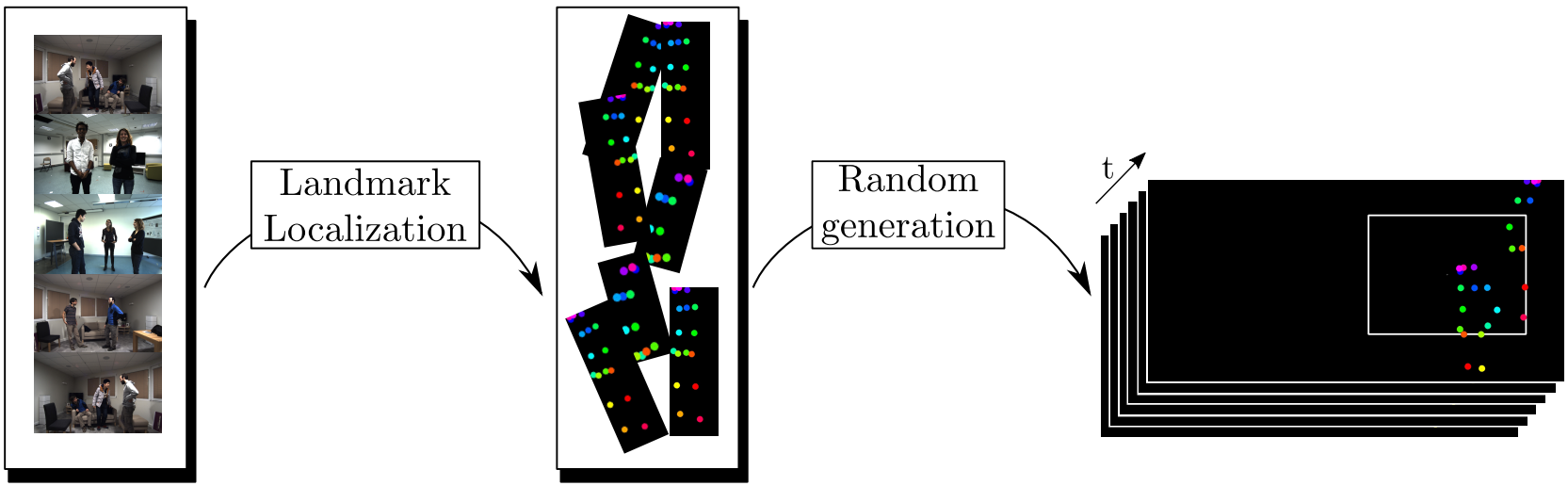
(b) *Late Fusion Net*

Figure: Proposed architectures to model the Q-function. Motor position are not displayed

with:

- $\boldsymbol{W}$: audio maps
- $\boldsymbol{V}$: visual maps

Simulated Environment for Training



Video

- Simulated environment
- Nao robot

Conclusion

- **We need learning for robot gaze control!**
- We propose an RL formulation.
- We show that it outperforms handcrafted strategies.
- Future works:
  - Add other cues such as people pose and gaze, speech recognition...
  - Continuous control