The Maersk Mc-Kinney Moller Institute

sdu.dk

Multi-view object pose distribution tracking for grasping on mobile robots

#sdudk

Lakshadeep Naik

Supervisor: Norbert Krüger

Co-supervisors/ Support group: Aljaz Kramberger, Thorbjørn M. Iversen, Jakob Wilm

FacilityCobot ReThiCare

SDU

May 2022

Mobile manipulation

Image source: [1]

+

Image source: [2]

=

Image source: [3]

[1] Universal Robots (UR)
[2] Mobile Industrial Robots (MiR)
[3] Enabled Robotics (ER)

Non-industrial / Welfare use-cases

- Time efficiency
- Avoiding failures

Image source: [1]

Image source: [2]

[1] Enabled Robotics (ER)
[2] SDU

Related work at SDU

Optimizing pose uncertainties for industrial applications

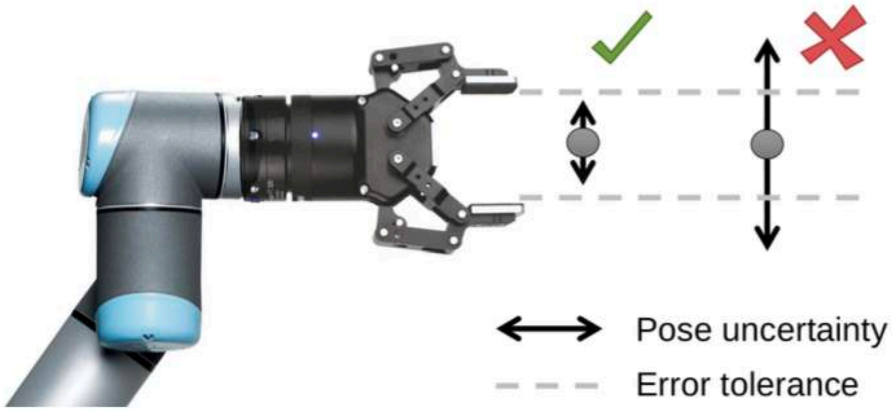
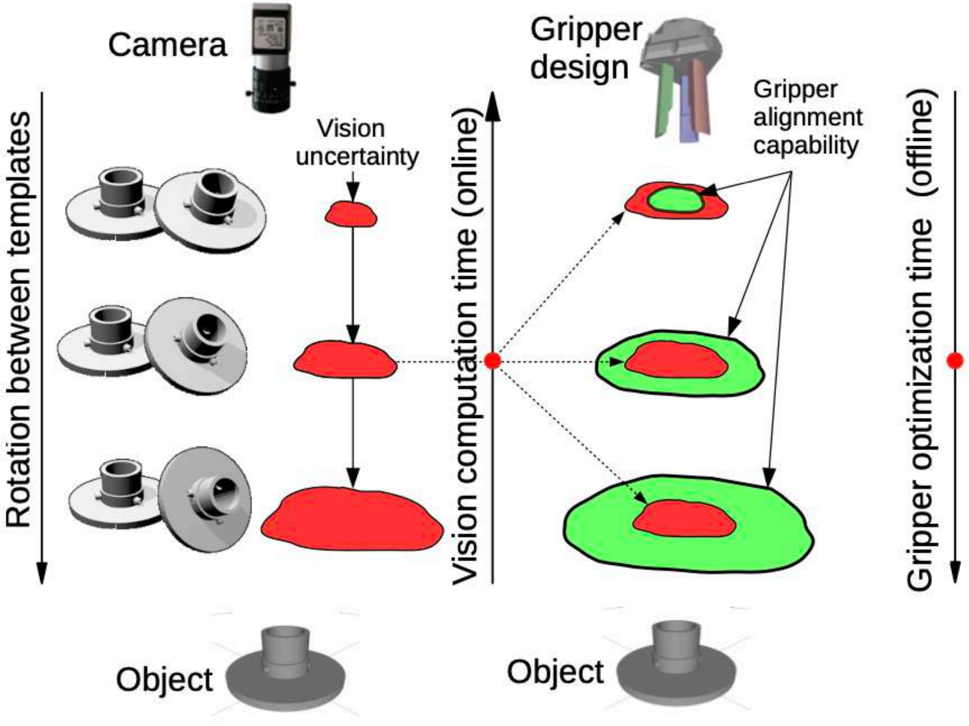


Image source: [1]



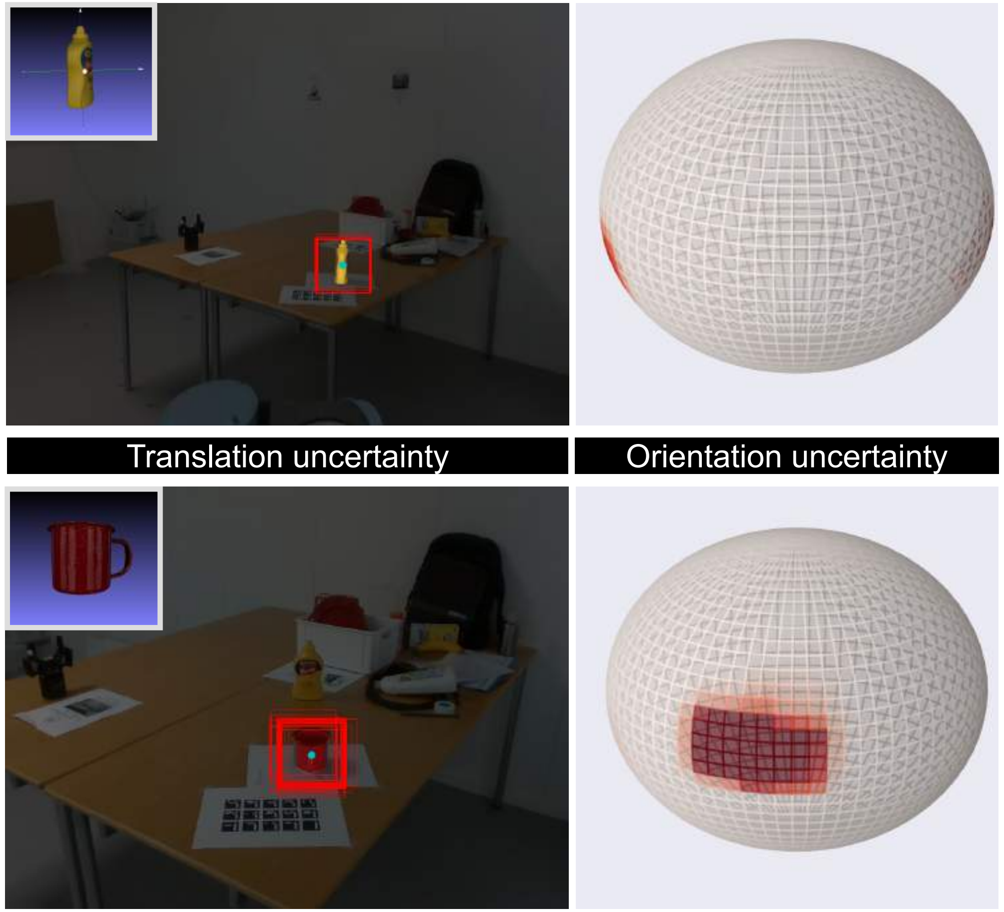
Img source: [2]

[1] Iversen, Thorbjørn Mosekjær. Automated configuration of vision sensor systems for industrial robotics. Diss. Syddansk Universitet, 2019.

[2] Hagelskjær, F., Kramberger, A., Wolniakowski, A., Savarimuthu, T. R., & Krüger, N. (2019, November). Combined Optimization of Gripper Finger Design and Pose Estimation Processes for Advanced Industrial Assembly. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2022-2029). IEEE.

Our approach

- Time efficiency
- Avoiding failures



Manipulator motion

Mobile base motion

Image source: Enabled Robotics

Improving time efficiency

Grasp planning using sequential navigation & manipulation actions

Go near table

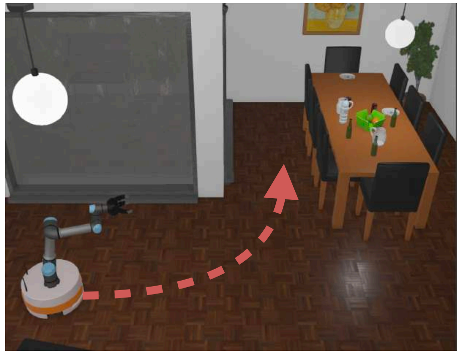
Navigation

Perception *Planning* *Control*

Grasp

Manipulation

Perception *Planning* *Control*



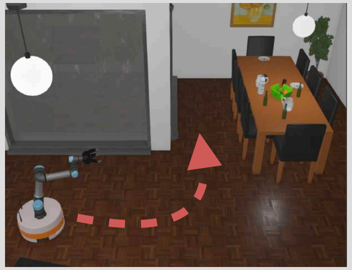
Grasp planning by coupling base and arm motion

Navigation

Perception *Planning* *Control*

Manipulation

Perception *Planning* *Control*



+

Reducing failures

- Estimation of the underlying uncertainty in object pose estimate
  - determine likelihood of success of the grasping task
  - take an action to reduce uncertainty in pose estimate before grasping

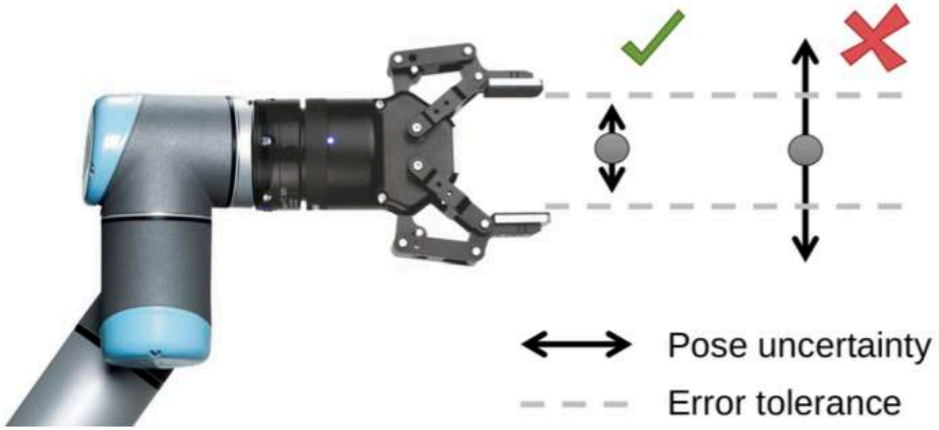
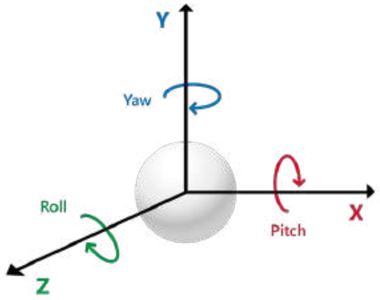


Image source: [1]



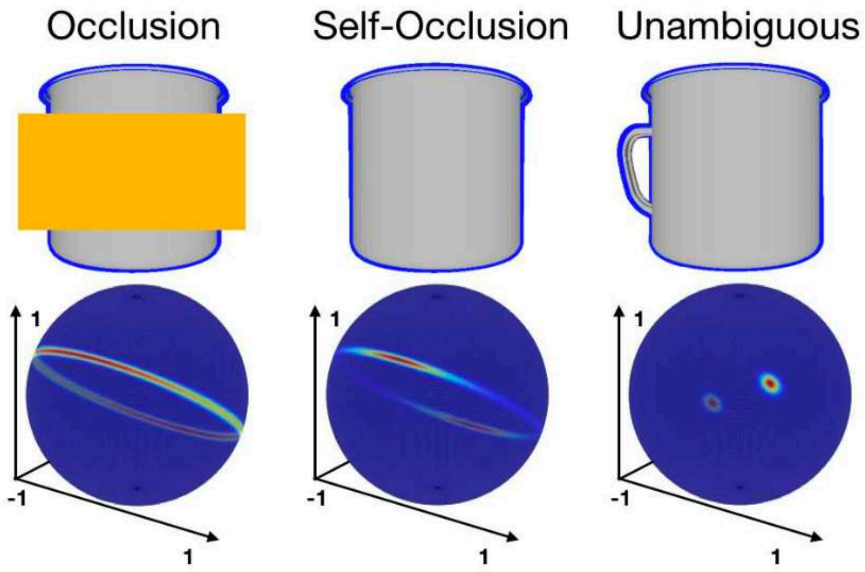
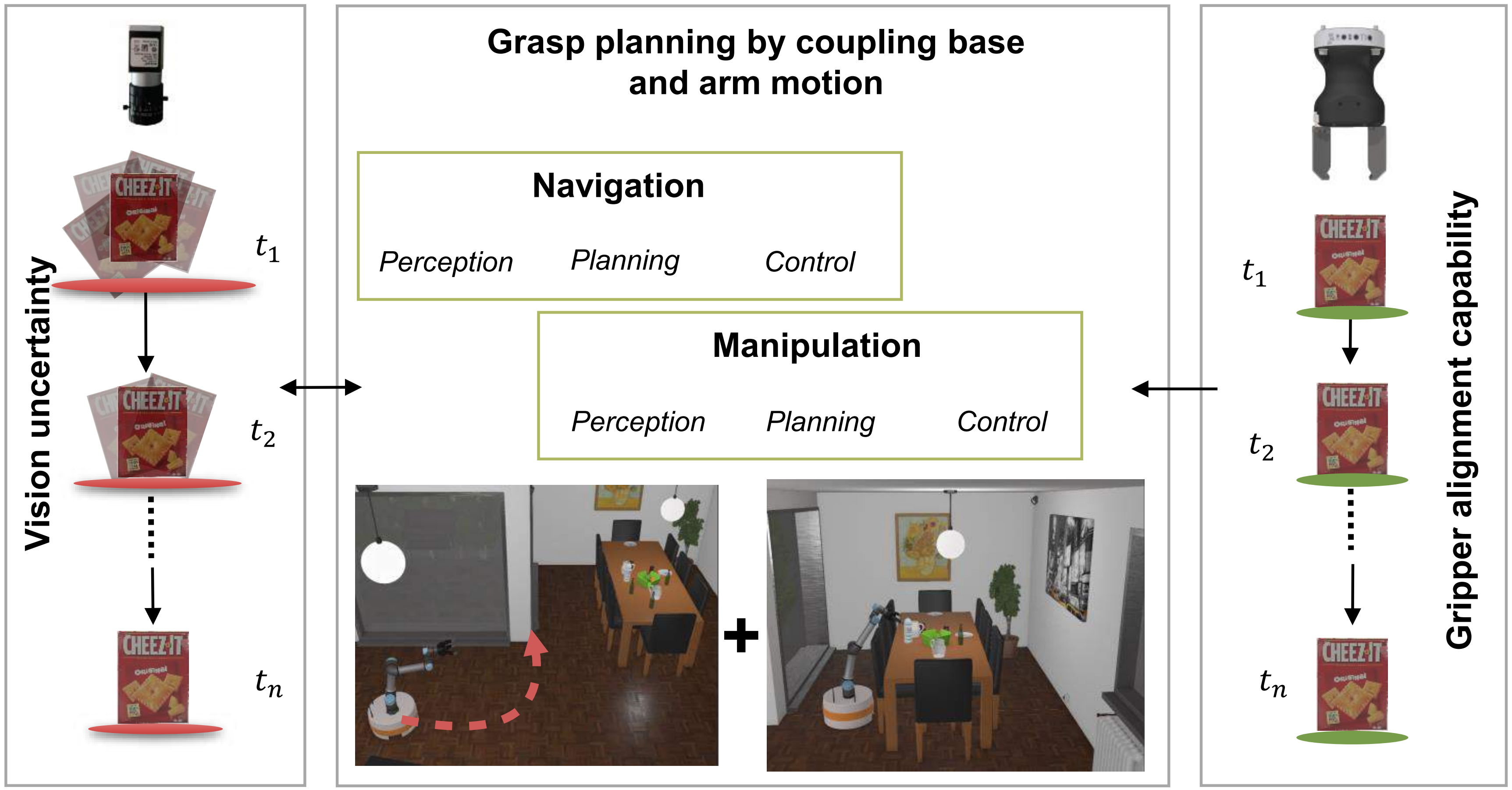


Image source: [2]

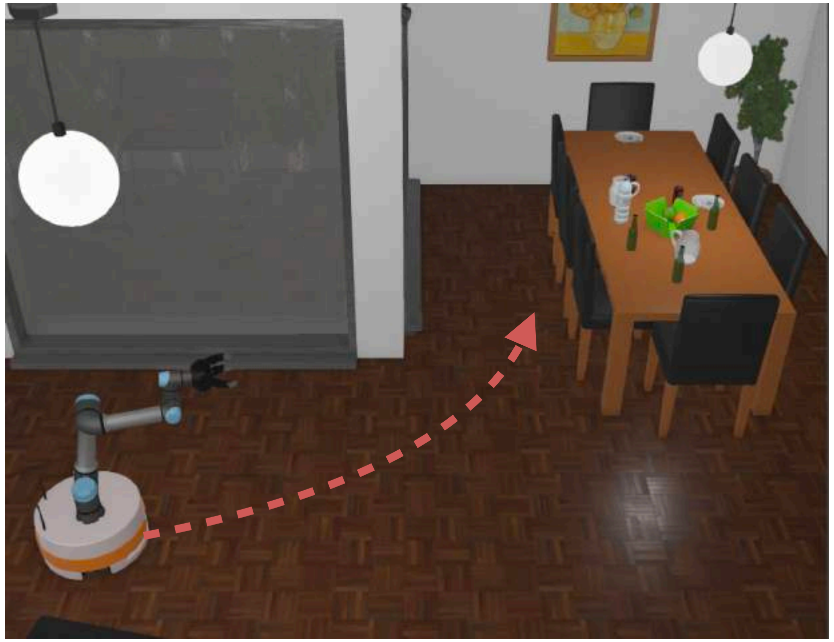
[1] Iversen, Thorbjørn Mosekjær. Automated configuration of vision sensor systems for industrial robotics. Diss. Syddansk Universitet, 2019.
[2] Manhardt, Fabian, et al. "Explaining the ambiguity of object detection and 6d pose from visual data." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

Problem formulation



Objectives

1

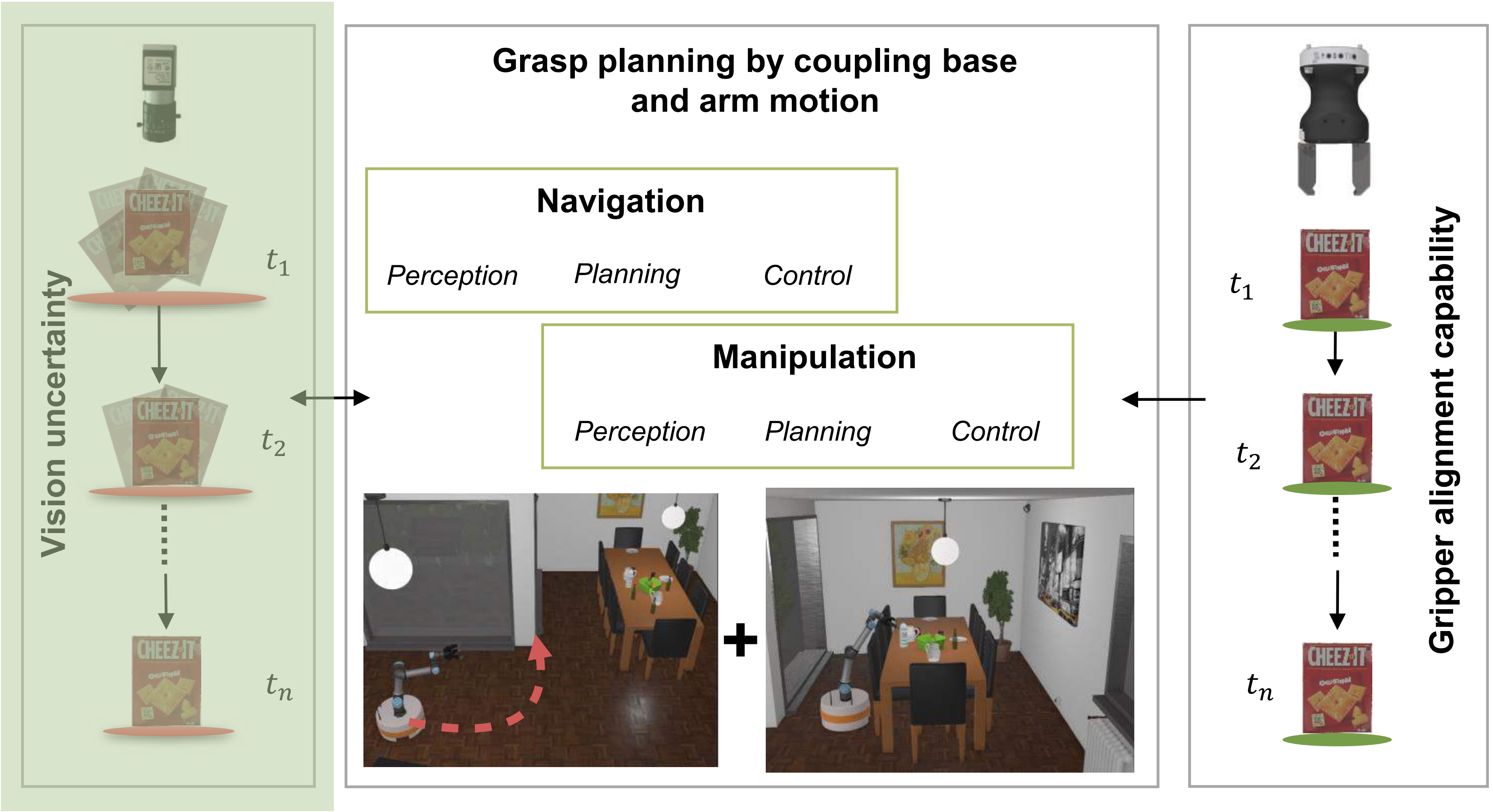


To enable object pose distribution tracking from distance for pre-grasp planning

2

To reduce the object pose uncertainties below that can be compensated by the gripper when robot is close enough to grasp the object

Problem formulation



Related work

Object pose distribution (uncertainty) estimation

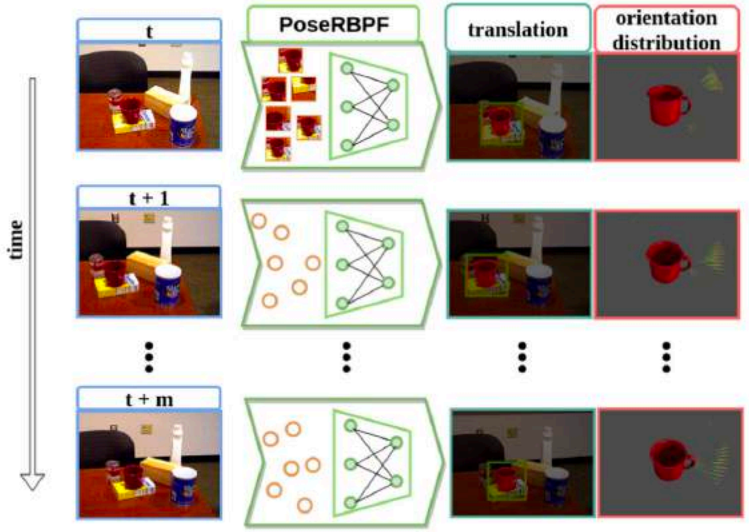


Image source: [1]

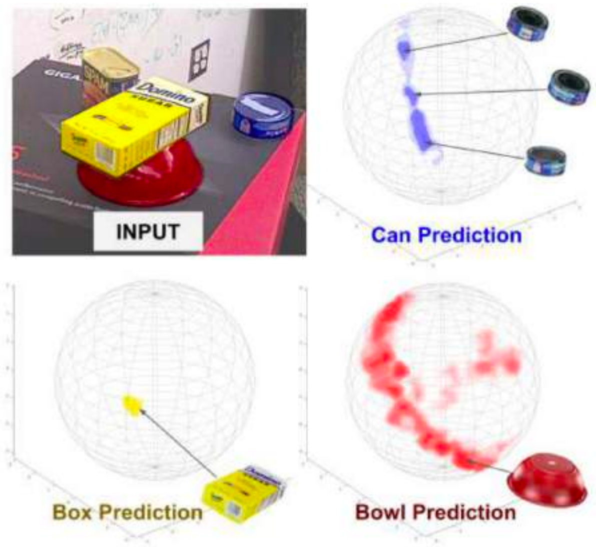


Image source: [2]

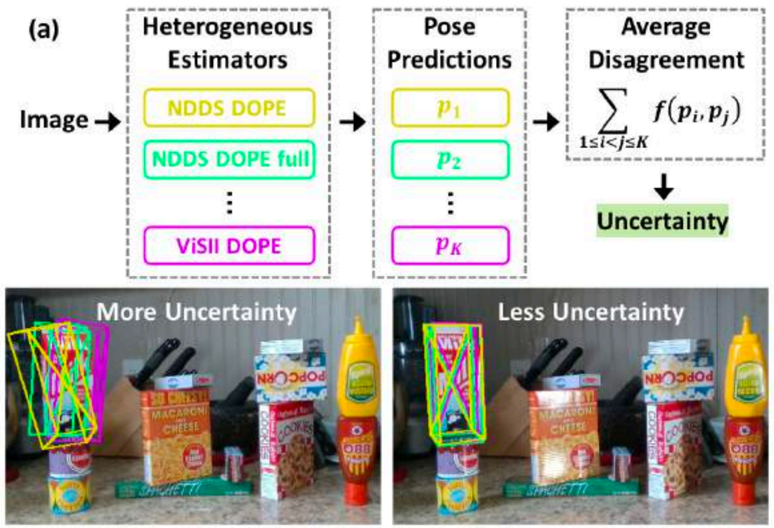


Image source: [3]

[1] Deng, Xinke, et al. "Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking." *IEEE Transactions on Robotics* 37.5 (2021): 1328-1342.
[2] Okorn, Brian, et al. "Learning orientation distributions for object pose estimation." *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.
[3] Shi, Guanya, et al. "Fast uncertainty quantification for deep object pose estimation." *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

Multi-view object pose distribution tracking

+

Related work

Multi-view pose distribution (uncertainty)

- Models posterior distribution as a uni-modal Gaussian distribution
- No temporal integration (tracking)

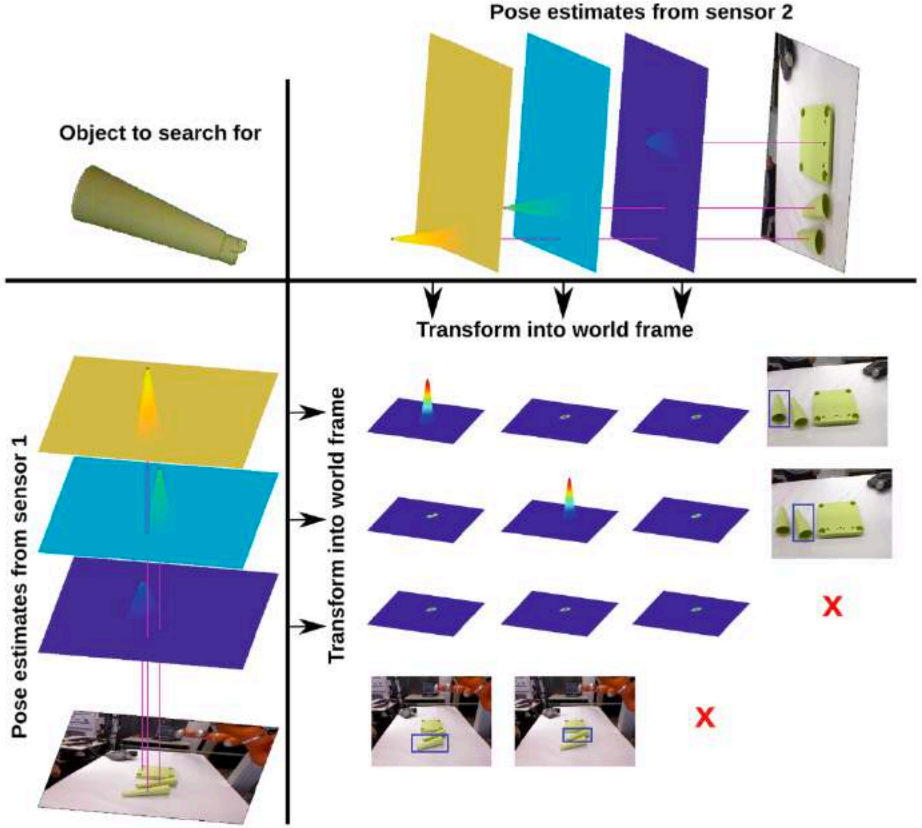
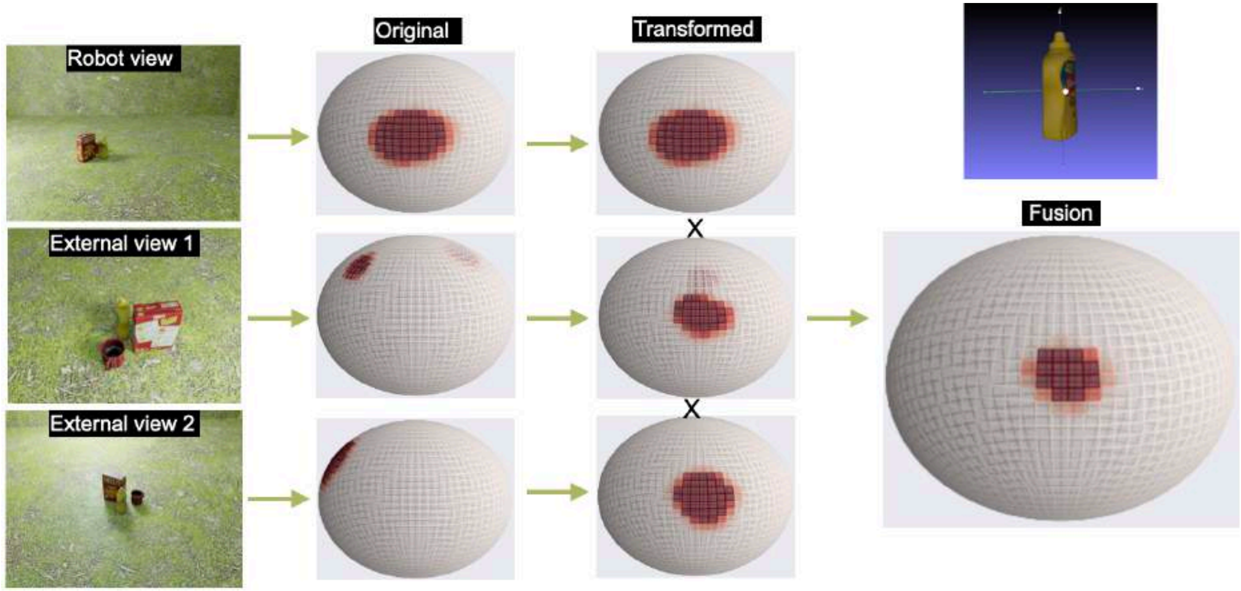


Image source: [1]

[1] Erkent, Özgür, Dadhichi Shukla, and Justus Piater. "Integration of probabilistic pose estimates from multiple views." *European Conference on Computer Vision*. Springer, Cham, 2016.

Our contribution

- Based on Rao-Blackwellized particle filter with de-noising auto-encoder for verifying observations [1]
- Extends [1] to fuse information from external cameras
- Both translation and orientation distributions are modeled as a multi-modal distributions

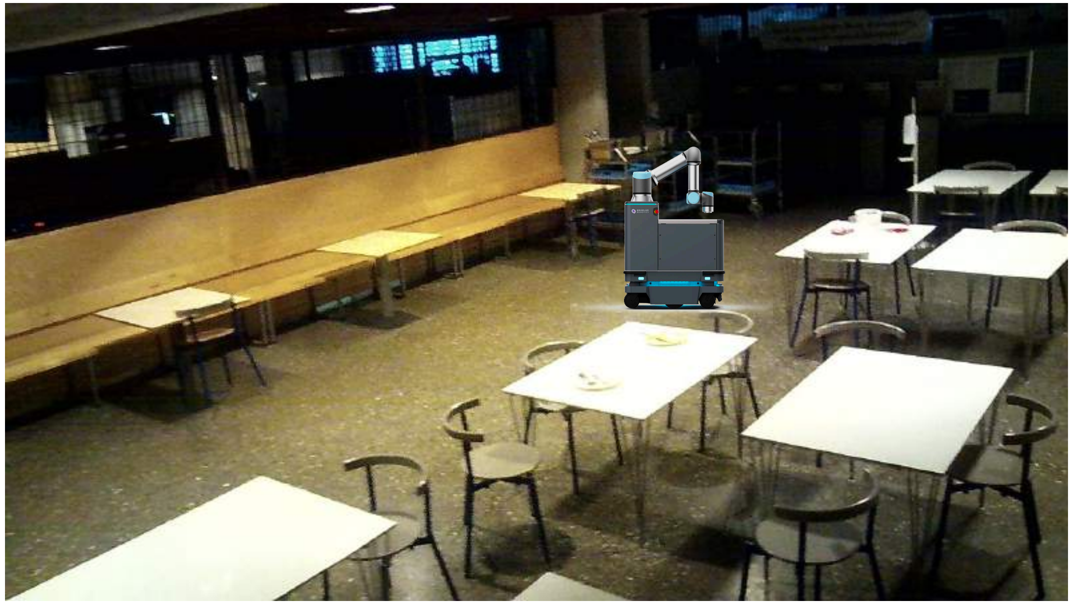


Naik, L., Iversen, T. M., Kramberger, A., Wilm, J., & Krüger, N. (Accepted/In press). Multi-view object pose distribution tracking for pre-grasp planning on mobile robots. In *2022 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1554-1561) IEEE.

[1] Deng, Xinke, et al. "Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking." *IEEE Transactions on Robotics* 37.5 (2021): 1328-1342.

FacilityCobot

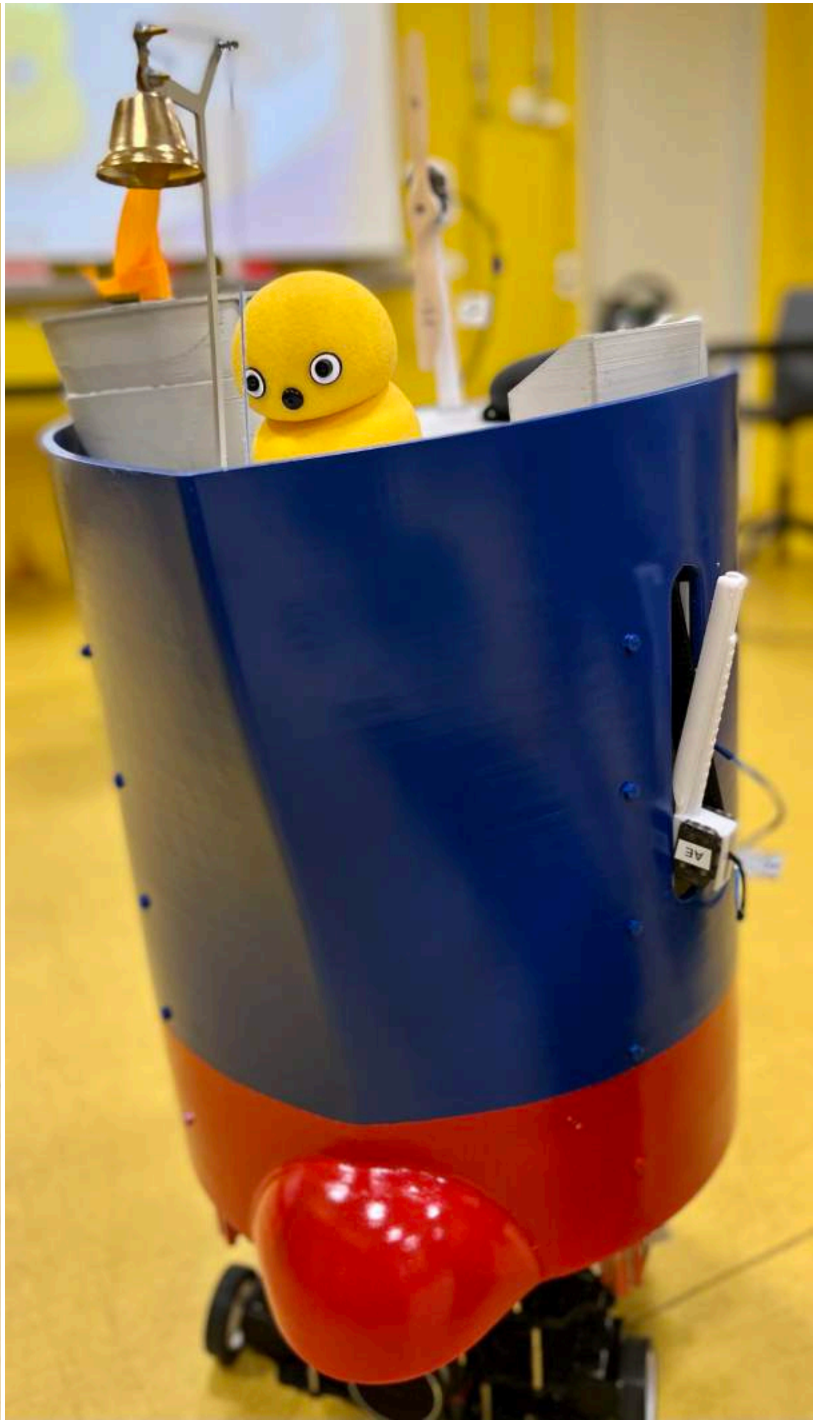
- A robot assistant for the cafeteria staff
- Mobile manipulator (Enabled Robotics) for cleaning and clearing cafeteria tables
- External stationary cameras (UbiqiSense Facility Sensors) for providing dynamic overview of the scene to the robot



ReThiCare

Plant Watering Robot (PWR)

- Re-thinking Care Robots
- Nautically designed plant watering robot to entertain people suffering from dementia at care homes and also carry out routine tasks such as plant watering



Particle filter

Posterior distribution model

Algorithm Particle_filter($\mathcal{X}_{t-1}, u_t, z_t$):

$\bar{\mathcal{X}}_t = \mathcal{X}_t = \emptyset$

for $m = 1$ *to* M *do*

 sample $x_t^{[m]} \sim p(x_t \mid u_t, x_{t-1}^{[m]})$

 $w_t^{[m]} = p(z_t \mid x_t^{[m]})$

 $\bar{\mathcal{X}}_t = \bar{\mathcal{X}}_t + \langle x_t^{[m]}, w_t^{[m]} \rangle$

endfor

for $m = 1$ *to* M *do*

 draw i *with probability* $\propto w_t^{[i]}$

 add $x_t^{[i]}$ *to* $\mathcal{X}_t$

endfor

return $\mathcal{X}_t$

State transition using motion model

Observation likelihood

Re-sampling

Source: **Probabilistic Robotics** by Sebastian Thrun, Wolfram Burgard and Dieter Fox

6D pose estimation

$$X_t = (x, y, z, \alpha, \beta, \gamma)$$

Posterior distribution

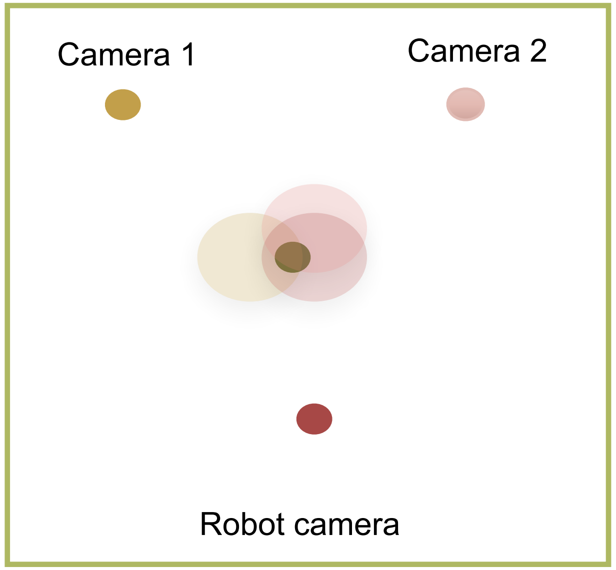
$$P(X_t \mid Z_{1:t}) = X_t^1, X_t^2, \ldots, X_t^M$$

Posterior distribution model

$$P(X_t \mid Z_{1:t}) = P(T_t, R_t \mid Z_{1:t}) = P(T_t \mid Z_{1:t}) P(R_t \mid T_t, Z_{1:t})$$

Rao-Blackwellisation

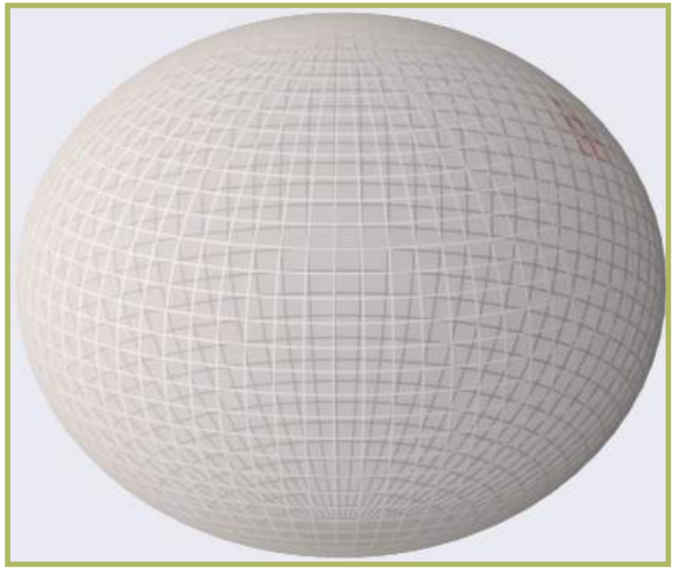
$$X_t = T_t \; R_t$$



Mixture of gaussian distribution

$$P(T_t) = \sum_{i=1}^{K} \phi_i \; \mathcal{N}(\mu_i, \sigma_i)$$

K = no. of cameras



Histogram distribution by discretizing orientation space

$$P(R_t) = \frac{p_{k,t}}{|R_{k,t}|}$$

State transition using motion model

Propagating particles originating from external cameras

$$ {}^{cr_t}_{o_t}T = {}^{cr_t}_{ce_t}T \cdot {}^{ce_t}_{ce_{t-1}}T \cdot {}^{ce_{t-1}}_{cr_{t-1}}T \cdot {}^{cr_{t-1}}_{o_{t-1}}T \cdot {}^{o_{t-1}}_{o_t}T $$

incorporates robot pose uncertainty

constant

constant from t-1 estimate

Brownian motion model

Propagating particles originating from robot camera

$$ {}^{cr_t}_{o_t}T = {}^{cr_t}_{cr_{t-1}}T \cdot {}^{cr_{t-1}}_{o_{t-1}}T \cdot {}^{o_{t-1}}_{o_t}T $$

odometry motion model

Brownian motion model

ce_t - external camera frame

cr_t - robot camera frame frame

Observation likelihood

De-noising auto-encoder training

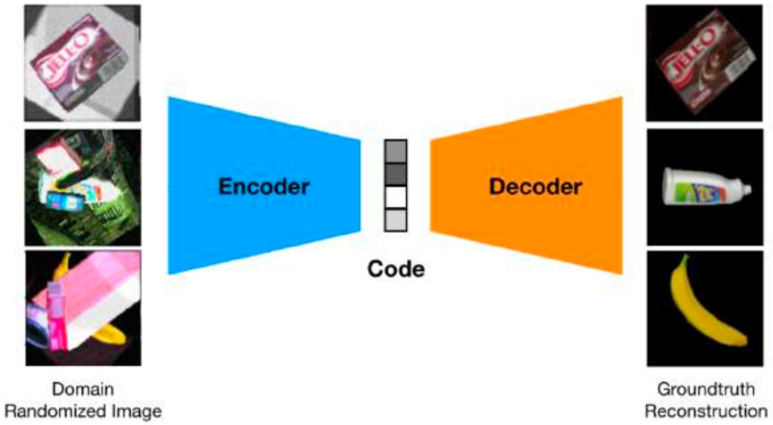
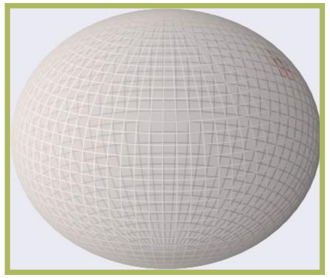


Image source: [1]

Offline codebook generation



Online codebook matching

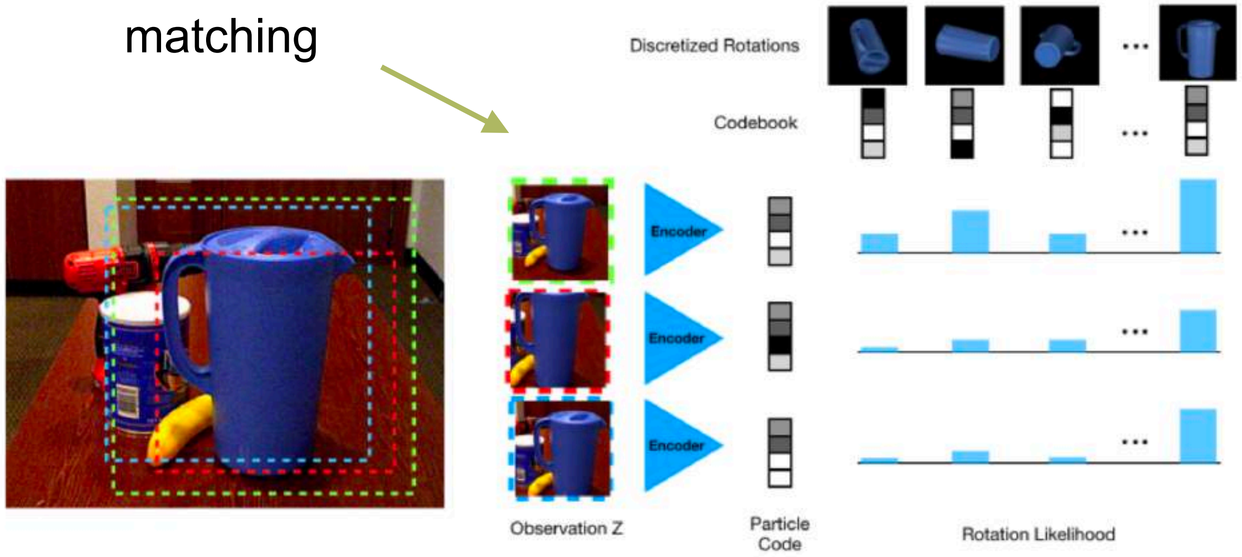


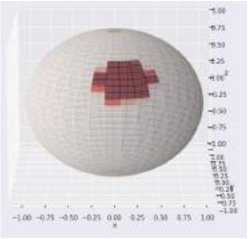
Image source: [1]

[1] Deng, Xinke, et al. "Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking." *IEEE Transactions on Robotics* 37.5 (2021): 1328-1342.

Observation likelihood

Single-view

$$P(Z_t \mid T_t, R_t) \prec P(R_t \mid T_t, Z_t) P(Z_t \mid T_t)$$



$f(Z(R_c^j, T_0))$ - codebook embeddings

$f(Z_t(T_t))$ - current embedding

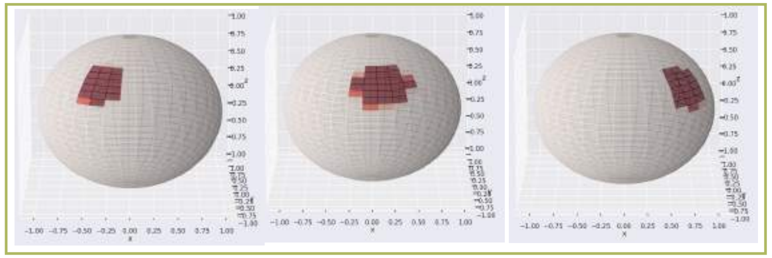
$$P(R_c^j \mid T_t, Z_t)) \prec \Phi\left(\frac{f(Z_t(T_t))\,.\,f(Z(R_c^j, T_0))}{||f(Z_t(T_t))||\,.\,||f(Z(R_c^j, T_0))||}\right)$$

(Cosine distance)

$$P(Z_t \mid T_t) \prec \sum_j P(R_c^j \mid T_t, Z_t))$$ - observation likelihood

Multi-view

$$P({}^1Z_t, {}^2Z_t, {}^rZ_t \mid {}^rT_t, {}^rR_t) \prec P({}^rR_t \mid {}^rT_t, {}^1Z_t, {}^2Z_t, {}^rZ_t) P({}^1Z_t, {}^2Z_t, {}^rZ_t \mid {}^rT_t)$$



=

Orientation distributions

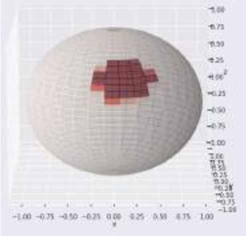
$P(R_c^j \mid {}^rT_t, {}^1Z_t))$ X $P(R_c^j \mid {}^rT_t, {}^2Z_t))$ X $P(R_c^j \mid {}^rT_t, {}^rZ_t))$ =

Observation likelihood
in each frames

Observation likelihood

Single-view

$$P(Z_t \mid T_t, R_t) \prec P(R_t \mid T_t, Z_t) P(Z_t \mid T_t)$$



$f(Z(R_c^j, T_0))$ - codebook embeddings

$f(Z_t(T_t))$ - current embedding

$$P(R_c^j \mid T_t, Z_t)) \prec \Phi\left(\frac{f(Z_t(T_t))\,.\,f(Z(R_c^j, T_0))}{||f(Z_t(T_t))||\,.\,||f(Z(R_c^j, T_0))||}\right)$$

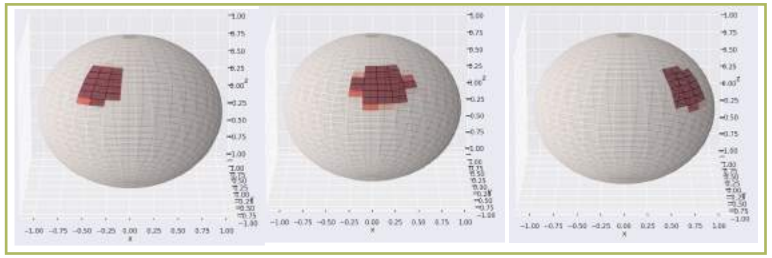
(Cosine distance)

$$P(Z_t \mid T_t) \prec \sum_j P(R_c^j \mid T_t, Z_t))$$

- observation likelihood

Multi-view

$$P({}^1Z_t, {}^2Z_t, {}^rZ_t \mid {}^rT_t, {}^rR_t) \prec P({}^rR_t \mid {}^rT_t, {}^1Z_t, {}^2Z_t, {}^rZ_t) P({}^1Z_t, {}^2Z_t, {}^rZ_t \mid {}^rT_t)$$



=

Orientation distributions

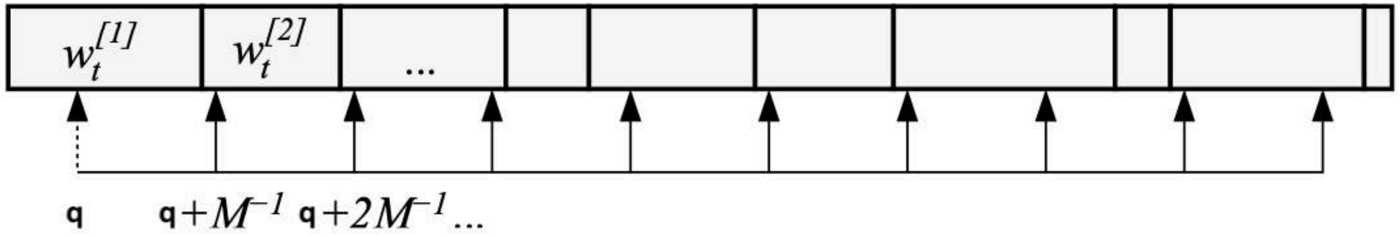
Laplace estimator

$$P(R_c^j \mid {}^rT_t, {}^1Z_t))^{+\lambda} \times P(R_c^j \mid {}^rT_t, {}^2Z_t))^{+\lambda} \times P(R_c^j \mid {}^rT_t, {}^rZ_t))^{+\lambda} =$$

Observation likelihood
in each frames

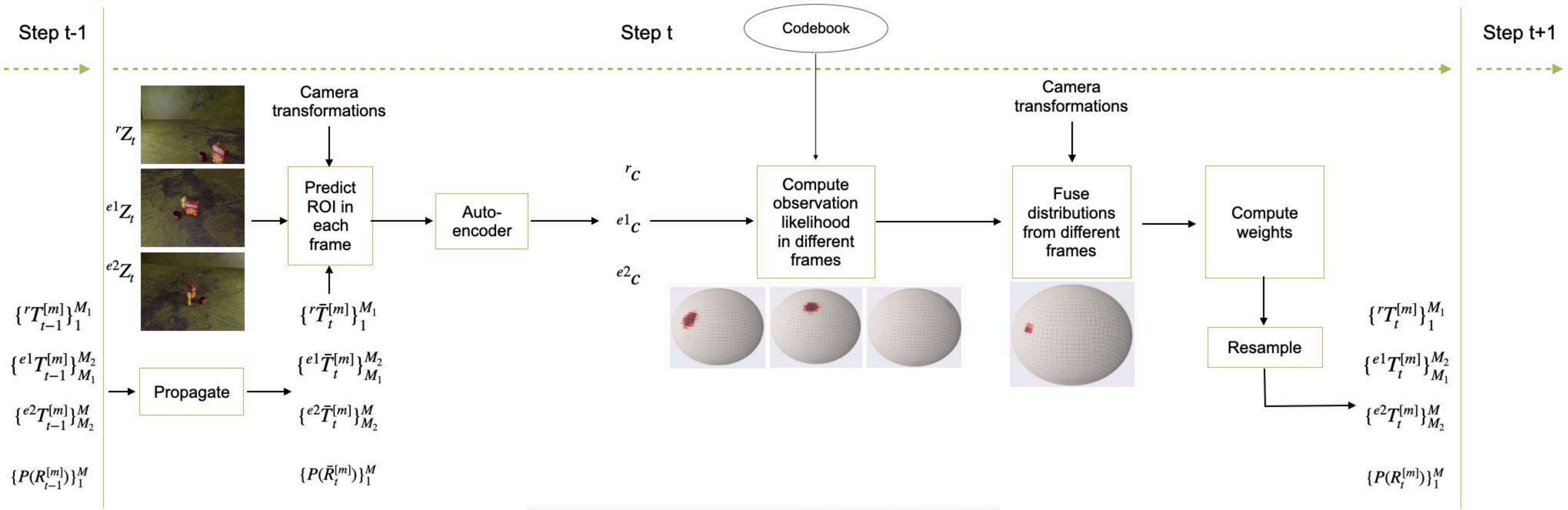
Re-sampling

- Weight of each (translation) particle is computed using **marginal probability** of histogram distribution
- Particles are re-sampled to increase the number of particles with good weights using **low-variance re-sampling**



Source: **Probabilistic Robotics** by Sebastian Thrun, Wolfram Burgard and Dieter Fox

Proposed approach

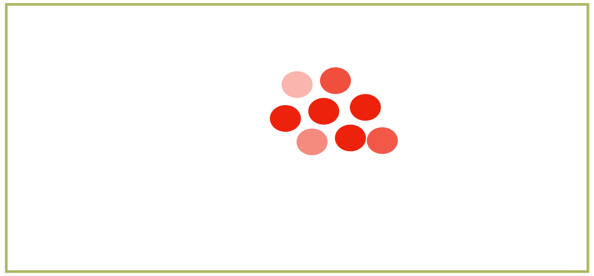


M - number of particles (1 < M1 < M2 < M)
r - robot camera frame
e1, e2 - external camera frames
T - translational component of pose estimate
P(R) - probability distribution of rotational component (3D histogram)
c - code generated by auto-encoder

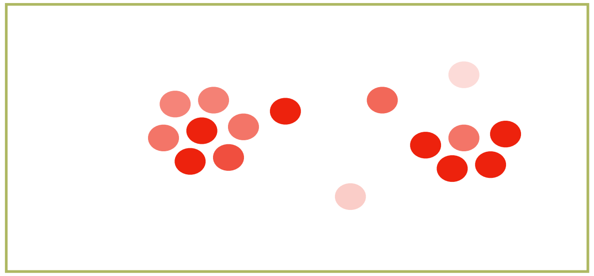
Translation expectation

$$E(T_t) = \begin{cases} \sum_1^M w_{m,t} T_{m,t} \; * & \text{if } P(T_t) \; is \; unimodal \\ max \; (P(T_t)) & \text{otherwise} \end{cases}$$

Uni-modal



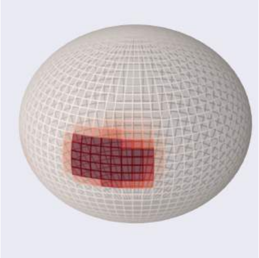
Multi-modal

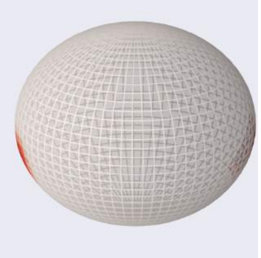


→ $P(T_t)$ - translation particles

→ Modality of the distribution is determined using Henze-Zirkler multivariate normality test

Orientation expectation





Initial estimate

$$R_{i,j,k,t} = \sum_{m=1}^{M} \frac{p_{m,i,j,k,t}}{|R_{m,i,j,k,t}|}$$

where:

i: 0 to 72 (bank)
j: 0 to 72 (azimuth)
k: 0 to 37 (elevation)
M: no of particles

$$R^{l}_{i,j,k,t} = argmax_{(i^{'},j^{'},k^{'})\subset(i,j,k),|(i^{'},j^{'},k^{'})|=L} \sum_{(i^{'},j^{'},k^{'})\subset(i,j,k)} R_{i,j,k,t}$$

$$E(R_t) = \begin{cases} \frac{\sum_1^L R^l_{i,j,k,t} * eulerToQuaternion(i,j,k)}{l} & \text{if } R^l_{i,j,k,t} is\ unimodal \\ max(R^l_{i,j,k,t}) & \text{otherwise} \end{cases}$$

Temporal fusion

$$E(\bar{R}) = {}^{r_t}_{r_{t-1}}T \circ R_{t-1}$$

$$R^p_{i,j,k,t} = arg_{(l^{'})\subset(l)} abs(R^l_{i,j,k,t} - E(\bar{R})) < r_{thresh}$$

r_{thresh} : threshold for difference between orientations

$$p = \frac{|R^p_{i,j,k,t}|}{|R^l_{i,j,k,t}|}$$

$$E(R_t) = \begin{cases} a_{t-1}\, E(\bar{R}) + a_t \frac{\sum_1^L R^l_{i,j,k,t} * eulerToQuaternion(i,j,k)}{l} & \text{if } p > p_{thres}\ and\ R^l_{i,j,k,t}\ is\ unimodal \\ \frac{\sum_1^L R^l_{i,j,k,t} * eulerToQuaternion(i,j,k)}{l} & \text{if } p < p_{thres}\ and\ R^l_{i,j,k,t}\ is\ unimodal \\ a_{t-1}\, E(\bar{R}) + a_t \frac{\sum_1^L R^p_{i,j,k,t} * eulerToQuaternion(i,j,k)}{l} & \text{if } p > p_{thres}\ and\ R^l_{i,j,k,t}\ is\ multimodal \\ max(R^l_{i,j,k,t}) & \text{if } p < p_{thres}\ and\ R^l_{i,j,k,t}\ is\ multimodal \\ E(\bar{R}) & \text{otherwise} \end{cases}$$

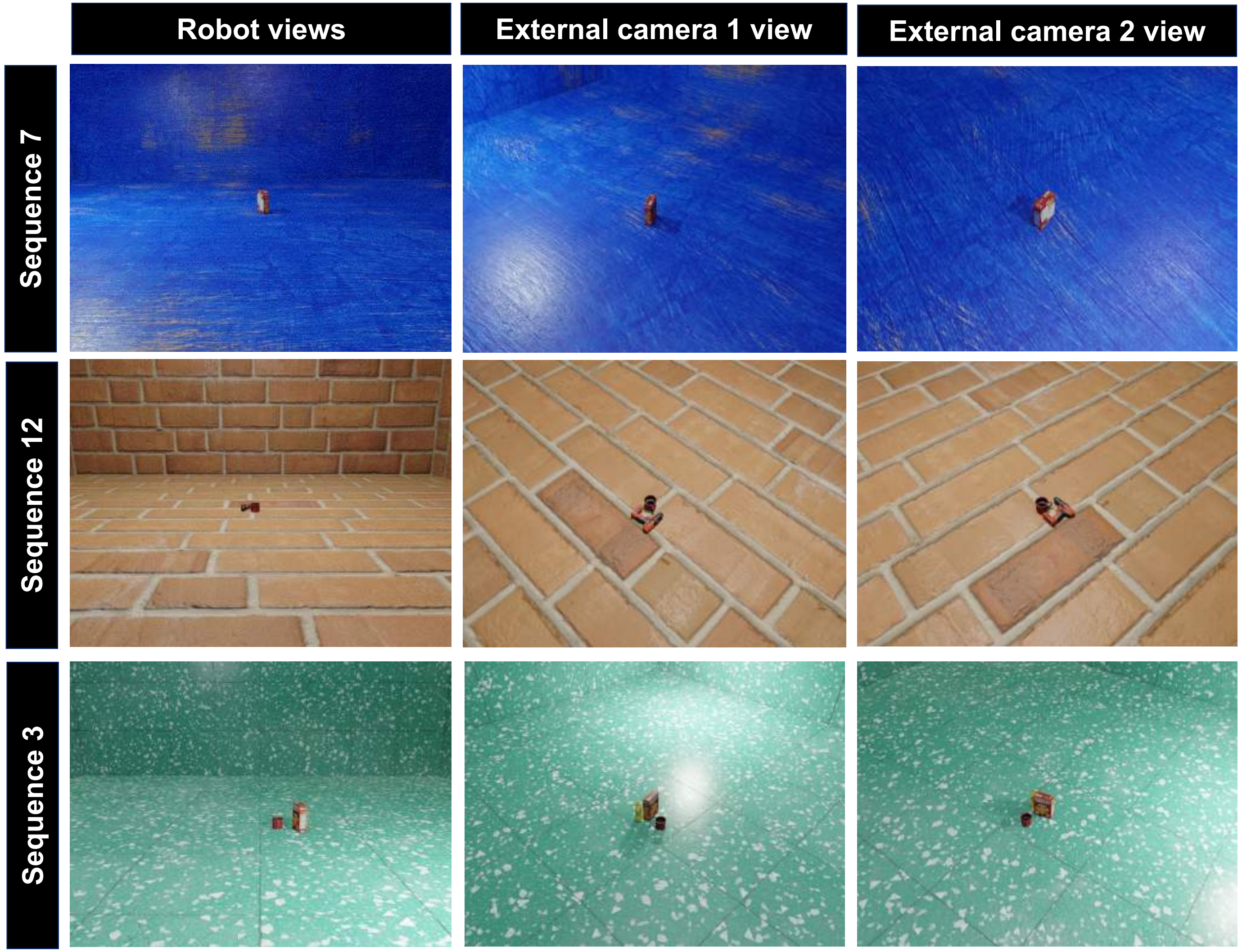
where:

a_{t-1} : scaling factor for rotation estimate at time t-1
a_t : scaling factor for rotation estimate at time t
$a_{t-1} + a_t = 1$
a_{thres} : difference between the angles to determine to if orientation estimate at time t should be incorporated in the tracked estimate
p_{thres} : threshold for determining percentage of orientations within the orientation expectation at previous time step

Experimental Evaluation

- Simulated dataset [1] containing 8 different YCB objects created using photo-realistic renderer

- Each sequence contains view from robot and external cameras with robot camera simulating robots base and arm motion



[1] L. Naik, "Multi-view rendered YCB dataset for mobile manipulation," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.

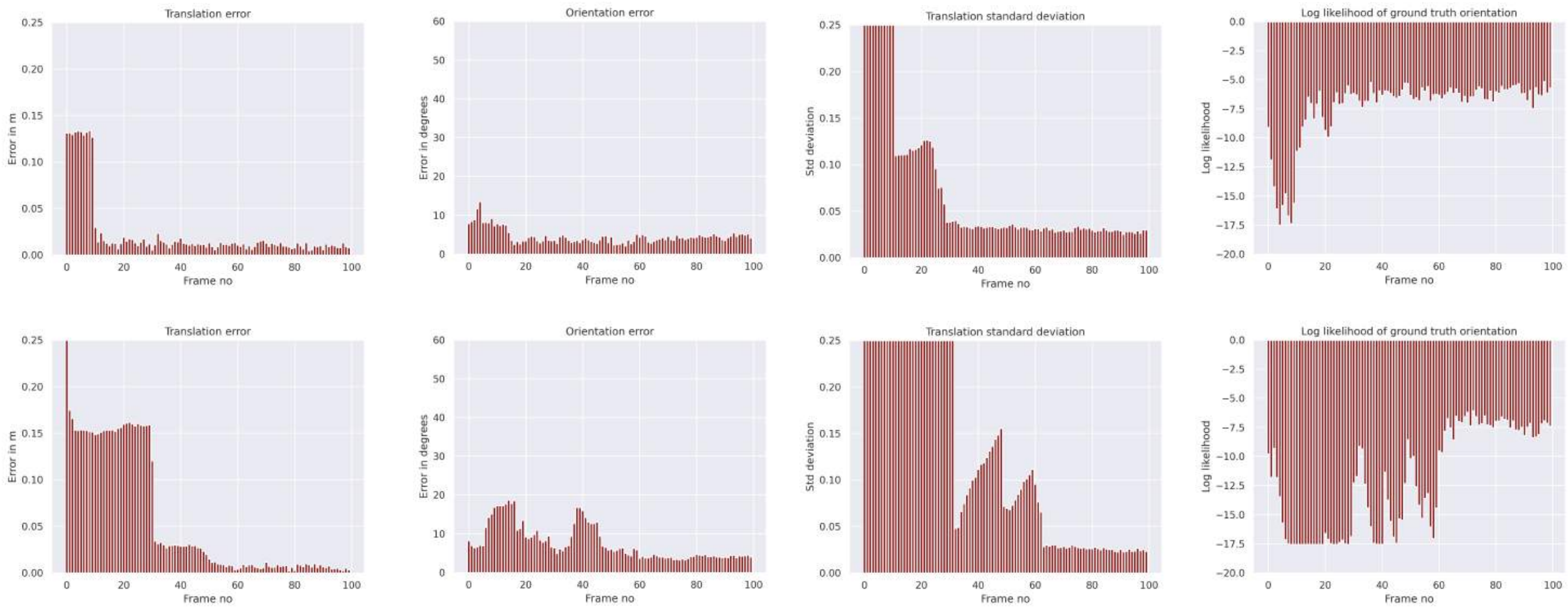
Qualitative results

Example 1: Robot moving closer towards the object

Translation error | Orientation error | Translation uncertainty | Orientation uncertainty

Multi-view

Single-view

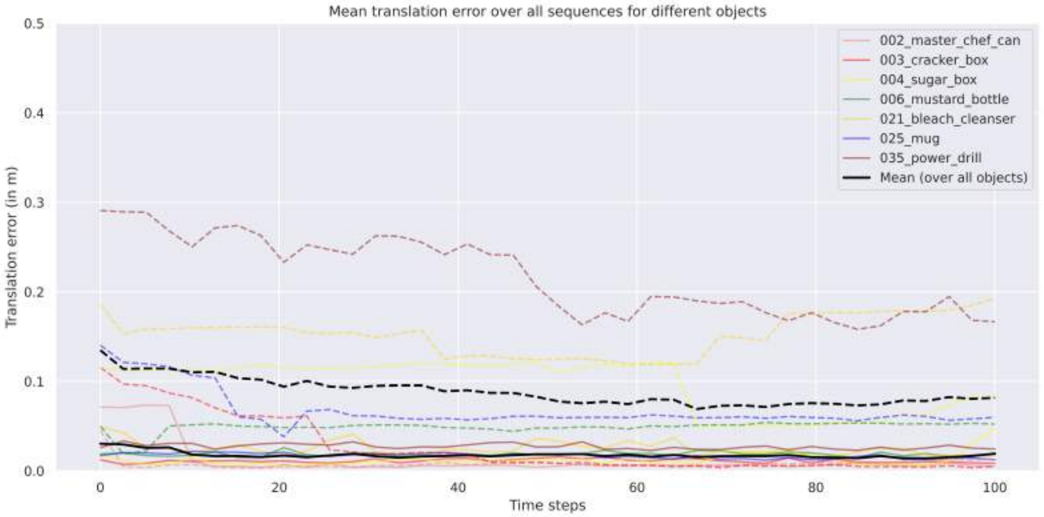


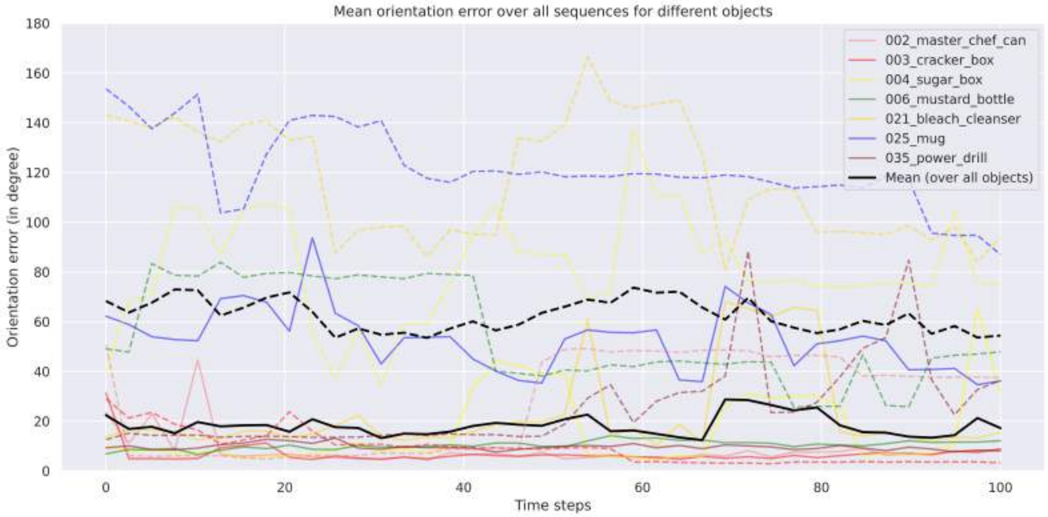
Original robot view

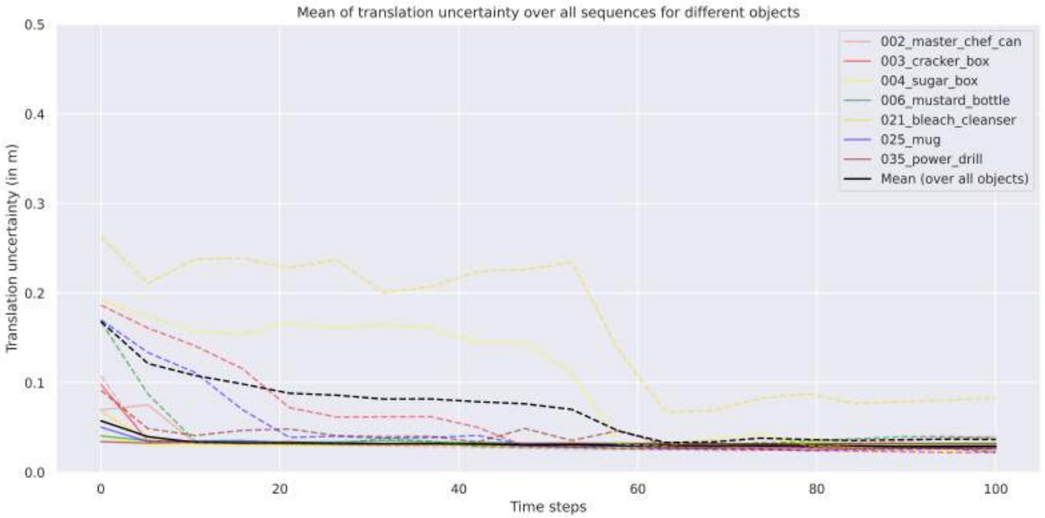
Multi-view result

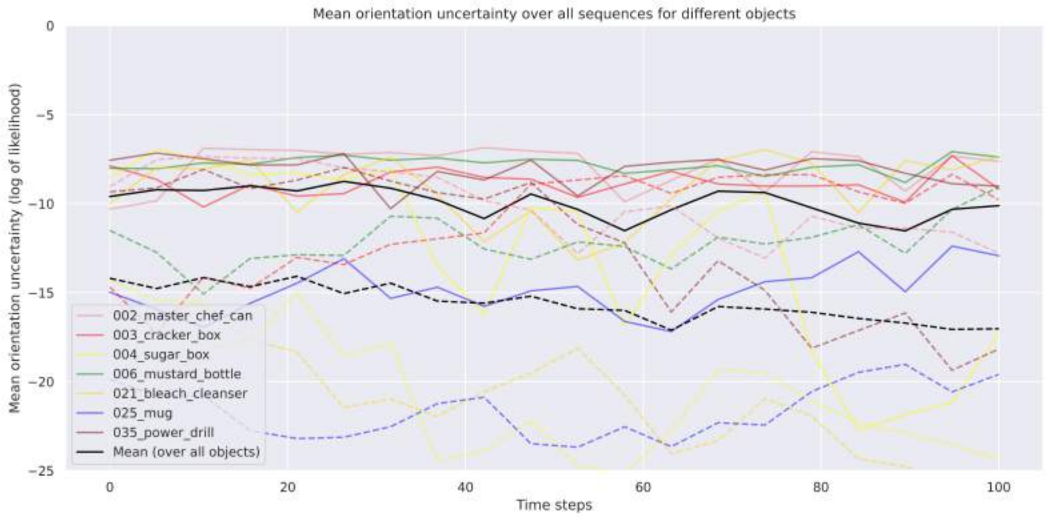
Single-view result

Results







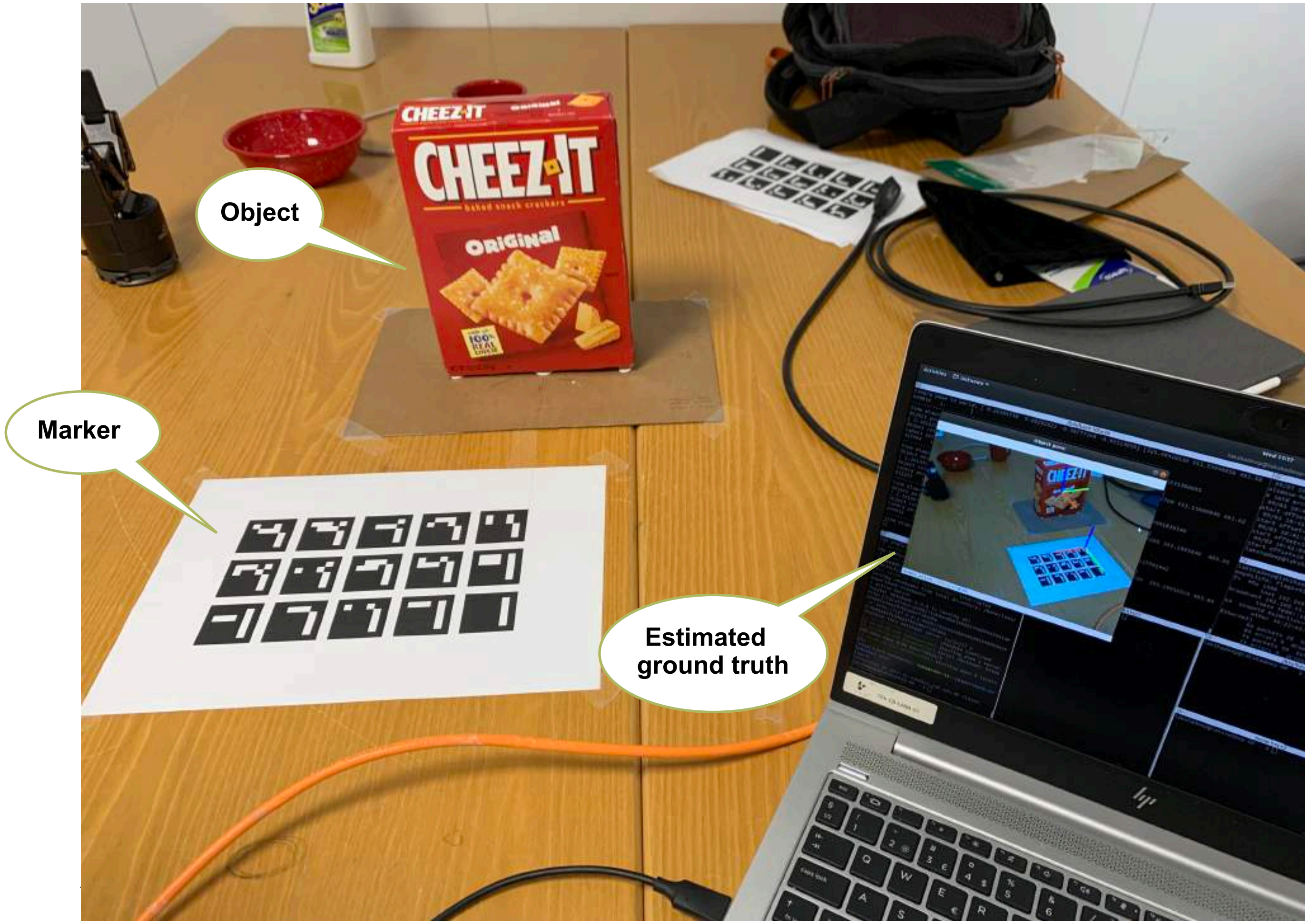


Multi-view
Single-view

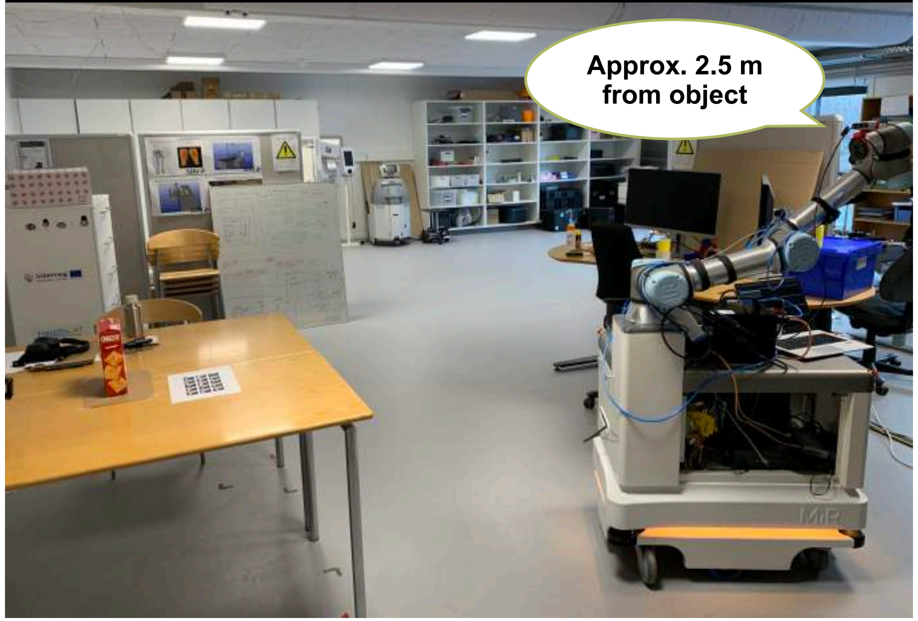
The Maersk Mc-Kinney Moller Institute

Real world evaluation

UR robot with marker setup



Start pose of the robot

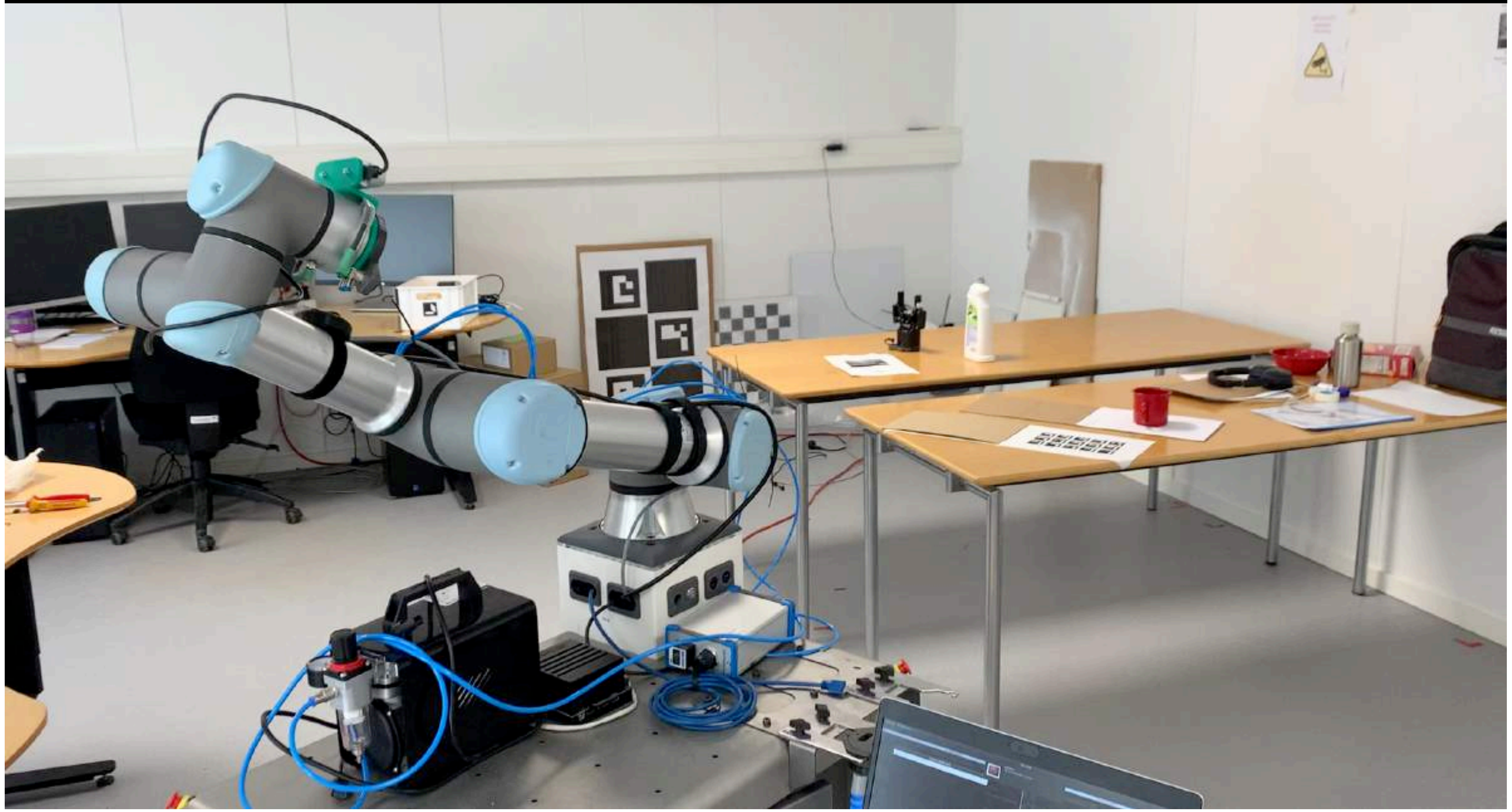


End pose of the robot

The Maersk Mc-Kinney Moller Institute

Example recording

Robot camera motion simulation



External camera simulation

External camera simulation

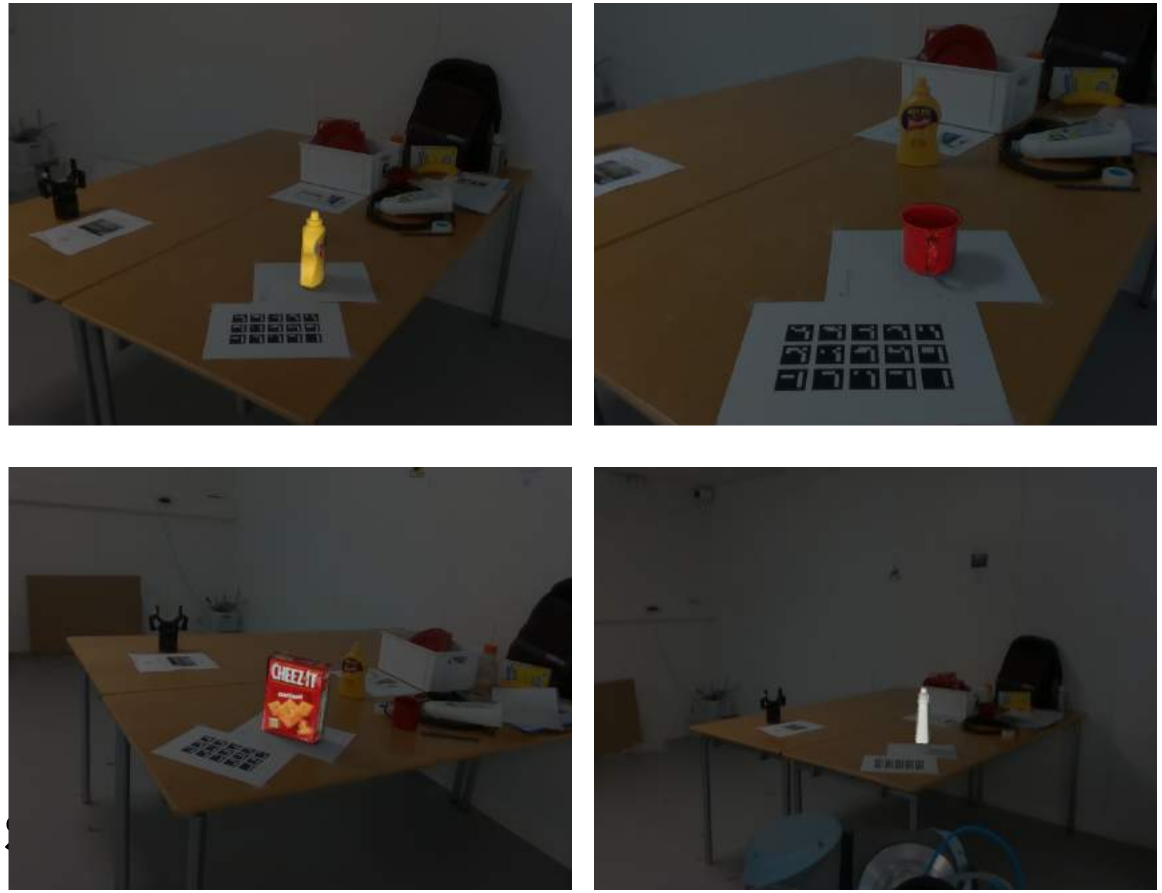
SDU

Real world evaluation

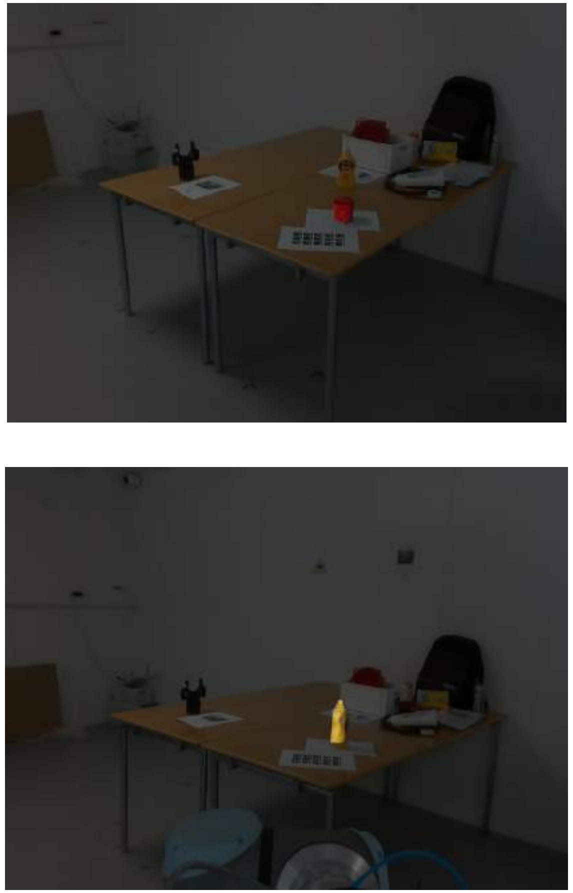
Quality of ground truths

Setup

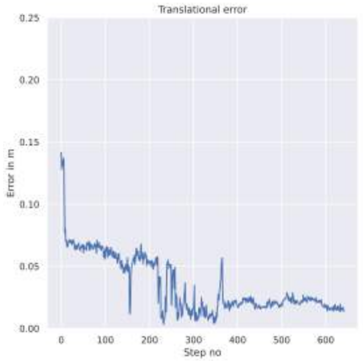
Good estimates



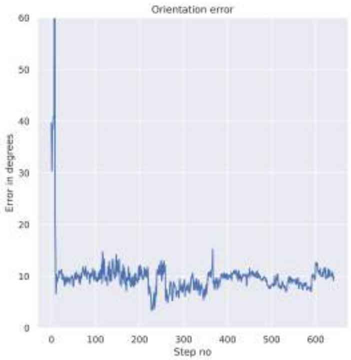
Bad estimates



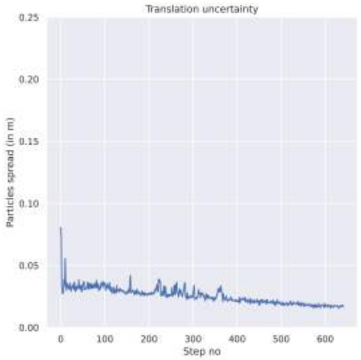
Multi-view (robot and external)



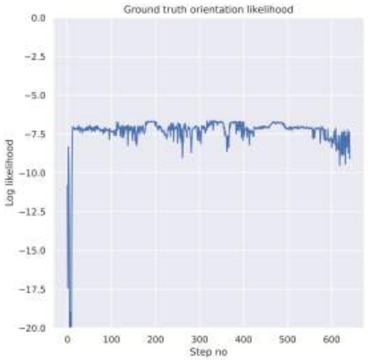
(a)



(b)

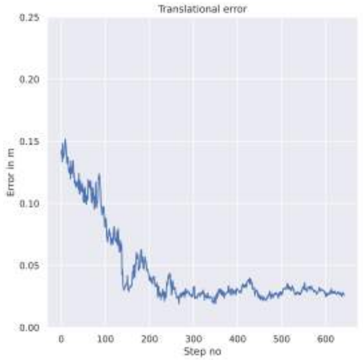


(c)

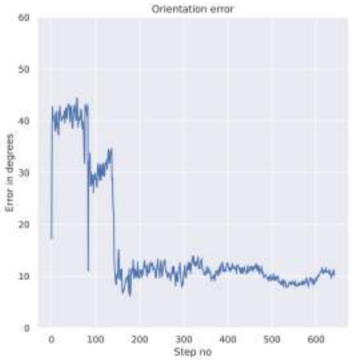


(d)

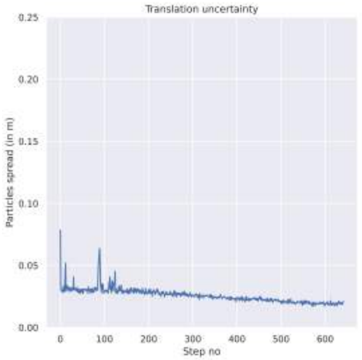
Single-view (robot only)



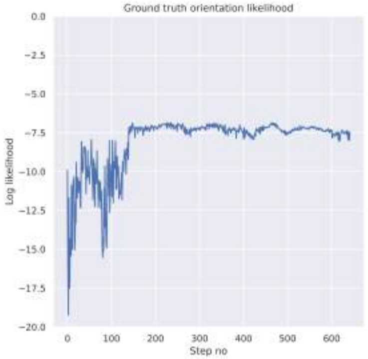
(e)



(f)



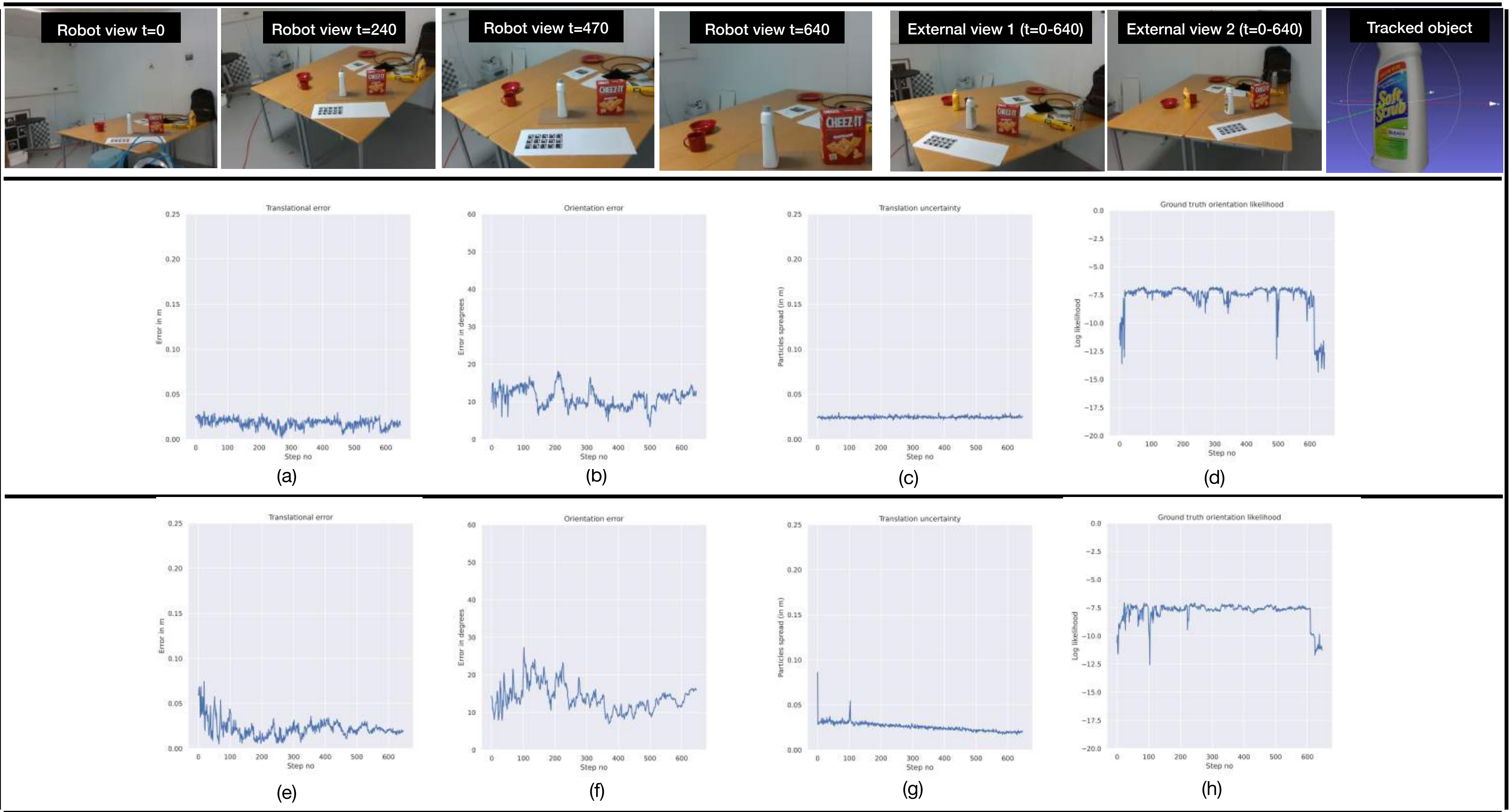
(g)



(h)

Multi-view
(robot and external)

Single-view
(robot only)

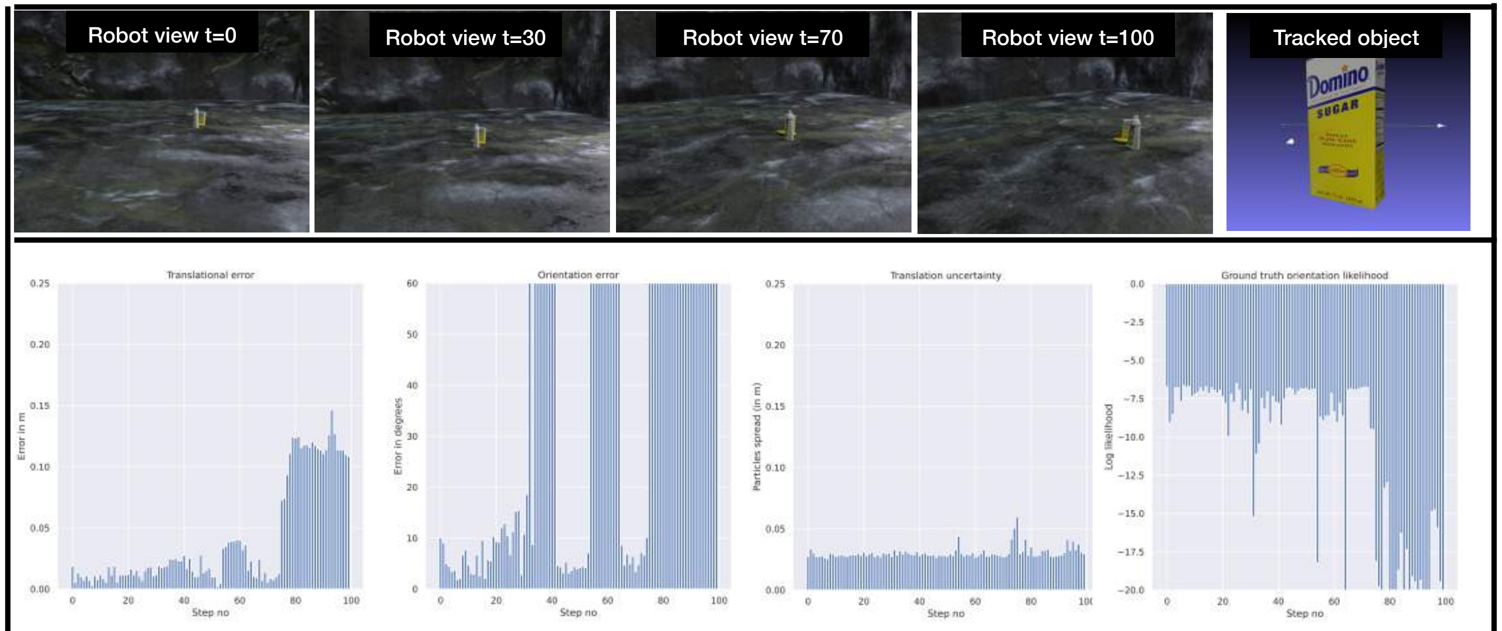


Conclusions

- The proposed approach generally results in faster convergence of translation and orientation errors and uncertainties compared to the single view baseline
- However, there are instances when single view approach performs better compared to multi-view as robot camera has much better observation compared to external camera views

Ongoing and future work

- Determining when to use robot and external cameras
- Maintaining multiple orientation expectations at each time step
- Planning robot camera views to improve estimates



(a) (b) (c) (d)

Thank you

Questions?