2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Perceptual Factors for Environmental Modeling in Robotic Active Perception

David Morilla-Cabello, Jonas Westheider, Marija Popović, and Eduardo Montijano

***Abstract*— Accurately assessing the potential value of new sensor observations is a critical aspect of planning for active perception. This task is particularly challenging when reasoning about high-level scene understanding using measurements from vision-based neural networks. Due to appearance-based reasoning, the measurements are susceptible to several environmental effects such as the presence of occluders, variations in lighting conditions, and redundancy of information due to similarity in appearance between nearby viewpoints. To address this, we propose a new active perception framework incorporating an arbitrary number of perceptual effects in planning and fusion. Our method models the correlation with the environment by a set of general functions termed *perceptual factors* to construct a perceptual map, which quantifies the aggregated influence of the environment on candidate viewpoints. This information is seamlessly incorporated into the planning and fusion processes by adjusting the uncertainty associated with measurements to weigh their contributions. We evaluate our perceptual maps in a simulated environment that reproduces environmental conditions common in robotics applications. Our results show that, by accounting for environmental effects within our perceptual maps, we improve the state estimation by correctly selecting the viewpoints and considering the measurement noise correctly when affected by environmental factors. We furthermore deploy our approach on a ground robot to showcase its applicability for real-world active perception missions.**

I. Introduction

Scene understanding is a core prerequisite for autonomous robotic planning and decision-making. In various tasks, an important capability is active perception, whereby a robot actively moves a sensor in order to acquire relevant information from the environment [1]. Advances in computer vision and deep learning have substantially improved scene understanding from visual data [2]–[5]. An open challenge is deciding how to interpret and reason about such sensor information, which depends on the environment's appearance, to improve global understanding and active perception [6], [7].

Traditional perception models [8], [9] used for active perception assume that observations are independent given the robot pose and a target state to estimate, e.g., the position or the class of an object. This assumption falls short for complex sensors such as vision-based neural networks [10].

This work was partially funded by Spanish grant FPU20-06563, project T45_23R and project PID2021-125514NB-I00, funded by MCIN/AEI/10.13039/501100011033, by ERDF A way of making Europe and by the European Union NextGenerationEU/PRTR and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 – 390732324. D. Morilla-Cabello and E. Montijano are with the Instituto de Investigación en Ingeniería de Aragón, Universidad de Zaragoza, Spain. J. Westheider and M. Popović are with the Institute of Geodesy and Geoinformation, University of Bonn, Germany. Corresponding: davidmc@unizar.es

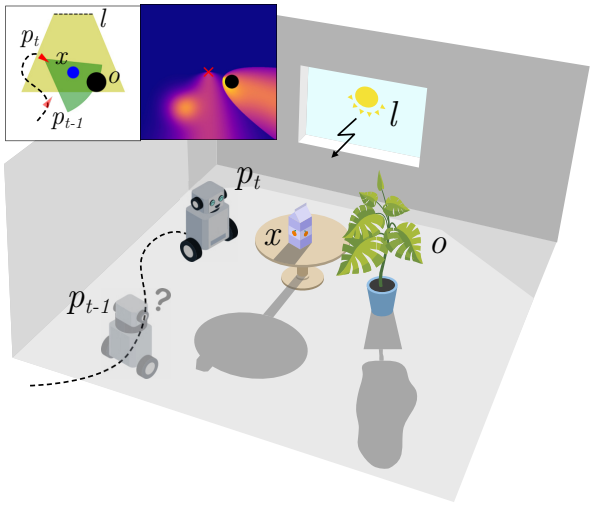


Fig. 1: Overview of the active perception task where a robot moves to poses $\mathbf{p}$ to observe a target object $\mathbf{x}$. A top-down view can be observed in the upper left corner. By integrating perceptual factors such as occluders $\mathbf{o}$, light sources $\mathbf{l}$, and previous poses $\mathbf{p}_{t-1}$, that condition potential observations shown in the heat map, our method enables accurately estimating the utility/value of potential viewpoints (blue is higher), which improves active perception performance

These sensors extract information about visually-based high-level features of the scene. Consequently, they have a strong dependency on environmental characteristics, such as the appearance of objects, partial occlusions or backlighting, and show important spatial correlations due to appearance similarities. Disregarding such correlations spoils the estimation about the information gain of candidate observations, hindering the performance of active perception solutions.

To overcome this limitation, our main contribution is a new perception model that improves the estimation of the information gain provided by candidate viewpoints. This is achieved by modeling an arbitrary number of effects with general functions termed *perceptual factors*.

These functions are combined to obtain a perceptual cost for the viewpoint that effectively weights the estimated information gain by increasing the measurement uncertainty when the information is redundant or affected by adverse environmental effects. The estimation is then used to evaluate the potential new viewpoints in active perception planning in order to acquire the most informative measurements. To showcase the generality of our approach, we also propose examples of perceptual factors relevant to robotic perception.

To validate the approach, we apply our active perception framework in a simulated environment that reproduces envi-

ronmental effects for the tasks of object pose estimation and semantic classification. Finally, in order to demonstrate the practicality of our approach we deploy the active perception framework in a real-world scenario using a ground robot to perform active object classification in the presence of occluders and a light source. The results show how accounting for perceptual factors enhances state estimation quality. This leads to improved consistency in uncertainty and error management, boosting the robustness and efficiency of information gathering in simulated and real scenarios.

II. Related Work

A. Active Perception

Considering how to actively select measurements that provide relevant information has been a long-standing problem in robotics. With a focus on the environment, exploration, and active reconstruction methods often use volumetric gain measurements that favor the observation of new geometry [11]–[13]. Other methods quantify the information gain in a more advanced fashion by including information about the angle between the observations, or surface normals [14], [15]. While these methods reason about how measurements are affected by the environment's appearance, they are exclusively based on geometry.

Traditional active perception approaches often assume measurement independence given the robot pose and the state to estimation to predict potential measurement values [8], [9], [16]. Consequently, the expected information gain of a viewpoint is purely a function of its location and the estimated state [8], [9]. These are still the base for many active perception methods that reason on high-level semantic information [3], [17]. Extending from traditional methods, new perception models have been proposed. For instance, works based on unmanned aerial vehicles (UAVs) [16], [18], [19] typically use empirically determined altitude-dependant sensor model. While these approaches can be practical, by not considering correlations with the environment, they neglect the reduced utility that redundant measurements and measurements affected by other factors have.

Velez et al. [6] first considered how measurements are correlated through the environment and approximated its effects by correlating measurements with previous viewpoints for object detection using Gaussian Processes. Several works stem from this formulation to incorporate planning [20], [21] and new models that include additional information about the viewpoint dependency in neural networks [7], [22]. In the present work, we consider a more general approach where we do not assume that all the influence in the environment is correlated with previous poses. Instead, we allow for more general models that include the influence of variables in the environment external to the object itself such as occluding objects, and light conditions.

B. Environmental modeling

We also consider works that model the environment despite they are not focused on active perception. Some methods apply learning methods to overcome the effects of the environment such as difficult light conditions [10], [23], or to remove occluding objects [24]. Rather than hardly reasoning about inadequate measurements to extract information, we focus on exploiting the positioning ability of robotic platforms to capture better and more informative measurements. Some of these environmental effects are solved by the application of specific solutions, such as in the case of occlusions, which are commonly considered in navigation [25], [26] and manipulation [27], [28]. While we aim to model diverse and complex relations in perception models, we follow a general, synergetic approach that unifies them under the same framework and allows us to integrate them into planning to obtain the most informative viewpoints.

III. Problem Formulation

Consider a robot moving in an environment equipped with a visual sensor to gather data. We denote $\mathbf{p}_t$ as the robot viewpoint at time t. The objective of the robot is to estimate the state of one or more targets, denoted by $\mathbf{x}_t$. This information can be of any kind, e.g., metric (target location) or semantic (target class). To simplify the notation, we assume the target's state does not change over time and drop the dependence on t, but this is not an actual limitation of our problem setup.

The robot has a perception algorithm available, such as a vision-based neural network, used to obtain measurements, denoted $\mathbf{z}_t$, of the target $\mathbf{x}$ at the viewpoints. Differently from classical setups, where the measurements are assumed to be conditionally independent (Fig. 2a), we consider that the value of $\mathbf{z}_t$ also depends on other environmental factors, Ψ (Fig. 2b),

$$\mathbf{z}_t = g(\mathbf{x}, \mathbf{p}_t, \Psi). \tag{1}$$

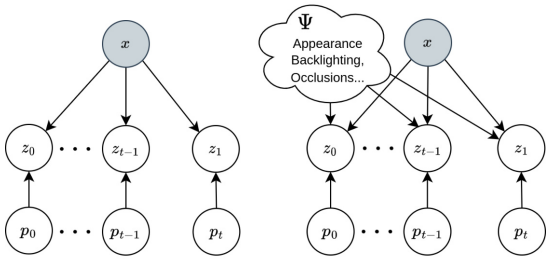
The measurements are used to infer a probability distribution over the target state, $\hat{\mathbf{x}}_t \equiv p(\mathbf{x} \mid \mathbf{z}_{0:t}, \mathbf{p}_{0:t}, \Psi)$.

In this setting, we consider an active perception task. Our goal is to move the robot sensor to maximize the information gathered from the environment. To this end, we leverage an informative path planning (IPP) algorithm, which involves finding the policy $\pi = (\mathbf{p}_1, \dots, \mathbf{p}_N), N > 0$ that optimizes a desired quality criterion, I, over the state estimation,

$$\begin{aligned} \pi^\star = \underset{\pi \in \Pi}{\arg\max}\, & I\big(\hat{\mathbf{x}}_N\big), \\ \text{s.t. } & \mathbf{z}_t = g(\mathbf{x}, \mathbf{p}_t, \Psi), \\ & \hat{\mathbf{x}}_t = h(\hat{\mathbf{x}}_{t-1}, \mathbf{p}_t, \mathbf{z}_t), \\ & C(\pi) \leq B, \end{aligned} \tag{2}$$

where $h(\cdot)$ is the estimation algorithm used for perception, $C(\pi)$ is the policy cost, $B \geq 0$ is the mission budget, e.g., traveled distance, and Π is the set of admissible viewpoints or paths of length N.

Our key contribution is a model for $h(\cdot)$ in Eq. (2) which captures Ψ in a general, simple and efficient way. Our new model enables accurately estimating the information gain associated with potential viewpoint candidates, thereby improving active perception performance.



(a) Traditional (b) General perception model

Fig. 2: Perception models considering different measurement dependencies. Traditional models (a) assume measurements are independent of each other. In reality (b), they are influenced by diverse environmental variables Ψ.

IV. Approach

A. Estimation Model

We restrict our analysis of h to Bayesian recursive estimation algorithms. Considering the effect of Ψ in the recursive estimation problem results in the following:

$$h(\hat{\mathbf{x}}_{t-1}, \mathbf{p}_t, \mathbf{z}_t) = \frac{p(\mathbf{z}_t|\mathbf{x}, \mathbf{p}_t, \Psi)\hat{\mathbf{x}}_{t-1}}{p(\mathbf{z}_t|\mathbf{p}_t, \Psi)}. \tag{3}$$

Accounting for all perceptual properties of the environment and their influence on the measurement likelihood in Eq. (3) is a very complex problem. To address this, we propose a map representation where multiple factors can be efficiently combined to approximate these effects. We then discuss how to use this map in different fusion mechanisms to compute the posterior estimation and provide examples of different factors.

B. Perceptual Maps

To model the influence of Ψ in the potential viewpoints we introduce a set of functions, $f(\mathbf{p}_t, \theta) \in [1, \infty)$, that we call *perceptual factors*, where θ are parameters that can be used to describe different influences. All the perceptual factors can be combined into a joint *perceptual map* that approximates the total influence simply by multiplying them,

$$\hat{\Psi}(\mathbf{p}_t, \theta) := \prod_j f_j(\mathbf{p}_t, \theta_j), \tag{4}$$

where $\theta_j \subseteq \theta$ are the parameters describing each perceptual factor j. Therefore, for each robot pose $\mathbf{p}_t$, the perceptual map function contains the *perceptual cost* associated with that viewpoint. Similar to other kinds of maps, the perceptual map can be built with prior information about the environment or updated online during the mission by including newly discovered information.

In the following, we explain how the perceptual cost can be applied to continuous variables using a standard Extended Kalman Filter (EKF) formulation and a categorical fusion for discrete classification. These examples represent typical robotic perception problems using image processing, i.e., object pose estimation and semantic classification.

1) Extended Kalman Filter (EKF): Assuming Gaussian distributions for all the probabilities in Eq. (3), and recalling that we assume a static state (i.e., no need for prediction), the EKF computes the mean $\hat{\mathbf{x}}_t$ and covariance $\mathbf{P}_t$ of the posterior distribution by

$$\begin{aligned} \mathbf{K}_t &= \mathbf{P}_{t-1}\mathbf{H}_t^T(\mathbf{R}_t + \mathbf{P}_{t-1}\mathbf{H}_t^T)^{-1}, \\ \hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_{t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t-1}), \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t-1}, \end{aligned} \tag{5}$$

where $\mathbf{H}_t$ is the Jacobian of the observation function linearized around the state estimate, and $\mathbf{R}_t$ is the sensor covariance matrix.

Our proposed framework includes the perceptual cost as part of the sensor covariance matrix via multiplication,

$$\mathbf{R}_t^{\Psi} = \mathbf{R}_t\hat{\Psi}(\mathbf{p}_t\,, \theta), \tag{6}$$

where the original value of $\mathbf{R}_t$ is obtained from the sensor calibration in standard conditions. Since $f(\mathbf{p}_t, \theta) \geq 1$, this modification preserves the principles of the EKF, such that $\mathbf{R}_t^{\Psi}$ is a symmetric and positive semi definite matrix, but also encodes potential environment uncertainties. A large value of the perceptual map at that viewpoint increases $\mathbf{R}_t^{\Psi}$, weighing down the effect that the measurement would normally have. Note that our perceptual factors can thus be applied to any existing IPP algorithm based on the EKF [8], [16].

2) Categorical Estimator: Perceptual factors are especially beneficial in deep-learning based semantic vision algorithms since they strongly depend on the measurement fusion process. In the case of semantic classification, $\hat{\mathbf{x}}_t$ and $\mathbf{z}_t$ are described by two probability vectors, where $\hat{x}_{i,t}$ is the probability of $\mathbf{x}$ being of class i (respectively z_i). Similarly to our previous work [29], we model the categorical semantic fusion as a Dirichlet distribution, where the concentration parameter expresses the relative importance between the prior and likelihood distributions to compute the posterior:

$$\hat{x}_{i,t} \propto (\hat{x}_{i,t-1})^{\frac{\bar{\alpha}_{i,t-1}}{\bar{\alpha}_{i,t}}} (z_i)^{\frac{\alpha_{i,t}}{\bar{\alpha}_{i,t}}}, \tag{7}$$

where $\bar{\alpha}_{i,t-1}$ and $\alpha_{i,t}$ are the concentration parameters of the prior and the measurement respectively, and $\bar{\alpha}_{i,t} = \max(\bar{\alpha}_{i,t-1}, \alpha_{i,t})$ is a normalizing term. We define

$$\alpha_{i,t} = \frac{1}{\hat{\Psi}(\mathbf{p}_t\,, \theta)} \in (0, 1], \text{ and } \bar{\alpha}_{i,t} = 1, \forall t, \tag{8}$$

to similarly reduce the influence of measurements with an associated large perceptual cost. Similar to the EKF, the way in which we introduce the perceptual factors in (8) permits using existing planners to solve Eq. (2) [17], [30].

C. Examples of Perceptual Factors

Our formulation offers a general framework that can support different mathematical representations of perceptual factors. Examples include analytically differentiable functions efficient for gradient-based planning or learned functions capturing intricate interactions with the environment. We introduce three hand-crafted examples relevant to robotic

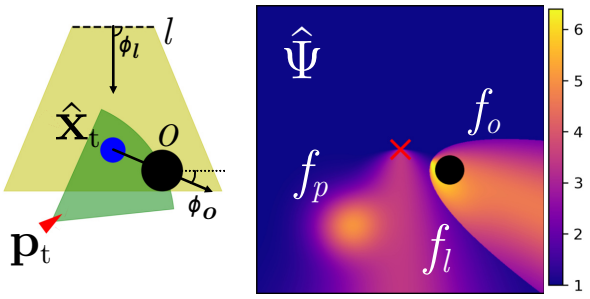


Fig. 3: Perceptual map combining three different perceptual factors. Left: Environment with one target (blue circle and estimation $\hat{\mathbf{x}}_t$), one obstacle ($\mathbf{o}$), one light source (l) and the last viewpoint ($\mathbf{p}_t$). Right: Perceptual map of the target combining the three factors: occlusions (f_o), backlighting (f_l) and past viewpoints (f_p). Larger values in the map identify less informative viewpoints.

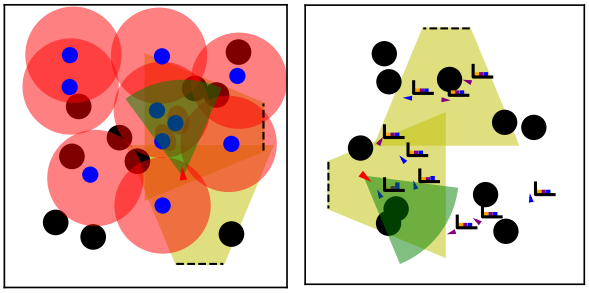


Fig. 4: Examples of the *metric* (left) and *semantic* (right) simulated experiments. A robot (red) moves to observe targets (blue). Occluders (black) and lights (yellow) may add noise to the measurements. The estimation covariance is shown in red for the *metric* experiment. The classification histogram is shown for the *semantic* experiment. The supplementary video shows some executions.

applications that model direct effects of the scene and redundancy due to local appearance.

Partial occlusions and backlighting are two environment elements that have a strong influence on the outcome of deep-learning vision algorithms. To keep computation simple and efficient, we model both factors with bounded quadratic, functions translated and rotated, with parameters $\theta = \{\delta, w\}$ defining the strength of the influence and the width of the parabola respectively.

For a 2D environment, given an obstacle located at $\mathbf{o} = (o_x, o_y)$, which position might change over time, the resulting factor is

$$
\begin{aligned}
f_o(\mathbf{p}_t, \mathbf{o}) &= \begin{cases} 1 + \delta \exp \frac{-x'^2}{y'} & \text{if } \frac{x'^2}{\sigma} < y', \\ 1 & \text{otherwise} \end{cases} \\
x' &= (p_x - o_x)\cos(\phi) - (p_y - o_y)\sin(\phi), \\
y' &= (p_x - o_x)\sin(\phi) + (p_y - o_y)\cos(\phi),
\end{aligned} \tag{9}
$$

where (p_x, p_y) is the position in $\mathbf{p}_t$ and $\phi = \arctan(\hat{y} - o_y, \hat{x} - o_x)$ with $(\hat{x}, \hat{y})$ the target position. The width is equal to the sum of the obstacle and target sizes in order to free the vision and avoid partial occlusions. The 3D case can be computed in a similar fashion.

The backlighting generated by an directional light source is very similar. To compute $f_l(\mathbf{p}_t, \theta)$ we follow the same computations as in Eq. (9), changing the origin of the parabola to the target, $\hat{\mathbf{x}}_t$, and defining ϕ as the direction of the light. The width, in this case, will depend on the light diffusion, the more concentrated the light source is in its direction, the narrower its effect.

The fact that measurements are captured by cameras from the scene makes the information of nearby viewpoints redundant due to similar appearance and does not contribute to improving the state estimate. We introduce a second type of perceptual factor that aims to reduce the influence of subsequent measurements captured at similar viewpoints:

$$
f_r = 1 + \delta \exp\left(\frac{-\|\mathbf{p}_t - \mathbf{p}_i\|_2^2}{2\sigma^2}\right). \tag{10}
$$

where $\mathbf{p}_i \in \{\mathbf{p}_0, \ldots, \mathbf{p}_{t-1}\}$ are previous viewpoints.

As an illustrative example, Fig. 3 combines the presented factors into a perceptual cost map, which corresponds to the physical environment shown in Fig. 1.

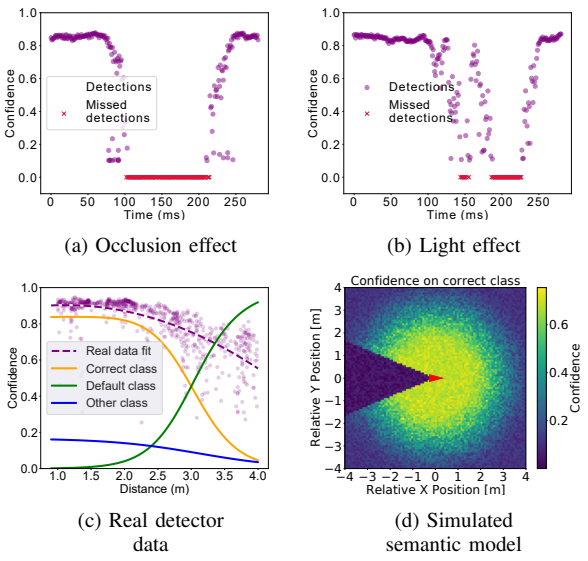
V. Experimental Results

A. Simulated Experiments

Setup. We evaluate our method in two active perception tasks: pose estimation, termed *metric*, and semantic classification, termed *semantic*. We consider 50 randomly generated 2D $10m \times 10m$ environments, with 10 targets to localize/classify, 10 occluders, and 2 light sources. Fig. 4 show examples of these environments. To simulate measurements that resemble the actual output of vision algorithms affected by environmental factors, we execute a characterization of the noise from a pre-trained object detector in different controlled experiments. Fig. 5a shows measurement confidences and missed detections when the target is incrementally occluded by a moving object. Similarly, Fig. 5b exemplifies a light source generating backlighting when moving behind the object. Based on the captured data, we model our simulator to add noise to the measurements depending on which, and how many factors affect the measurement. Each partial occlusion adds noise incrementally with the occlusion amount. Finally, lights add a fixed increment when the target is between the light and the robot.

For the *metric* problem, the robot is equipped with a range-bearing sensor. The sensor returns measurements corrupted with Gaussian noise with standard deviation for distance $\sigma_d = 0.3$ and bearing $\sigma_b = 0.1$. The standard deviation is scaled upon the effect of perceptual factors by $\gamma = [1, 6]$. The noise also increases the probability of a missed detection.

For the *semantic* experiment, the sensor returns classification confidences for each measured target. In this case, we also characterize the confidence response from the real object detector. According to Fig. 5c, we model the distance effect by decaying confidence in the detected class and increasing noise, which results in outliers. We also add a region where the object is classified as the wrong class, resulting in the model shown in Fig. 5d. Similarly to the *metric* case, we include the perceptual factors' noise increment in the confidence noise and the probability of outliers.



(a) Occlusion effect (b) Light effect

(c) Real detector data (d) Simulated semantic model

Fig. 5: Characterization of real-world perception models used in our simulator. We capture real detections from a classifier (in purple) in order to model the effects of (a) occlusions and (b) backlighting. The capturing process is included in the supplementary video. We use the response to semantic detections to model (c) simulated confidence. Together with noise, and a region where the target is misclassified, we build (d) a simulated semantic model.

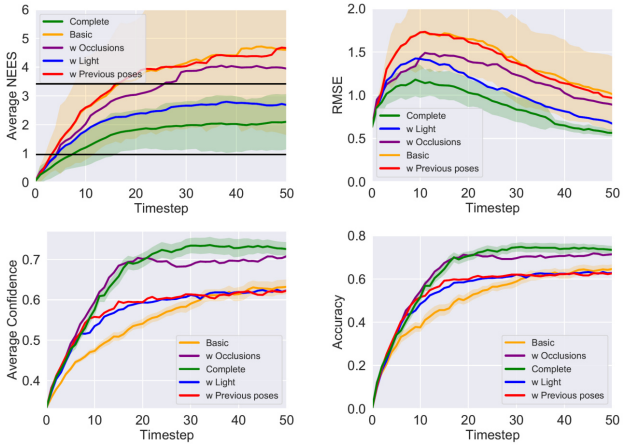


Fig. 6: Averaged results (50 experiments) for the *metric* (left) and *semantic* (right) experiments. Introducing perceptual factors reduces average NEES and RMSE. Additionally, by accounting for the noise induced by perceptual factors, our method stays within the confidence interval for a significance level of 0.05 (horizontal black lines), thus ensuring consistency. Besides, we increase average confidence and accuracy by reducing the influence of outliers.

Baselines. To carry out the perception tasks, we use Eq. (2) to decide which next viewpoints for the robot to take measurements. We consider the entropy of $\hat{\mathbf{x}}_t$ as the quality criterion, $I(\hat{\mathbf{x}}_t)$, the Euclidean distance for $C(\pi)$, the EKF in Eq. (5) as estimation model for the *metric* task and the categorical estimator of Eq. (7) for the *semantic* task. We consider a planning horizon of $N = 1$, use random sampling to generate 100 candidate viewpoints with $B = 2m$ and 2π rad for the orientation. Then, we choose the best candidate viewpoint in terms of $I(\hat{\mathbf{x}}_t)$ and move the robot to acquire a new measurement there. We repeat this process 50 times in each simulation.

The IPP algorithm is designed in a simple way to emphasize the analysis of the perceptual factors, the core contribution of our paper. We compare different versions of $\hat{\Psi}(\mathbf{p}_t, \theta)$ for the perceptual factors. The *Basic* algorithm is the baseline in which perceptual factors are neglected, i.e., $\hat{\Psi} = 1$ everywhere. We also analyze each of the three proposed factors separately (*Occlusions*, *Light* and *Previous Poses*), as ablation versions of our *Complete* method. The perceptual factors are accounted in Eq. (6) and Eq. (8) for the *metric* and *semantic* experiments respectively. For modeling these factors, we use the proposed functions in Sec. IV-C. The influence values and widths for occluders, lights, and correlation with previous poses are $\delta_o = 3$, $\delta_l = 2$, $w_l = 3$, $\delta_p = 3$, and $\sigma_p = 0.1$ respectively.

Evaluation metrics. Since the analyzed methods have different measurement models, in terms of their uncertainties, the amount of information gain of each method, $I(\hat{\mathbf{x}}_t)$, is not a representative metric to compare them. Moreover, entropy reduction does not necessarily imply a good estimation in the presence of very noisy measurements. Instead, our evaluation focuses on the consistency of the estimation, understood as how the uncertainty represents the true error of the estimator.

In the case of the EKF, we use the Normalized Estimated Error Squared (NEES) averaged over all targets, defined as

$$NEES = (\hat{\mathbf{x}}_t - \mathbf{x})^T \mathbf{P}_t^{-1} (\hat{\mathbf{x}}_t - \mathbf{x}). \tag{11}$$

NEES values closer to the target dimension (2 in our experiments) indicate better filter estimation in terms of consistency. We also report Root Mean Squared Error (RMSE) of the position estimation with respect to ground truth.

In the case of the categorical semantic fusion, we evaluate the average confidence of the true class, which represents the confidence of the posterior and also encodes the concept of consistency over the prediction. Additionally, we report the overall accuracy of the classification as the percentage of correctly classified targets.

Results. As illustrated in Fig. 6, the consistency of the estimation for the basic method increases during the mission execution beyond accepted levels of confidence. This is primarily attributed to noisy measurements resulting from partial occlusions and lighting conditions. Our method addresses these challenges via two significant enhancements. First, it selects measurement points that avoid these effects, as shown in Fig. 7. Second, our method reduces the confidence of noisy measurements during the fusion process. This diminishes their influence and reduces the estimated uncertainty, thereby ensuring consistent stability around the desired NEES value, which is 2. These effects are also reflected in the RMSE graph, where our method effectively minimizes estimation errors by mitigating the impact of outliers on state estimation.

For the *semantic* problem, by including all the perceptual factors we obtain more confident measurements and a better accuracy in the classification (Fig. 6). Similarly, the number of bad measurements, affected by occlusions or light is lower

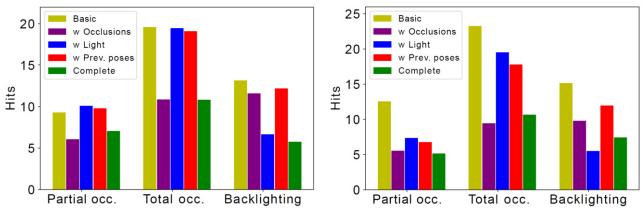


Fig. 7: Numbers of perceptual factors hit in the (left) *metric* and (right) *semantic* experiments. When we introduce perceptual factors, the planner estimates lower information value in the viewpoints they affect, guiding the robot to avoid these locations.

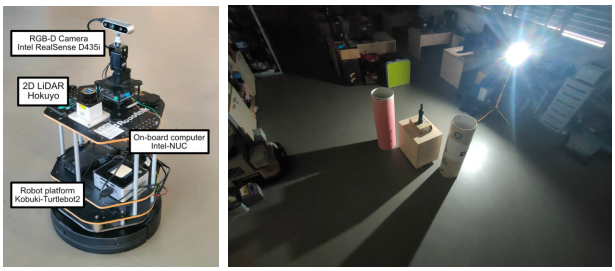


Fig. 8: Real experiment setup. (Left) Ground robot platform. (Right) Working environment with occluders and a directional light. The robot moves to obtain measurements and decide whether there is an object at the target position.

than the baseline (Fig. 7).

B. Real-World Experiments

The aim of this experiment is to showcase the applicability of our approach in real-world scenarios. Fig. 8 shows our experimental setup. Our ground platform is a Turtlebot2 equipped with a Hokuyo 2D LiDAR to allow navigation in an occupancy map using the ROS *navigation stack*. The robot is equipped with an Intel RealSene D435i RGB-D camera to perform object detection using the pre-trained YOLOv5 model from Ultralytics [31]. All the processing is performed onboard the robot on an Intel NUC computer. The experiments are conducted in the room where there are two obstacles and a strong directional light.

The goal of the mission is to decide if there is a *bottle* or *no object* in a known position by taking multiple measurements. For that, the robot uses the active perception planner presented in Sec. V-A to select viewpoints. Then, it navigates to them and captures an image. If the target object is detected by the neural network in the image, the detection confidence is integrated into the belief. Otherwise, a *no object* observation is integrated with confidence 0.75. When the confidence for *bottle* or *no object* are > 0.99, the mission ends. We compare the performance of two active planners, with and without all the perceptual factors described in the paper. Different from the simulation, candidate viewpoints are chosen from a deterministic set instead of using random sampling.

Fig. 9 shows the result for the first viewpoint estimation values and the subsequent poses for the two planners. After four poses, due to the effect of occlusions and backlighting, the baseline only reaches an object detection confidence of 0.62 and gets stuck moving between two viewpoints. In contrast, our planner obtains successful detections for all poses, integrating confidence over 0.99. This shows

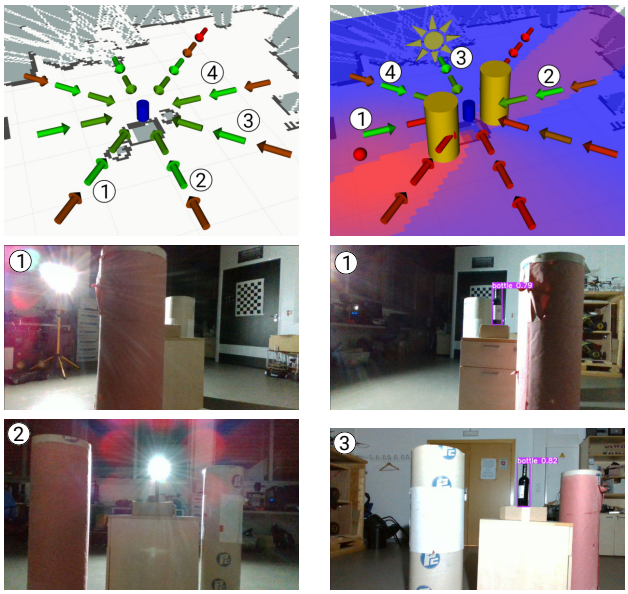


Fig. 9: Real experiment results. (Left) Basic IPP without perceptual factors. (Right) IPP considering perceptual factors for occluders (yellow cylinders), light sources (sun icon), and previous poses (red sphere), which contribute to the perceptual map projected on the floor. The arrows on the top figures show the candidate viewpoints to observe the blue target with colors indicating their estimated information value for the first step (green is better). The numbers indicate viewpoint selection order during the mission. The baseline leverages only a distance-based model for planning, while our approach includes perceptual information. As a result, the baseline obtains wrong measurements from the neural network (left 1 and 2) affected by occlusion (top) and backlighting (bottom). In contrast, our planner selects viewpoints that ensure good detections (right 1 and 3), resulting in better mission performance.

the benefit of using our perceptual factors for informative viewpoint selection by considering correlations with the environment. We refer the reader to the supplementary video for a complete visualization of the experiment.

VI. Conclusion and Future Work

This paper presented a new perception model to characterize the impact that different environmental factors have on high-level measurements, like those provided by deep-learning vision algorithms. To achieve this, we propose using perceptual factors, which are general functions used to quantify the aggregated influence of the environment on future measurements. We demonstrated the integration of our perceptual factors into well-known recursive estimation filters and devised an active perception framework for informative viewpoint selection. Further, we propose concrete mathematical examples of perceptual factors relevant to robotics scenarios. We evaluated our complete active perception pipeline in simulation for robotic object localization and semantic classification tasks. By incorporating perceptual factors, we obtain consistent state estimation and robust, informative viewpoint selection for predictive planning in active perception. Potential avenues for future work include using deep learning to model perceptual factors and incorporating additional sensor modalities.

References

[1] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[2] S. Papatheodorou, N. Funk, D. Tzoumanikas, C. Choi, B. Xu, and S. Leutenegger, "Finding things in the unknown: Semantic object-centric exploration with an mav," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2023, pp. 3339–3345.

[3] X. Liu, A. Prabhu, F. Cladera, I. D. Miller, L. Zhou, C. J. Taylor, and V. Kumar, "Active metric-semantic mapping by multiple aerial robots," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2023, pp. 3282–3288.

[4] M. Dharmadhikari and K. Alexis, "Semantics-aware exploration and inspection path planning," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2023, pp. 3360–3367.

[5] T. Dang, C. Papachristos, and K. Alexis, "Autonomous exploration and simultaneous object search using aerial robots," in *Proc. of the IEEE Aerospace Conference*, 2018, pp. 1–7.

[6] J. Velez, G. Hemann, A. S. Huang, I. Posner, and N. Roy, "Modelling observation correlations for active exploration and robust object detection," *Journal of Artificial Intelligence Research (JAIR)*, vol. 44, pp. 423–453, 2012.

[7] V. Tchuiev and V. Indelman, "Inference over distribution of posterior class probabilities for reliable bayesian classification and object-level perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4329–4336, 2018.

[8] Y. Kantaros, B. Schlotfeldt, N. Atanasov, and G. J. Pappas, "Sampling-based planning for non-myopic multi-robot information gathering," *Autonomous Robots*, vol. 45, no. 7, pp. 1029–1046, 2021.

[9] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, and R. Fitch, "Dec-mcts: Decentralized planning for multi-robot active perception," *Intl. Journal of Robotics Research*, vol. 38, no. 2-3, pp. 316–337, 2019.

[10] Z.-F. Xu, R.-S. Jia, Y.-B. Liu, C.-Y. Zhao, and H.-M. Sun, "Fast method of detecting tomatoes in a complex scene for picking robots," *IEEE Access*, vol. 8, pp. 55 289–55 299, 2020.

[11] J. Delmerico, S. Isler, R. Sabzevari, and D. Scaramuzza, "A comparison of volumetric information gain metrics for active 3d object reconstruction," *Autonomous Robots*, vol. 42, no. 2, pp. 197–208, 2018.

[12] S. Song, D. Kim, and S. Choi, "View path planning via online multiview stereo for 3-d modeling of large-scale structures," *IEEE Trans. on Robotics*, vol. 38, no. 1, pp. 372–390, 2021.

[13] L. Schmid, M. Pantic, R. Khanna, L. Ott, R. Siegwart, and J. Nieto, "An efficient sampling-based method for online informative path planning in unknown environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1500–1507, 2020.

[14] B. Hepp, M. Nießner, and O. Hilliges, "Plan3d: Viewpoint and trajectory optimization for aerial multi-view stereo reconstruction," *ACM Trans. on Graphics*, vol. 38, no. 1, pp. 1–17, 2018.

[15] D. Morilla-Cabello, L. Bartolomei, L. Teixeira, E. Montijano, and M. Chli, "Sweep-your-map: Efficient coverage planning for aerial teams in large-scale environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 810–10 817, 2022.

[16] M. Popović, T. Vidal-Calleja, G. Hitz, J. J. Chung, I. Sa, R. Siegwart, and J. Nieto, "An informative path planning framework for uav-based terrain monitoring," *Autonomous Robots*, vol. 44, pp. 889–911, 2020.

[17] A. Asgharivaskasi and N. Atanasov, "Semantic octree mapping and shannon mutual information computation for robot exploration," *IEEE Trans. on Robotics*, vol. 39, no. 3, pp. 1910–1928, 2023.

[18] L. Qingqing, J. Taipalmaa, J. P. Queralta, T. N. Gia, M. Gabbouj, H. Tenhunen, J. Raitoharju, and T. Westerlund, "Towards Active Vision with UAVs in Marine Search and Rescue: Analyzing Human Detection at Variable Altitudes," in *Proc. of the IEEE Intl. Symposium on Safety, Security, and Rescue Robotics*, 2020, pp. 65–70.

[19] A. A. Meera, M. Popović, A. Millane, and R. Siegwart, "Obstacle-aware Adaptive Informative Path Planning for UAV-based Target Search," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2019, pp. 718–724.

[20] W. L. Teacy, S. Julier, R. De Nardi, A. Rogers, and N. R. Jennings, "Observation modelling for vision-based target search by unmanned aerial vehicles," in *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2015, p. 1607–1614.

[21] V. Tchuiev and V. Indelman, "Epistemic uncertainty aware semantic localization and mapping for inference and belief space planning," *Artificial Intelligence*, vol. 319, 2023.

[22] Y. Feldman and V. Indelman, "Bayesian viewpoint-dependent robust classification under model and localization uncertainty," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2018, pp. 3221–3228.

[23] A. Kuznetsova, T. Maleva, and V. Soloviev, "Using YOLOv3 algorithm with pre-and post-processing for apple detection in fruit-harvesting robot," *Agronomy*, vol. 10, no. 7, 2020.

[24] R. Cheng, A. Agarwal, and K. Fragkiadaki, "Reinforcement learning of active vision for manipulating objects under occlusions," in *Proc. of the Conf. on Robot Learning (CoRL)*, vol. 87, 2018, pp. 422–431.

[25] K. Schlegel, P. Weissig, and P. Protzel, "A blind-spot-aware optimization-based planner for safe robot navigation," in *Proc. of the Europ. Conf. on Mobile Robotics (ECMR)*, 2021.

[26] B. Gilhuly, A. Sadeghi, P. Yedmellat, K. Rezaee, and S. L. Smith, "Looking for trouble: Informative planning for safe trajectories with occlusions," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2022.

[27] D. Morrison, P. Corke, and J. Leitner, "Multi-view picking: Next-best-view reaching for improved grasping in clutter," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2019, pp. 8762–8768.

[28] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz, "Nbv-sc: Next best view planning based on shape completion based on fruit mapping and reconstruction," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2023.

[29] D. Morilla-Cabello, L. Mur-Labadia, R. Martinez-Cantin, and E. Montijano, "Robust Fusion for Bayesian Semantic Mapping," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2023.

[30] A. Serra-Gómez, E. Montijano, W. Böhmer, and J. Alonso-Mora, "Active classification of moving targets with learned control policies," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3717–3724, 2023.

[31] G. Jocher, "Yolov5 by ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5