Efficient Greedy Algorithms for Feature Selection in Robot Visual Localization

Vivek Pandey, Amirhossein Mollaei, and Nader Motee

***Abstract*—Robot localization is a fundamental component of autonomous navigation in unknown environments. Among various sensing modalities, visual input from cameras plays a central role, enabling robots to estimate their position by tracking point features across image frames. However, image frames often contain a large number of features, many of which are redundant or uninformative for localization. Processing all features can introduce significant computational latency and inefficiency. This motivates the need for intelligent feature selection—identifying a subset of features that are most informative for localization over a prediction horizon. In this work, we propose two fast and memory-efficient feature selection algorithms that enable robots to actively evaluate the utility of visual features in real time. Unlike existing approaches with high computational and memory demands, the proposed methods are explicitly designed to reduce both time and memory complexity while achieving a favorable trade-off between computational efficiency and localization accuracy.**

I. Introduction

Navigation is a fundamental aspect of robotics, essential for enabling robots to autonomously perform a wide range of tasks. Effective navigation relies on continuous and accurate position estimation, commonly referred to as localization. To achieve reliable and precise localization, robots employ advanced probabilistic estimation techniques such as the Kalman filter [20], particle filter [18, 19], and information filter [21]. These algorithms integrate sensor measurements with predictive motion models to iteratively refine and update the robot's position estimates.

In the context of perception-based navigation, robots employ sensors such as cameras and LiDAR for localization. The large volume of features captured by these sensors makes the estimation process computationally demanding. Consequently, it is crucial to efficiently select a smaller subset of the most informative features to optimize localization performance. Extensive research has been devoted to this problem, focusing on developing efficient algorithms for perception-based feature selection [12, 7, 6, 10].

In [3], the authors use convex relaxation and a greedy strategy to select the most informative features. Zhao *et al.* [24] propose a lazier-than-lazy greedy algorithm to address the feature selection problem by maximizing the log-determinant as an information measure for optimizing Visual SLAM performance. Zhang *et al.* [23] approach Visual SLAM from a system-theoretic perspective, defining observability scores to quantify feature informativeness and employing a greedy algorithm for feature selection.

In [14], the authors propose an efficient randomized algorithm for multi-agent localization that offers probabilistic performance guarantees. It is well established that subset selection is a combinatorial optimization problem. To address this challenge, several algorithmic frameworks have been explored, including convex relaxation [8] and greedy methods [13]. In this work, we focus on greedy algorithms, which provide practical and computationally efficient approaches for obtaining near-optimal solutions to submodular maximization problems.

V. Pandey, A. Mollaei and N. Motee are with the Department of Mechanical Engineering and Mechanics, Lehigh University. {vkp219,ammb23,motee}@lehigh.edu.

A. Greedy Algorithms for Submodular Maximization

Greedy algorithms have long been recognized for their effectiveness in addressing combinatorial optimization problems, particularly those involving submodular maximization. The seminal work of Nemhauser *et al.* [13] established fundamental performance guarantees for greedy algorithms applied to monotone submodular functions, demonstrating that such algorithms can achieve a $(1 - 1/e)$-approximation of the optimal solution.

Greedy algorithms have since been widely adopted across diverse domains, including sensor scheduling [17], feature selection [3], and anchor selection in SLAM [4], among others. Recent advancements have focused on improving the computational and memory efficiency of greedy methods, thereby enhancing both their theoretical guarantees and practical applicability in resource-constrained settings [11, 2, 9].

Building on these developments, we leverage modern variants of greedy algorithms to address the feature selection problem in robot visual localization, aiming to balance computational efficiency with localization accuracy.

B. Our Contribution

In this paper, we address the feature selection problem under computational and memory constraints commonly encountered in robotic systems. Our main contributions are summarized as follows:

- **Stochastic Greedy Feature Selection:** We propose a stochastic greedy algorithm that significantly reduces the time complexity of feature selection while maintaining near-optimal performance. This approach enables efficient operation in real-time scenarios without substantial loss in accuracy.
- **Approximate Greedy Feature Selection:** We develop a mathematically justified approximate greedy algorithm that simultaneously improves both computational and memory efficiency. This formulation leverages tractable surrogates for information-based metrics to further accelerate the selection process while preserving theoretical soundness.

Our contributions are theoretical and analytical in nature, aimed at improving memory efficiency and fast computation. Some of the results in this paper have been developed independently in [22].

C. Mathematical Notation

Vectors and matrices are denoted by lowercase and uppercase letters, respectively. Sequences of vectors or matrices stacked over time are represented by boldface lowercase and uppercase letters. The cones of symmetric positive semidefinite and positive definite $n \times n$ matrices are denoted by S_+^n and S_{++}^n. For a matrix X, $X^\top$ denotes its transpose. The special orthogonal group in $\mathbb{R}^3$ is $\mathrm{SO}(3)$, the cardinality of a set A is $|A|$, and $\mathrm{Tr}(\cdot)$ denotes the trace. $I_{(\cdot)}$ represents the identity matrix, with size inferred from context. For $X_1, X_2 \in S_{++}^n$, $X_2 \preceq X_1$ if $X_1 - X_2 \in S_+^n$. A Gaussian random vector z with mean μ and covariance Σ is $z \sim \mathcal{N}(\mu, \Sigma)$. The sets of nonnegative integers and reals are denoted by $\mathbb{Z}_+$ and $\mathbb{R}_+$, respectively.

II. Problem Statement

We address the problem of sparse feature selection to estimate the positions of a robot navigating in an unknown environment. Using efficient greedy algorithms, we select a set of the most informative features by anticipating their importance over a fixed time horizon.

Let $\boldsymbol{x}_t \in \mathbb{R}^3$ be the position vector of robot at time $t \in \mathbb{Z}_+$. Consider the vector of discrete time horizon given by $[t : t+M] = [t, t+1, \cdots, t+M]$. Then vector $\mathbf{x}_{t:t+M}$ which contains states of the robot for all $t \in [t : t+M]$ can be written as

$$\mathbf{x}_{t:t+M} = \left[x_t^T, x_{t+1}^T, \cdots, x_{t+M}^T\right]^T \in \mathbb{R}^{3N(M+1)}. \quad (1)$$

Let Θ_t represent the set of all features f available at a given time step t. Furthermore, let the matrix $\mathbf{H}_{t:t+M}^f$ be the contribution of new features to the information matrix, then the information matrix is updated as follows

$$\mathbf{H}_{t:t+M}(\Theta_t) = \bar{\mathbf{H}}_{t:t+M} + \sum_{f \in \Theta_t} \mathbf{H}_{t:t+M}^f. \quad (2)$$

Given the constraints on onboard computational resources, it is essential for robot to select a smaller subset of features Φ_t from the available set Θ_t, which prioritizes the most informative features, maximizing the contribution to the information gain while minimizing the computational complexity. The feature selection problem can be formulated as the following optimization problem:

$$\Phi_t^* = \underset{\Phi_t \subseteq \Theta_t : |\Phi_t| \leq q}{\operatorname{argmax}} \rho\left(\Phi_t\right), \quad (3)$$

where the goal is to maximize a non-negative scalar function ρ over the subsets of size q.

The rest of the paper is organized as follows. Section III presents the models for robot motion and the vision system. We formulate the feature selection problem as optimal experiment design problem in Section IV. In Section V, we formally introduce our algorithms, providing a detailed description along with theoretical guarantees and analysis. Section VI concludes the paper with a discussion of future research directions. The proofs of all theoretical results are provided in the Appendix.

III. Model for Robot Motion and Vision System

To achieve accurate position estimates for the robot, the robot continuously updates the initial estimates obtained from the robot dynamics using information gathered from sensor measurements. This section begins by introducing the robot motion model, which mathematically describes how the robot positions evolve over time. This model lays the foundation for incorporating measurement models, which will be used to refine the position estimates based on sensor data.

A. Model for Robots Motion

We consider that dynamics of the robot's motion is governed by the following linear dynamical model:

$$x_\tau = Ax_{\tau-1} + Bu_\tau + \delta_\tau, \quad (4)$$

for all $\tau \in [t : t+M]$, where x_τ is the state (position vector) of the robot at time τ, u_τ is the control input of the robot at time τ and $\delta_\tau \sim \mathcal{N}(0, \Lambda_\tau)$ such that $\mathbb{E}[\delta_{\tau_1}\delta_{\tau_2}] = 0$ for all $\tau_1 \neq \tau_2$.

Let $\bar{\boldsymbol{\mu}}_t = \mathbb{E}[\boldsymbol{x}_\tau]$ and $\bar{\boldsymbol{\Sigma}}_t = \mathbb{E}[(\boldsymbol{x}_\tau - \bar{\boldsymbol{\mu}}_t)(\boldsymbol{x}_\tau - \bar{\boldsymbol{\mu}}_t)^T]$, then the mean and covariance of $\mathbf{x}_{t:t+M}$ can be expressed as

$$\bar{\boldsymbol{\mu}}_{t:t+M} = \begin{bmatrix} \bar{\boldsymbol{\mu}}_t & \bar{\boldsymbol{\mu}}_{t+1} & \cdots & \bar{\boldsymbol{\mu}}_{t+M} \end{bmatrix} \quad (5)$$

$$\bar{\boldsymbol{\Sigma}}_{t:t+M} = \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_t & \bar{\boldsymbol{\Sigma}}_{t,t+1} & \cdots & \bar{\boldsymbol{\Sigma}}_{t,t+M} \\ \bar{\boldsymbol{\Sigma}}_{t,t+1}^T & \bar{\boldsymbol{\Sigma}}_{t+1} & \cdots & \bar{\boldsymbol{\Sigma}}_{t+1,t+M} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\boldsymbol{\Sigma}}_{t,t+M}^T & \bar{\boldsymbol{\Sigma}}_{t+1,t+M}^T & \cdots & \bar{\boldsymbol{\Sigma}}_{t+M} \end{bmatrix} \quad (6)$$

where $\bar{\boldsymbol{\Sigma}}_\tau = \boldsymbol{A}\bar{\boldsymbol{\Sigma}}_{\tau-1}\boldsymbol{A}^T + I_N \otimes \Lambda_\tau$ for all $\tau \in [t : t+M]$ and

$$\bar{\boldsymbol{\Sigma}}_{\tau_1,\tau_2} = \boldsymbol{A}^{(\tau_2-\tau_1)}\bar{\boldsymbol{\Sigma}}_{\tau_1}$$

for all $\tau_1, \tau_2 \in [t : t+M]$ with $\tau_1 < \tau_2$.

The information matrix of $\mathbf{x}_{t:t+M}$ using the dynamical model can be written as

$$\bar{\mathbf{H}}_{t:t+M} = \bar{\boldsymbol{\Sigma}}_{t:t+M}^{-1}. \quad (7)$$

The information matrix $\bar{\mathbf{H}}_{t:t+M}$ is a measure of information content of $\mathbf{x}_{t:t+M}$ when the system dynamics is predicted over the fixed time horizon M. The following subsections delve into the process of sensor fusion. Here, we explore how information matrices, derived from the robot motion model, are dynamically updated by incorporating measurements from the robot's sensors.

B. Model for Vision System

Let us denote the set of all features available at any time step by Θ_t, the rotation matrices describing the orientation of the camera and robot by $R_c \in \mathrm{SO}(3)$ and $R_\tau \in \mathrm{SO}(3)$ respectively, the position vector of the feature $f \in \Theta_t$ by $y_f \in \mathbb{R}^3$, the position vector of the camera with respect to robot by x_c, and the unit vector (with respect to camera) corresponding to pixel measurement of feature $f \in \Theta_t$ at time τ by $(u_\tau^f)^T \in \mathbb{R}^3$. For a given vector $(u_\tau^f)^T$, its cross product with another vector v can be written as the product of a skew-symmetric matrix U_τ^f and vector v as

$$(u_\tau^f)^T \times v = U_\tau^f v.$$

Then the noisy vision model for the robot proposed in [3] is given by

$$U_\tau^f (R_\tau R_c)^T \left(y_f - (x_\tau + R_\tau x_c)\right) = \eta_\tau^f. \tag{8}$$

The model in (8) can be written equivalently as

$$z_\tau^f = U_\tau^f (R_\tau^i R_c)^T \left(x_\tau - y_f\right) + \eta_\tau^f, \tag{9}$$

where $\eta_{\tau,T}^f$ is the measurement noise such that $\eta_{\tau,T}^f \sim \mathcal{N}(0, \sigma^2 I_3)$ and $z_\tau^f = (U_\tau^f)^T (R_c)^T x_c$.

The vision model for the team of robots can be written by stacking the model (9) for all robots as

$$z_\tau^f = F_\tau^f \mathbf{x}_{t:t+M} + E_\tau^f y_f + \eta_\tau^f, \tag{10}$$

for appropriate matrices F_τ^f and E_τ^f.

We assume that robots are capable of running $M-$step forward simulations to determine the number of frames n_f in which the feature $f \in \Theta_t$ is visible. The overall $M-$step vision model for a given feature f can be written by vertically stacking (10) as

$$\mathbf{z}_{t:t+M}^f = \mathbf{F}_{t:t+M}^f \mathbf{x}_{t:t+M} + \mathbf{E}_{t:t+M}^f y_f + \boldsymbol{\eta}_{t:t+M}^f. \tag{11}$$

Let us denote the covariance matrix of $\boldsymbol{\eta}_{t:t+M}^f$ by

$$\mathbb{E}\left[\boldsymbol{\eta}_{t:t+M}^f \left(\boldsymbol{\eta}_{t:t+M}^f\right)^T\right] = \sigma^2 I_{3n_f} \in S_+^{3n_f}. \tag{12}$$

The information matrix of the parameters $\mathbf{x}_{t:t+M}$ and y_f is given by

$$\Omega_\star^f = \sigma^{-2} \begin{bmatrix} \left(\mathbf{F}_\star^f\right)^T \mathbf{F}_\star^f & \left(\mathbf{F}_\star^f\right)^T \mathbf{E}_\star^f \\ \left(\mathbf{E}_\star^f\right)^T \mathbf{F}_\star^f & \left(\mathbf{E}_\star^f\right)^T \mathbf{E}_\star^f \end{bmatrix}, \tag{13}$$

where $\star = t : t+M$ is introduced for simplicity of notation. The information matrix corresponding to the parameters $\mathbf{x}_{t:t+M}$ is obtained by taking the Schur complement of the block $\left(\mathbf{E}_\star^f\right)^T \mathbf{E}_\star^f$ and can be written as

$$\mathbf{H}_\star^f = \sigma_i^{-2} \left(\left(\mathbf{F}_\star^f\right)^T \mathbf{F}_\star^f - \left(\mathbf{F}_\star^f\right)^T \mathbf{E}_\star^f \left(\left(\mathbf{E}_\star^f\right)^T \mathbf{E}_\star^f \right)^{-1} \left(\mathbf{E}_\star^f\right)^T \mathbf{F}_\star^f \right). \tag{14}$$

The information matrix $\mathbf{H}_\star^f = \mathbf{H}_{t:t+M}^f$ specified in (A) is a measure of amount of information contained by a feature f for estimation of $\mathbf{x}_{t:t+M}$.

IV. Theory of Optimal Experiment Design

We forumlate the feature selection problem (3) for robot's localization as parameter estimation problem given a set of observations. Specifically, for estimating robot's position given n visual features, we choose the features that decreases the covraince of robot's position the most. Using the Cramer-Rao bound [16, 5]

$$\text{var}\left(\hat{\theta}\right) \succeq I(\theta)^{-1}, \tag{15}$$

which states that variance in the eastimated parameter $\hat{\theta}$ is lower bounded by information matrix $I(\theta)$. Equivalently, to minimize a scalar measure of covaraince matrix, theory of optimal design of experiments suggests maximixing the scalar measure of information matrix [15]. For any new observations by tracking a feature f, the information matrix of $\mathbf{x}_{t:t+M}$, is updated as

$$\mathbf{H}_\star\left(\{f\}\right) = \tilde{\mathbf{H}}_\star + \mathbf{H}_\star^f. \tag{16}$$

TABLE I: Performance Measures

Performance Measures	Matrix Operator Form
Variance of the Error $\rho_v(\mathbf{H}_\star(\Phi_t))$	$\text{Tr}\left(\mathbf{H}_\star\left(\Phi_t\right)^{-1}\right)$
Differential Entropy $\rho_e(\mathbf{H}_\star(\Phi_t))$	$-\log\left(\det\left(\mathbf{H}_\star(\Phi_t)\right)\right)$
Spectral Variance $\rho_\lambda(\mathbf{H}_\star(\Phi_t))$	$\lambda_{\min}\left(\mathbf{H}_\star(\Phi_t)^{-1}\right)$

Since each feature's contribution is independent, the updates to the corresponding information matrices for a set Φ_t of selected features can be done by adding the individual information matrices according to the following:

$$\mathbf{H}_\star\left(\Phi_t\right) = \tilde{\mathbf{H}}_\star + \sum_{f \in \Phi_t} \mathbf{H}_\star^f. \tag{17}$$

After accounting for all visible features in the set Φ_t, the information matrix can be inverted to obtain the covariance matrix according to

$$\boldsymbol{\Sigma}_\star(\Phi_t) = \mathbf{H}_\star\left(\Phi_t\right)^{-1}. \tag{18}$$

To select a feature, it must be triangulated and the information matrix $\left(\mathbf{E}_\star^f\right)^T \mathbf{E}_\star^f$ corresponding to its position vector y_f must be invertible.

To identify the most informative features, we define suitable metrics that quantify information content or estimated accuracy. To evaluate the information content of features, several performance measures can be employed. We discuss some of these measures next.

A. Performance Measures

We quantify the informativeness of a feature subset, $\Phi_t \subset \Theta_t$, by employing established performance measures [12, 1, 3]. These performance measures are monotonic map of the covariance matrix, which is defined in (18). For accurate robot team localization, lower values of these performance measures indicate a better estimate of the robots' positions. Specific examples of these performance measures are listed in Table I.

V. Efficient Greedy Algorithms

In this section, we present the details of the feature selection algorithms designed to offer an efficient and approximate solution to maximization problem (3). To this end, we first introduce a few key definitions that will be used throughout the discussion.

Definition 1. *[13] A set function $g : 2^V \to \mathbb{R}$, where V is a ground set, is submodular if, for any two sets $A \subseteq B \subseteq V$ and any element $e \in V \setminus B$, the following inequality holds:*

$$\rho(A \cup \{e\}) - \rho(A) \geq \rho(B \cup \{e\}) - \rho(B), \tag{19}$$

for all $A, B \subseteq V$.

Submodular set functions exhibit the property of diminishing returns, meaning that the marginal gain from adding an element to a set decreases as the set grows.

Definition 2. *[13] A set function $\rho : 2^V \to \mathbb{R}$, where V is a ground set, is monotone if :*

$$A \subseteq B \implies \rho(A) \leq \rho(B), \tag{20}$$

for all $A, B \subseteq V$.

Remark 1. *A set function ρ is said to be normalized if $\rho(\emptyset) = 0$.. We employ normalized function $\rho(\cdot) = \log\det(\cdot)$, which is non negative, and satisfies the properties of submodularity and monotonicity defined in Definitions 1 and 2.*

The choice of $\rho(\cdot)$ facilitates the application of greedy algorithms, which are known to perform well with non-negative, monotone submodular functions and offer strong theoretical guarantees. For simplicity of notation, we represent marginal gain of adding a singleton $\{e\}$ to set A as

$$\rho\left(\{e\}|A\right) = \rho\left(A \cup \{e\}\right) - \rho\left(A\right) \tag{21}$$

A. Algorithms Description

Having established the fundamental properties of normalized monotone submodular set functions, we are now positioned to discuss the specific algorithms utilized to efficiently solve the feature selection problem (3) in a decentralized multi-agent setting. The following sections provide a detailed description of the algorithms, including their implementation and theoretical guarantees.

1) Algorithm 1: The Stochastic-Greedy Algorithm [11] offers an efficient solution to the feature selection problem (3) accompanied by theoretical guarantees. The algorithm begins with an empty feature set and the initial information matrix $\bar{\mathbf{H}}_\star$ (Line 3). It iterates up to q times (Line 4), where each iteration involves randomly sampling a subset $\mathcal{S}$ of size s from the remaining features (Line 5). The parameter ε is crucial as it determines the size of the sampled subset and affects the trade-off between the algorithm's computational efficiency and its approximation quality. The algorithm then selects the feature f that maximizes the marginal gain(Line 6) and adds it to the set while updating the information matrix accordingly (Line 7). This process continues until the feature set reaches the desired size, resulting in the final set Φ_t and final information matrix (Line 8).

2) Algorithm 2: The Approximate Greedy Algorithm provides a memory-efficient approach to the feature selection problem defined in (3). Unlike conventional methods that require storing full information matrices, this algorithm only maintains their traces, significantly reducing memory usage. It begins with an empty selected feature set and an initial empty score vector (Line 3). The algorithm iterates up to n times (Line 4); in each iteration, a feature $f \in \Theta_t$ is randomly sampled from the remaining candidates (Line 5), and its score—quantified by the number of frames in which it is observed—is appended to the score vector. The evaluated feature is then removed from the candidate set (Line 7). Subsequently, the score vector $\mathbf{z}$ is sorted in descending order (Line 8). Finally, the selected feature set Φ_t is formed by choosing the features corresponding to the top-q entries of the sorted score vector (Line 9).

Algorithm 1: Stochastic-Greedy Algorithm [11]

1: **Input**: $\bar{\mathbf{H}}_\star$, Θ_t, q, ε.
2: **Output**: Set Φ_t with $|\Phi_t| = q$.
3: Initialize: $\Phi_t \leftarrow \emptyset$, $\mathbf{H}_\star = \bar{\mathbf{H}}_\star$.
4: **for** ($k = 1$ to $q; k \leq q; k \leftarrow k + 1$) **do**
5: $\quad \mathcal{S} \leftarrow$ a random subset obtained by sampling $s = \frac{|\Theta_t|}{q} \log\left(\frac{1}{\varepsilon}\right)$ random elements from $\Theta_t \setminus \Phi_t$
6: $\quad f \leftarrow \underset{f \in \mathcal{S}}{\operatorname{argmax}}\ \rho\left(\{f\} \mid \Phi_t\right)$
7: $\quad \Phi_t \leftarrow \Phi_t \cup \{f\}$ and $\mathbf{H}_\star \leftarrow \mathbf{H}_\star + \mathbf{H}_\star^f$
8: **return** Φ_t.

B. Performance Analysis

In the next section, we thoroughly evaluate the performance of the algorithm 1 and provide rigorous analysis of algorithm 2.

1) Algorithm 1: To demonstrate the theoretical efficacy of the Stochastic-Greedy Algorithm 1, we assess its approximation quality relative to the optimal solution Φ_t^* in Theorem 1.

Theorem 1. *[11] For a non-negative monotone submodular function ρ and $s = \frac{|\Theta_t|}{q} \log\left(\frac{1}{\varepsilon}\right)$, the Stochastic-Greedy Algorithm 1 achieves the following approximation guarantee:*

$$\mathbb{E}\left[\rho_e\left(\Phi_t\right)\right] \geq \left(1 - \frac{1}{e} - \varepsilon\right) \rho_e\left(\Phi_t^*\right), \tag{22}$$

with only $O\left(|\Theta_t| \log \frac{1}{\varepsilon}\right)$ function evaluations.

The theorem 1 guarantees near-optimal performance in expectation for an appropriately chosen sampling parameter ε, with computational complexity independent of cardinality constraint q^i. It underscores the algorithm's efficiency in terms of function evaluations, highlighting its practicality for large-scale feature selection.

2) Algorithm 2: The greedy algorithm sequentially selects the element that maximizes the *local expected information gain*, quantified as

$$\log\det(A + B) - \log\det(A).$$

Although the greedy algorithm is effective and enjoys theoretical performance guarantees, it is computationally and memory intensive. Specifically, it requires storing the information matrices of all candidate elements and has a time complexity of $O(nq)$.

To mitigate these limitations and accelerate the selection process, we employ a tractable approximation of the log-determinant function, given by

$$\log\det(I + A) \leq \operatorname{tr}(A), \tag{23}$$

which serves as a surrogate for $\log\det(\cdot)$. In conjunction with the inequality

$$\operatorname{tr}(A^{-1}) \leq \frac{n}{\operatorname{tr}(A)}, \tag{24}$$

we use these relations to design a computationally efficient heuristic. Our objective is to minimize the uncertainty measure $\operatorname{tr}\left(\mathbf{H}_\star(\Phi_t)^{-1}\right)$. To achieve this efficiently, we instead maximize its surrogate $\operatorname{tr}(\mathbf{H}_\star(\Phi_t))$, which significantly reduces computation time and memory usage.

Algorithm 2: Surrogate Greedy Algorithm

Input: Θ_t, q, score vector $\boldsymbol{z} \in \mathbb{R}^n$
Output: Set Φ_t with $|\Phi_t| = q$
Initialize: $\Phi_t \leftarrow \emptyset$.
for ($k = 1$ to $q; k \leq n; k \leftarrow k + 1$) **do**
    Select a feature $f \in \Theta_t$
    $\boldsymbol{z}(k) = n_f$
    $\Theta_t \leftarrow \Theta_t \setminus \{f\}$
$\boldsymbol{z}_{\text{sort}}$ = sort ($\boldsymbol{z}$)
$\Phi_t \leftarrow \Phi_t \cup \{f : f \in \boldsymbol{z}(1:k)\}$
return Φ_t.

As established in Theorem 2, the trace of the information matrix associated with a visual feature depends solely on the number of frames, making this surrogate an excellent candidate for efficient feature selection.

Theorem 2. *Let the vision measurement model noise satisfy the assumption in* (12)*. Then, for a given feature $f \in \Theta_t$, the trace of its information matrix $\mathbf{H}_\star^f$ is given by*

$$\operatorname{tr}\left(\mathbf{H}_\star^f\right) = 2n_f - 3, \tag{25}$$

where n_f denotes the number of frames in which feature f is observed.

This Theorem states the trace of information matrix of a feature depends only on the number of frames in which it is visible.

VI. Conclusion

In this work, we presented two fast and memory-efficient algorithms for visual feature selection aimed at improving robot localization in unknown environments. The proposed methods enable robots to evaluate the utility of visual features without incurring the substantial computational and memory overhead typical of existing approaches. Our theoretical results demonstrate that a carefully chosen subset of features can preserve localization accuracy while significantly reducing processing latency, making the techniques well suited for onboard deployment in resource-constrained robotic systems. More broadly, this work highlights the importance of principled feature selection in visual perception pipelines and opens the door to future extensions, including multi-robot collaboration, adaptive horizon selection, and integration with modern learning-based visual front-ends.

References

[1] A. Amini, H. K. Mousavi, Q. Sun, and N. Motee. "Space Time Sampling for Network Observability". In: *IEEE Transactions on Control of Network Systems* 10.3 (2023), pp. 1159–1171.

[2] A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, and A. Krause. "Streaming Submodular Maximization: Massive Data Summarization on the Fly". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 3311–3320.

[3] L. Carlone and S. Karaman. "Attention and Anticipation in Fast Visual-Inertial Navigation". In: *IEEE Transactions on Robotics* 35.1 (2019).

[4] Y. Chen, L. Zhao, Y. Zhang, S. Huang, and G. Dissanayake. "Anchor Selection for SLAM Based on Graph Topology and Submodular Optimization". In: *IEEE Transactions on Robotics* 38.1 (2022), pp. 329–350.

[5] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[6] T. Guadagnino, X. Chen, M. Sodano, J. Behley, G. Grisetti, and C. Stachniss. "Fast Sparse LiDAR Odometry Using Self-Supervised Feature Selection on Intensity Images". In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 7597–7604.

[7] J. Jiao, Y. Zhu, H. Ye, H. Huang, P. Yun, L. Jiang, L. Wang, and M. Liu. "Greedy-Based Feature Selection for Efficient LiDAR SLAM". In: *IEEE International Conference on Robotics and Automation*. IEEE. 2022, pp. 5222–5228.

[8] S. Joshi and S. Boyd. "Sensor Selection via Convex Optimization". In: *IEEE Transactions on Signal Processing* 57.2 (2009), pp. 451–462.

[9] E. Kazemi, M. Mitrovic, M. Zadimoghaddam, S. Lattanzi, and A. Karbasi. "Submodular Streaming in All Its Glory: Tight Approximation, Minimum Memory and Low Adaptive Complexity". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 3311–3320.

[10] R. Lerner, E. Rivlin, and I. Shimshoni. "Landmark Selection for Task-oriented navigation". In: *IEEE Transactions on Robotics* 23.3 (2007), pp. 494–505.

[11] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrak, and A. Krause. "Lazier Than Lazy Greedy". In: *Twenty-Ninth AAAI Conference on Artificial Inteliigence*. 2015, pp. 182–1818.

[12] H. K. Mousavi and N. Motee. "Estimation with Fast Feature Selection in Robot Visual Navigation". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3572–3579.

[13] G. Nemhauser, L. Wolsley, and M. Fischer. "An Analysis of Approximations for Maximizing Submodular Set Functions-I". In: *Mathematical Programming* 14.1 (1978), pp. 40–53.

[14] V. Pandey, A. Amini, G. Liu, U. Topcu, Q. Sun, K. Daniilidis, and N. Motee. "Scalable Networked Feature Selection with Randomized Algorithm for Robot Navigation". In: *arXiv preprint arXiv:2403.12279* (2024).

[15] F. Pukelsheim. *Optimal Design of Experiments*. Vol. 50. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2006.

[16] C. R. Rao. "Information and the accuracy attainable in the estimation of statistical parameters". In: *Bulletin of the Calcutta Mathematical Society* 37 (1945), pp. 81–91.

[17] M. Shamaiah, S. Banerjee, and H. Vikalo. "Greedy Sensor Selection: Leveraging Submodularity". In: *IEEE International Conference on Decision and Control*. IEEE. 2010, pp. 2572–2577.

[18] C. Stachniss and W. Burgard. "Particle Filters for Robot Navigation". In: *Foundations and Trends in Robotics* 3.4 (2014), pp. 211–282.

[19] S. Thrun. "Particle Filter in Robotics". In: *Proceedings of Uncertainty in AI*. 2002.

[20] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[21] S. A. Thrun and Y. Liu. "Multi-robot SLAM with Sparse Extended Information Filters". In: *Proceedings of the th International Symposium of Robotics Research (ISRR'03)*. 2003.

[22] R. Vafaee, K. Behzad, M. Siami, L. Carlone, and A. Jadbabaie. *Non-submodular Visual Attention for Robot Navigation*. 2025. arXiv: 2510.00942 [cs.RO]. URL: https://arxiv.org/abs/2510.00942.

[23] G. Zhang and P. A. Vela. "Good Features to Track". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2015, pp. 1373–1382.

[24] Y. Zhao and P. A. Vela. "Good Feature Matching: Towards Accurate, Robust VO/VSLAM With Low Latency". In: *IEEE Transactions on Robotics* 36.3 (2020), pp. 657–675.

APPENDIX

Proof of Theorem 1: The proof of this Theorem is adapted from [11]. Given the optimal set Φ_t^* and current selected set Φ_t^k, and the set of randomly sampled elements $\mathcal{S}$, the probability that there exists an element in $\mathcal{S}$, which lies in optimal set Φ_t^* and is not currently present in Φ_t^k is given by

$$\begin{aligned}\mathbb{P}\left[\mathcal{S}\cap\left(\Phi_t^*\setminus\Phi_t^k\right)\neq\emptyset\right] &= 1-\mathbb{P}\left[\mathcal{S}\cap\left(\Phi_t^*\setminus\Phi_t^k\right)=\emptyset\right]\\ &= 1-\left(1-\frac{|\Phi_t^*\setminus\Phi_t^k|}{|\Theta_t\setminus\Phi_t^k|}\right)^s\\ &\geq 1-\exp\left(-s\frac{|\Phi_t^*\setminus\Phi_t^k|}{|\Theta_t\setminus\Phi_t^k|}\right)\\ &\geq 1-\exp\left(-s\frac{|\Phi_t^*\setminus\Phi_t^k|}{|\Theta_t|}\right)\\ &\geq \left(1-e^{-\frac{sq}{|\Theta_t|}}\right)^{(\frac{|\Phi_t^*\setminus\Phi_t^k|}{q})}\end{aligned}$$

Substituting $s=\frac{|\Theta_t^i|}{q^i}\ \log\left(\frac{1}{\varepsilon}\right)$ leads to

$$\mathbb{P}\left[\mathcal{S}\cap\left(\Phi_t^*\setminus\Phi_t^k\right)\neq\emptyset\right]\geq(1-\varepsilon)\frac{|\Phi_t^*\setminus\Phi_t^k|}{q}$$

Using the above analysis, any element in $\mathcal{S}$, which is also in $\Phi_t^*\setminus\Phi_t^k$ has a nonzero probability. Thus, the expected marginal gain of a feature $f\in\Phi_t^*\setminus\Phi_t^k$ is given by

$$\begin{aligned}\mathbb{E}\left[\rho\left(\{f\}\mid\Phi_t\right)\right] &\geq \frac{\mathbb{P}\left[\mathcal{S}\cap\left(\Phi_t^*\setminus\Phi_t^k\right)\neq\emptyset\right]}{|\Phi_t^*\setminus\Phi_t^k|}\sum_{f\in\Phi_t^*\setminus\Phi_t^k}\rho\left(\{f\}\mid\Phi_t\right)\\ &\geq\frac{(1-\varepsilon)}{q}\sum_{f\in\Phi_t^*\setminus\Phi_t^k}\rho\left(\{f\}\mid\Phi_t\right)\end{aligned}$$

The proof can be completed as follows: In iteration k, given Φ_t^k

$$\mathbb{E}_k\left[\rho\left(\{f\}\mid\Phi_t^k\right)\right]\geq\frac{(1-\varepsilon)}{q}\sum_{f\in\Phi_t^*\setminus\Phi_t^k}\rho\left(\{f\}\mid\Phi_t^k\right),$$

where $\mathbb{E}_k[\cdot]=\mathbb{E}\left[\cdot\mid\Phi_t^k\right]$ represents expectation given Φ_t^k. Since $\log\det$ is submodular, it follows that,

$$\sum_{f\in\Phi_t^*\setminus\Phi_t^k}\rho\left(\{f\}\mid\Phi_t^k\right)\geq\rho(\Phi_t^*)-\rho(\Phi_t^k),$$

which leads to

$$\begin{aligned}\mathbb{E}_k\left[\rho(\Phi_t^{k+1})-\rho(\Phi_t^k)\right] &= \mathbb{E}_k\left[\rho\left(\{f\}\mid\Phi_t^k\right)\right]\\ &\geq\frac{1-\varepsilon}{q}\left(\rho(\Phi_t^*)-\rho(\Phi_t^k)\right).\end{aligned}$$

Finally, by taking expectation over Φ_t^k leads to

$$\mathbb{E}\left[\rho(\Phi_t^{k+1})-\rho(\Phi_t^k)\right]\geq\frac{1-\varepsilon}{q}\mathbb{E}\left[\rho(\Phi_t^*)-\rho(\Phi_t^k)\right].$$

The result then follows by induction as follows:

$$\begin{aligned}\mathbb{E}\left[\rho(\Phi_t)\right] &\geq\left(1-\left(1-\frac{1-\varepsilon}{q}\right)^q\right)\rho(\Phi_t^*)\\ &\geq\left(1-e^{-(1-\varepsilon)}\right)\rho(\Phi_t^*)\\ &\geq\left(1-\frac{1}{e}-\varepsilon\right)\rho(\Phi_t^*)\end{aligned}$$

□

Proof of Theorem 2: Consider the information matrix of feature given in (A)

$$\mathbf{H}_\star^f=\sigma_i^{-2}\left(\left(\mathbf{F}_\star^f\right)^T\mathbf{F}_\star^f-\left(\mathbf{F}_\star^f\right)^T\mathbf{E}_\star^f\left(\left(\mathbf{E}_\star^f\right)^T\mathbf{E}_\star^f\right)^{-1}\left(\mathbf{E}_\star^f\right)^T\mathbf{F}_\star^f\right).$$

For simplicity of notation, we drop the subscripts and superscripts, and consider $\sigma_i=1$.

$$\mathrm{Tr}\left(H\right)=\mathrm{Tr}\left((\boldsymbol{F})^T\boldsymbol{F}\right)-\mathrm{Tr}\left((\boldsymbol{F})^T\boldsymbol{E}\left((\boldsymbol{E})^T\boldsymbol{E}\right)^{-1}(\boldsymbol{E})^T\boldsymbol{F}\right)$$

$$\begin{aligned}\mathrm{Tr}\left((\boldsymbol{F})^T\boldsymbol{F}\right) &= \mathrm{Tr}\left(\sum_{i=1}^{n_f}(\boldsymbol{E}_i)^T\boldsymbol{E}_i\right)\\ &=\sum_{i=1}^{n_f}\mathrm{Tr}\left((\boldsymbol{E}_i)^T\boldsymbol{E}_i\right)=2n_f,\end{aligned}$$

where $\boldsymbol{E}_i=U_iR_i^T$.

$$\begin{aligned}&\mathrm{Tr}\left((\boldsymbol{F})^T\boldsymbol{E}\left((\boldsymbol{E})^T\boldsymbol{E}\right)^{-1}(\boldsymbol{E})^T\boldsymbol{F}\right)\\ &\quad=\mathrm{Tr}\left(\sum_{i=1}^{n_f}\left(((\boldsymbol{E}_i)^T\boldsymbol{E}_i\right)^2\left((\boldsymbol{E})^T\boldsymbol{E}\right)^{-1}\right)\\ &\quad=\mathrm{Tr}\left(\sum_{i=1}^{n_f}\left(((\boldsymbol{E}_i)^T\boldsymbol{E}_i\right)\left((\boldsymbol{E})^T\boldsymbol{E}\right)^{-1}\right)\\ &\quad=3.\end{aligned}$$

The result follows by combining the two pieces together. □