A Theory of Active Object Localization

Alexander Andreopoulos, John K. Tsotsos
Dept. of Computer Science and Engineering & the Centre for Vision Research
York University, Toronto, Ontario, Canada
{alekos, tsotsos}@cse.yorku.ca

Abstract

We present some theoretical results related to the problem of actively searching for a target in a 3D environment, under the constraint of a maximum search time. We define the object localization problem as the maximization over the search region of the Lebesgue integral of the scene structure probabilities. We study variants of the problem as they relate to actively selecting a finite set of optimal viewpoints of the scene for detecting and localizing an object. We do a complexity-level analysis and show that the problem variants are NP-Complete or NP-Hard. We study the trade-offs of localizing vs. detecting a target object, using single-view and multiple-view recognition, under imperfect deadreckoning and an imperfect recognition algorithm. These results motivate a set of properties that efficient and reliable active object localization algorithms should satisfy.

1. Introduction

In one of the earliest known treatises on vision [1], Aristotle describes vision as a passive process that is mediated by what he refers to as the “transparent” ($\delta \alpha \omega \phi \rho \omega \varepsilon \xi$), an invisible property that allows the sense organ to become like the actual form of the visible object. Much has been learned since then and today, a popular definition is that vision is the process of discovering from images what is present in the world and where it is [11]. Within this context, four levels of tasks in the vision problem are discernible [16]:

- **Detection**: is a particular item present in the stimulus?
- **Localization**: detection plus accurate location of item.
- **Recognition**: localization of the items present in the stimulus plus their accurate description through their association with linguistic labels.
- **Understanding**: recognition plus role of stimulus in the context of the scene.

The concept of *active perception* or *active vision* was first introduced by Bajcsy [2], as “a problem of intelligent control strategies applied to the data acquisition process”. Active control of a vision based sensor offers a number of benefits [18]. It allows us to: (i) Bring into the sensor’s field of view regions that are hidden due to occlusion and self-occlusion. (ii) Foveate and compensate for spatial non-uniformity of the sensor. (iii) Increase spatial resolution through sensor zoom and observer motion that brings the region of interest in the depth of field of the camera. (iv) Disambiguate degenerate views due to finite camera resolution, lighting changes and induced motion [5]. (v) Deal with incomplete information and complete a task.

An active vision system’s benefits must outweigh the associated execution costs [18]. The associated costs in an active vision system include: (i) Deciding the actions to perform and their execution order. (ii) The time to execute the commands and bring the actuators to their desired state. (iii) Adapt the system to the new viewpoint, find the correspondences between the old and new viewpoint and deal with the inevitable ambiguities due to sensor noise.

A number of active object detection, localization and recognition algorithms have been proposed over the years [3, 4, 6, 8, 10, 12, 13, 14, 15, 19, 20]. A smaller number of papers have dealt with the complexity and reliability of such systems [5, 9, 17, 18, 21]. Limited work exists on the complexity of search tasks and the effect that imperfect recognition and imperfect dead-reckoning has on object localization. In this paper, we argue that the problem is likely intractable, by proving that the active object localization problem is NP-Hard and by showing that the problem remains difficult at best, even under certain simplifying variants of the main problem. We study the tradeoffs of localizing vs. detecting a target object under single-view and multiple-view recognition schemes and show that there are a number of bias/variance/entropy relationships and trade-offs between the reliability of target localization and target detection, that depend on the quality of the recognition algorithm used and the magnitudes of the correspondence or dead-reckoning errors. The results motivate a set of first-principles based properties that efficient and reliable active object localization algorithms should satisfy.
2. Problem Formulation

Assumption 1. We assume that exactly one instance of the target object exists in the scene.

Definition 1. (Search Space) The search space consists of a 3D region whose coordinates are expressed with respect to an inertial coordinate frame.

Definition 2. (Target Map) The target map is a discretization of the inertial coordinate frame into non-overlapping 3D cells coinciding with the search space. Each cell is assigned the probability of containing the target centroid.

We use a set of positive integers, $C = \{1, 2, ..., |C|\}$, to index each cell in the target map. Notice that, since we assume a single target object exists in the scene, the target map cell values sum to one.

Definition 3. (Scene Sample Function) A scene sample function $\mu_v(\vec{x})$ denotes the sensor output, where $v$ represents the values assigned to the controllable sensor parameter state (e.g., coordinate frame, zoom, focus) and $\vec{x}$ is an index into the scene sample function (e.g., in the case of greyscale images $\vec{x} = (i, j)$ can denote a pixel index).

We define a probability space $\Upsilon = (X_1, \Sigma_1, p_1)$ for the sensor parameter states, where $v \in X_1$ denotes a sensor parameter state, $\Sigma_1$ is a $\sigma$-algebra of $X_1$ and $p_1$ is a probability measure on $X_1$ whose support includes all states $v$ that have a non-zero probability of occurring in the search space. Similarly, for each $v$, we define a probability space $\Upsilon(v) = (X_v, \Sigma_v, p_v)$ with $p_v(\mu_v(\vec{x})) > 0$ for each $\mu_v(\vec{x}) \in X_v$, denoting the probability of occurrence of the corresponding scene sample function given sensor parameter state $v$. The underlying probability measure, models the sensed scene uncertainty (e.g., image noise, varying illumination conditions, dead-reckoning errors, etc.) and it is largely unknown and difficult to model in practice. Since we do not know the distributions of $p_1, p_v$, we approximate them by using a finite sample of optimally selected $v, \mu_v$.

Definition 4. (Sequence Cost) Given a sequence $v_1, ..., v_n$ of sensor parameter states, the cost $T(n)$ associated with executing the sequence is given by $T(n) = T(n-1) + t_0(v_1, ..., v_n)$, where $t_0(v_1, ..., v_n) > 0$ denotes the cost of moving from $v_{n-1}$ to $v_n$, given all previous states, and $T(1)$ is the cost of reaching state $v_1$ from the initial sensor state.

We define the 3D object localization and constrained active object localization (CAOL) problems as follows:

Definition 5. (3D Object Localization) Find the cell $i_0 = \arg \max_x \int p(c_i | \mu_v(\vec{x})) d\mu_v(\vec{x}) d\gamma(v)$, where we are taking the Lebesgue integrals over $\Upsilon$ and $\Upsilon(v)$ and $c_i$ denotes the event that the target object’s centroid is in cell $i$. $p(c_i | \mu_v(\vec{x}))$ is a recognition algorithm depending on $v$, $\mu_v$. If $p(c_i | \mu_v(\vec{x}))$ is a “good” algorithm with respect to $\Upsilon$, $\Upsilon(v)$, then $i_0 = i_i$, where $i_i$ is defined as the cell containing the target’s centroid.

Definition 6. (Constrained Active Object Localization) Find the cell $i_t \in C$ maximizing $p(c_i | \mu_v(\vec{x}), ..., \mu_v(\vec{x}))$ across all $n > 0$, all sequences $v_1, ..., v_n$ of sensor states and all corresponding $\mu_v(\vec{x}), ..., \mu_v(\vec{x})$, under the constraint $T(n) \leq T'$, where $T'$ is a search cost bound.

Solutions to the CAOL problem must compensate for (i) our limited knowledge on $\Upsilon(v)$ and the optimal $\Upsilon$ and (ii) the need to minimize sensor movements, by finding a finite sample $\mu_v(\vec{x}), ..., \mu_v(\vec{x})$ that best samples the unknown probability spaces without exceeding the maximum allotted search cost. Even if we know the distributions of the probability spaces $\Upsilon$, $\Upsilon(v)$, eliminating point (i) and potentially even making $p(c_i | \mu_v(\vec{x}), ..., \mu_v(\vec{x}))$ a function of $v_1, ..., v_n$, the problem remains intractable. As we show later, the CAOL problem belongs to the class of NP-Hard problems [7], implying that there is no known polynomial time algorithm that solves the problem. One can attempt to make it tractable by using variants of the problem:

Definition 7. (Constrained Active Object Localization: Variant 1) Find a sequence $v_1, ..., v_n$ of sensor states and the cells $i_t \in C$ satisfying $p(c_i | \mu_v(\vec{x}), ..., \mu_v(\vec{x})) \geq \theta$ and $T(n) \leq T'$ for some $\mu_v(\vec{x}), ..., \mu_v(\vec{x})$, where $T'$ is a search cost bound and each movement cost $t_0(v_1, ..., v_1)$ is bounded from below by a positive non-zero constant $C'$.

Definition 8. (Constrained Active Object Localization: Variant 2) Find the cell $i_t \in C$ maximizing $p(c_i | \mu_v(\vec{x}), ..., \mu_v(\vec{x}))$ across all $n > 0$, all sequences $v_1, ..., v_n$ of sensor states and all corresponding $\mu_v(\vec{x}), ..., \mu_v(\vec{x})$, under the constraint $T(n) \leq T'$, where $T'$ is a search cost bound and each movement cost $t_0(v_1, ..., v_1)$ is bounded from below by a positive non-zero constant $C'$.

Theorem 1. (Simplified Bayesian Updating) Assume $p(\mu_v | c_i, \mu_{v_{n-1}}, ..., \mu_v) = p(\mu_v | c_i)$. Then

$$ p(c_i | \mu_{v_{n-1}}, ..., \mu_v) = \frac{p(c_i | \mu_{v_{n-1}}, ..., \mu_v) p(\mu_v | c_i)}{\sum_j p(c_j | \mu_{v_{n-1}}, ..., \mu_v) p(\mu_v | c_j)}. $$

Proof. $p(c_i | \mu_{v_{n-1}}, ..., \mu_v) p(\mu_v | c_i) \Leftrightarrow p(c_i | \mu_{v_{n-1}, ..., \mu_v}) = p(\mu_v | c_i)$. Notice also that $\sum_j p(c_j | \mu_{v_{n-1}, ..., \mu_v}) = \sum_j p(\mu_v | c_j) p(c_j | \mu_{v_{n-1}, ..., \mu_v})$. $\square$

When we are not using the simplifying assumption stated in Theorem 1, we say we are using normal Bayesian updating. Theorem 1 assumes that the scene sample functions are conditionally independent given the cell $i$ where the target is centred. By Assumption 1, exactly one instance of the target exists in the scene, which implies that event $c_i$ is sufficient to determine which regions of $\mu_v$ (if any) correspond to the projection of the target object on the image plane and which
regions correspond to the background. We are implicitly assuming that $p(\mu_{v_n} | c_i)$ denotes a generative modeling of the recognition algorithm’s resultant segmentation into the foreground (target position) and the background, based on a single view. Similarly $p(c_i | \mu_{v_0}, \ldots, \mu_{v_n})$ denotes the corresponding probability of event $c_i$, based on the bayesian fusion of multiple-views $\mu_{v_0}, \ldots, \mu_{v_n}$. Notice that for a uniform prior $p(c_i), \arg \max_i p(\mu_{v_n} | c_i) = \arg \max_i p(c_i | \mu_{v_n})$. The greater the uncertainty implicit in spaces $T(v)$, the weaker the assumption of conditional independence becomes, due to increased sources of error. Nevertheless, it is convenient to use Theorem 1 to model various localization tradeoffs.

In the next section, we prove that if we know the distributions of $\Upsilon, \Upsilon(v)$ and under normal Bayesian updating, Variant 1 of the CAOL problem (Def.7) and the corresponding detection problem are NP-Hard and NP-Complete respectively. It is easy to see that Def.7 is reducible to the similarly discretized version of Def.6 and thus, the CAOL problem is NP-Hard. For constant $T', C'$, Variant 2 of the problem, may have a high-order polynomial solution: Since there are at most $\binom{m}{T'}$ sensor settings to execute within time $T'$, an enumeration and evaluation of all candidate solutions, runs in $\Omega(m^{T'/3})$, where $m$ is the total number of possible states. But for non-constant $T', C'$, the running time increases exponentially as the length of $T'$ increases. We see that for the set of problem instances where the cost function $t_{kn}$ is defined such that we can efficiently set $C'$ to the minimum achievable sensor state pair cost, by a trivial reduction from Def.7 we can show that Def.8 is NP-Hard. We could also approach the localization problem by thresholding the generative probability $p(\mu_{v_n}(x) | c_i)$ rather than the discriminative probability $p(c_i | \mu_{v_n}(x), \ldots, \mu_{v_n}(x))$. Ye [21] uses a binary classifier with a presumed zero false positive rate, to show that a similar problem is NP-Complete.

3. The Constrained Active Object Localization Problem: Variant 1, is NP-Hard

To analyze the complexity of the constrained active object localization problem when we know the distributions of $\Upsilon, \Upsilon(v)$, we first reformulate the problem into the corresponding detection problem, taking into account the finite precision of floating point arithmetic, and the finite set $V$ that is necessary to represent the space of scene sample functions $(X_u)$ achievable across the sensor parameter states $(X_1)$. Let $Q^+ \triangleq \{ \frac{p}{q} : p, q \in \mathbb{Z}^+ \}$ denote the set of positive rational numbers. We model each probability by a non-negative rational in $Q^+_1 \triangleq \{ x \in Q^+ \cup \{0\} : x \leq 1 \}$.

**Definition 9. (Valid Sequence)** Let $v'_i = (v_{\pi_i(1)}, \ldots, v_{\pi_i(l(i))})$ denote an ordered set of length $l(i)$, where $\pi_i : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ is a one-to-one mapping. A sequence $v'_i, v''_i$ of ordered sets is valid if $l(i_1) = 1, l(i_{k+1}) = l(i_k) + 1$ for each $1 \leq k < n - 1$ and $\pi_{i_k}(j) = \pi_{i_{k+1}}(j), \forall j 1 \leq j \leq l(i_k)$.

We define an ordered set of length zero as $v'_0 \triangleq ()$. For any ordered set $v'_i = (v_{\pi_i(1)}, \ldots, v_{\pi_i(l(i))})$ let $v'_i(v_c) \triangleq (v_{\pi_i(1)}, \ldots, v_{\pi_i(l(i))}, v_c)$. Also $v'_0(v_c) \triangleq (v_c)$ and $i_0 \triangleq 0$.

**Definition 10. (Π$_\Pi$: Constrained Active Object Localization Problem: Variant 1)**

**INSTANCE:** A finite set $V = \{v_1, \ldots, v_N\}$. A cost constraint $B'$ is $\in \mathbb{Q}^+$, and a cost function $C(v'_i) \in \mathbb{Q}^+$ where $v'_i = (v_{\pi_i(1)}, \ldots, v_{\pi_i(l(i))}), i \in \mathbb{Z}^+, v_{\pi_i(1)}, \ldots, v_{\pi_i(l(i))} \in V$. $S' = \in \mathbb{Z}^+$ denoting the number of cells in the target map. A function $f_1(v'_i, j) \in \mathbb{Q}^+_1$ such that for any ordered set $v'_i$ and any $1 \leq j \leq S'$, $\sum_{v_c \in V} f_1(v'_i(v_c), j) = 1$.

A function $f_2(v'_i, j) \in \mathbb{Q}^+_1$ defined for $1 \leq j \leq S'$, such that $\sum_{j=1}^{S'} f_2(v'_i, j) = 1$ for all ordered sets $v'_i$ and $f_2(v'_{i_n}, j) \triangleq f_2(v'_i, j) \prod_{k=1}^{n} f_1(v'_i, j_k)$.

A recognition threshold $\theta \in \mathbb{Q}^+_1$. A query cell $1 \leq j \leq S'$.

**QUESTION:** Is there a valid sequence $v'_i, \ldots, v'_{i_n}$ so that $\sum_{k=1}^{n} C(v'_{i_k}) \leq B'$ and $f_2(v'_{i_n}, j) \geq \theta$?

As $\theta, \bar{\theta}$ decrease, the expected running times of $\Pi_{\Pi}, \mathbb{Π}$ do not increase (e.g., for $\theta, \bar{\theta} = 0$, solutions in $O(|V|)$ are trivial to find). Notice that for $\theta > \frac{1}{2}$, there is at most one cell that $\mathbb{Π}$ can output. We quote the Knapsack problem (an NP-Complete problem) as given by Garey and Johnson [7]:

**Definition 12. (Π$^\mathbb{Π}$: Knapsack Problem)**

**INSTANCE:** A finite set $U$, a "size" $s(u) \in \mathbb{Z}^+$ and a "value" $w(u) \in \mathbb{Z}^+$ for each $u \in U$, a size constraint $B \subset \mathbb{Z}^+$, and a value goal $K \in \mathbb{Z}^+$.

**QUESTION:** Is there a subset $U' \subset U$ such that $\sum_{u \in U'} s(u) \leq B$ and $\sum_{u \in U'} w(u) \geq K$?

$\Pi_{\Pi}$ is in NP, since any candidate solution is verifiable in polynomial time. We assume $\sum_{u \in U} w(u) \leq K$ since otherwise $\#U' \subset U$ that satisfies $\Pi_{\Pi}$. We define a mapping $f$ from $\Pi_{\Pi}$ to $\Pi_{\Pi}$, for which $\Pi_{\Pi}$ is true if and only if $\Pi_{\Pi}$ is true:

1. $V \leftarrow U$
2. $B' \leftarrow B$
3. $C(v'_i) = s(v_{\pi_i(l(i))})$
4. $S' \leftarrow 2$
5. $\theta \leftarrow \sum_{u \in U} w(u)$
6. We need to define $f_1(v'_i, j)$ and $f_2(v'_i, j)$ for all ordered sets $v'_i$ that are composed of elements in $V$ and all $j$, $1 \leq j \leq S'$, such that $f_1, f_2$ satisfy their preconditions stated in $\Pi_{\Pi}$.
For each distinct set \( U' \subseteq V \) and each distinct ordering \( o \) of the elements in \( U' \), we assume \( d(U', o) \in \mathbb{Z}^+ \) is unique and define the identifier of the corresponding ordered set \( v'_{d(U', o)} = (v_{\pi_{d(U', o)}(1)}, ..., v_{\pi_{d(U', o)}(|d(U', o)|)}) \) where \( l(d(U', o)) = |U'| \). Furthermore, \( d(U', o, k) \), for \( 1 \leq k \leq l(d(U', o)) \), denotes the ordered set composed of the first \( k \) elements of \( v'_{d(U', o)} \) — i.e., \( v'_{d(U', o, k)} = (v_{\pi_{d(U', o)}(1)}, ..., v_{\pi_{d(U', o)}(k)}) \) and \( v'_{d(U', o, k)} = v'_j \) iff \( d(U', o, k) = j\). For any ordering \( o \) and set \( U' = \{v_{\pi_{d(U', o)}(1)}, ..., v_{\pi_{d(U', o)}(|d(U', o)|)}\} \subseteq V \), we need to define \( f_1(v'_{\tilde{d}(U', o, k)}, j) \) and \( f_2(v'_{d(U', o, k)}, j) \) for all \( 1 \leq k \leq l(d(U', o)) \). We also need to make sure \( f_1(v'_{d(U', o, k)}, j) \) and \( f_2(v'_{d(U', o, k)}, j) \) satisfy the requirements set in the definition of \( \Pi_j \) and only depend on \( j \) and the first \( k \) parameters of \( v'_{d(U', o)} \). For each instance of \( \Pi \) we define \( f_2 \) in \( \Pi_j \) by

\[
f_2(v'_j, j) = \begin{cases} \\
\sum_{i=1}^{l(o)} w(v_{\pi_i}(k)) \quad \text{if } j = j_t \\
\frac{1}{S' - 1} \left( 1 - \sum_{i=1}^{l(o)} w(v_{\pi_i}(u)) \right) \quad \text{otherwise}
\end{cases}
\]

Since \( \sum_{i=1}^{S'} f_2(v'_j, j) = 1 \), \( f_2(v'_j, j) \) satisfies the requirements in \( \Pi_j \). Notice from Def.10 that if \( f_2(v'_j, j) = 1 \), then \( f_1(v'_{i_k}, j') \neq 0 \). Also, if \( 0 < f_2(v'_j, j) < 1 \), then \( 0 < f_2(v'_{i_{k-1}}, j) < 1 \). From the definition of \( \Pi_j \), for each subset \( U' \), each ordering \( o \) and each \( 1 \leq k \leq l(d(U', o)) \), we want to define \( f_1 \) so that

\[
f_2(v'_j, j) = \frac{f_2(v'_{i_{k-1}}, j) f_1(v'_{i_k}, j)}{\sum_{j'} f_2(v'_{i_{k-1}}, j') f_1(v'_{i_k}, j')} \tag{2}
\]

where \( i_k = d(U', o, k) \), \( 1 \leq k \leq l(d(U', o)) \), is used to denote a valid sequence of ordered sets. From Lemma 1 below, we know that for each sensor setting \( v'_{i_k} \) and \( v'_{j} \), there exists an assignment function \( f_1(v'_{i_k}, v'_{j}) \) that satisfies Eq.(2) and depends only on the parameters \( v'_{i_k}, v'_{j} \) — i.e., given parameters \( v'_{i_k} \) and \( v'_{j} \), \( f_1 \) is independent of set \( U' \). Also Eq.(2) is independent of scaling factors applied on \( f_1 \), implying that we can assume that \( \sum_{v'_{i_k} \in V} f_1(v'_{i_k}, v'_{j}) = 1 \) as wanted. We see that mapping \( f \) runs in polynomial time.

We now show that there exists a valid sequence \( v'_{i_1}, ..., v'_{i_n} \) that satisfies \( \Pi_j \) iff \( \exists \Pi' \subseteq \Pi \) that satisfies \( \Pi' \): If \( \Pi_j \) holds, \( f_2(v'_{i_n}, j) \geq \theta \Rightarrow \sum_{u \in U'} w(u) \geq K \) where \( U' = \{v_{\mu_{1}(1)}, ..., v_{\mu_{n}(l(i_n))}\} \subseteq U \). Conversely, assume that for a subset \( U' \subseteq U \) problem \( \Pi' \) holds. Choose an arbitrary ordering \( o \) and let \( i_k = d(U', o, k) \). We see that \( f_2(v'_{i_k}, j) \geq \theta \). The converse direction of the proof holds regardless of the ordering assigned to \( U' \). Regardless of the ordering \( o \) assigned to \( U' \), \( \sum_{k=1}^{l(o)} C(v'_{i_k}) \leq B' \) iff \( \sum_{u \in U'} w(u) \leq B \), which proves that there is a subset \( U' \) satisfying \( \Pi' \) if an ordered set satisfies \( \Pi_j \). This proves that \( \Pi_j \), under normal Bayesian updating, is NP-Complete.

To prove that \( \Pi \) is NP-Hard, we define a mapping from \( \Pi_j \) to \( \Pi \) as follows: \( \tilde{V} \leftarrow V, B' \leftarrow B', C(v') = C(v'_j), S' \leftarrow 2, \theta \leftarrow \frac{\theta}{2}, f_2(v'_{i_k}, j) = \frac{1}{2} I_{A(k)} + \frac{1}{2} I_{A(k)}, f_2(v'_{i_k}, j) = \frac{1}{2} I_{A(k)} + \frac{1}{2} I_{A(k)} \), where \( I_X = \{0, 1\} \) is an indicator function that takes a value of 1 iff boolean variable \( X \) is true and \( A(k) \) are true iff \( f_2(v'_{i_k}, j) \geq \theta \) or \( f_2(v'_{i_k}, j) < \theta \) respectively. By Lemma 1, this also implicitly defines \( f_1 \). We see that \( \Pi_j \) holds iff \( \Pi \) finds a valid sequence that is satisfied by cell \( j = 1 \). This shows that \( \Pi \) is NP-Hard.

In the reduction from \( \Pi' \) to \( \Pi_j \), each call to \( f_2 \) is in \( O(|V|) \) and takes \( O(|V| \cdot S') \) space to encode. We are making the implicit assumption that \( f_1, f_2 \) in \( \Pi_j \) and \( f_1, f_2 \) in \( \Pi \) have running times and encoding sizes that are polynomial functions of \( |V|, S' \) and \( |V|, S' \) respectively, implying that the scene structure must exhibit a minimum degree of “non-randomness”. From the above proofs and Lemma 1, we notice that \( f_1(v'_{i_k}, j) \) and \( f_1(v'_{i_k}, j) \) correspond to \( p(\mu_1, \mu_2, ..., \mu_{l(o)}) \). Only if \( f_1(v'_{i_k}, j) \) depended exclusively on \( j \) and \( v_{\mu_{i_k}(l(i_k))} \) would this constitute a proof that Def.11 is NP-Hard under simplified Bayesian updating, \( f_2(v'_j, j) \) and \( f_2(v'_j, j) \) denote the prior distributions of the target maps and are typically set to a uniform distribution.

**Lemma 1.** Let \( \beta, \alpha_1, ..., \alpha_m \in \mathbb{Q}^+ \) such that \( \sum_{i=1}^{m} \alpha_i = 1 \), if \( \beta = 1 \), then \( \alpha_1 \neq 0 \) and if \( 0 < \beta < 1 \), then \( 0 < \alpha_1 < 1 \). If \( m > 1 \), \( \exists x_1, ..., x_m \in \mathbb{Q}^+ \) such that \( \frac{\alpha_1}{\sum_{i=1}^{m} \alpha_i} = \beta \).

**Proof.** If \( \beta = 1 \), let \( x_1 = 1 \) and let \( x_i = 0 \) for \( i \neq 1 \). If \( \beta = 0 \), let \( x_2 = 1 \) and let \( x_i = 0 \) for \( i \neq 2 \). Otherwise, if \( 0 < \beta < 1 \), assume \( x_1 > 0 \) and notice that \( \sum_{i=1}^{m} \alpha_i = \beta \Leftrightarrow \alpha_1 - \beta \alpha_1 = \sum_{i=2}^{m} (\beta \alpha_i) y_i \), a linear equation of \( y_i = \frac{x_i}{x_1} \). Since \( 0 < \beta < 1 \), \( 0 < \alpha_1 < 1 \) and consequently \( \sum_{i=2}^{m} \alpha_i > 0 \), which implies \( \alpha_1 - \beta \alpha_1 > 0 \) and \( \sum_{i=1}^{m} \alpha_i > 0 \). Therefore, there exist \( y_2, ..., y_n \) which satisfy the linear equation. We leave it as an exercise for the reader to verify that for any \( y_2, ..., y_n \in \{0, 1\} \cap \{0\} \), \( \exists x_1, x_2, ..., x_n \in \mathbb{Q}^+ \) \((x_1 \neq 0)\) which satisfy \( y_i = \frac{x_i}{x_1} \).

4. Localization vs. Detection

We formalize some of the tradeoffs of single-view and multiple-view recognition schemes for localizing and detecting a target object under simplified Bayesian updating and under a number of different sources of errors. In Sec. 4.1 we define and discuss the problems and in Sec. 4.2-4.3 we prove the respective theorems.

4.1. Definitions and Discussion

**Definition 13. (Correspondence Error)** Any error in the calculation of the correspondence(s) between the index value \( \bar{x} \) of a scene sample function \( \mu_v(\bar{x}) \) and the target map cell indices whose structure projects on \( \bar{x} \).
Definition 14. (Dead-Reckoning Errors) We are dealing with dead-reckoning errors when there exists a reflection-free rigid transformation $RT(\cdot)$ of the sensor’s estimated coordinate frame with respect to the inertial coordinate frame of the search space, that corrects all correspondence errors without introducing new correspondence errors.

Definition 15. (Visibility) Cell $i$ is visible for state $v_n$, if it falls in the sensor’s field of view and satisfies a set of necessary conditions for localizing a target centered in $i$, that may only depend on (i) the coordinates of the points in cell $i$ and the depth map of $v_n$, with respect to the sensor coordinate frame, and, (ii) the parameters of $v_n$, which do not affect the sensor coordinate frame.

Definition 16. (Good Single-View Recognition) We have good single-view recognition at step $n$ if $p(\mu_{v_n}|c_i)$ is not affected by changes to the inertial coordinate frame. Also, under dead-reckoning errors, $p(\mu_{v_n}|c_i) \geq p(\mu_{v_n}|\neg c_i)$ for all target map distributions at step $n - 1$ if $i \in \hat{V}(v_n)$ and $RT(i) = i$, or $i \notin \hat{V}(v_n)$ and $RT(i) \notin \hat{V}(v_n)$.

$RT(i_t)$ denotes the cell containing the transformation of the target’s centroid under $RT(\cdot)$ (Def. 14). $p(\mu_{v_n}|\neg c_i)$ is defined in Sec. 4.2. $V(v_n)$ is the ground truth of visible cells for $\mu_{v_n}$, $v_n$ and no correspondence errors, while $\hat{V}(v_n)$ denotes the calculated visible cells based on our estimate of the sensor coordinate frame and under no guaranty of perfect correspondences. Under perfect correspondences $\hat{V}(v_n) = V(v_n)$, but the converse does not hold. For good single-view recognition, as the correspondence errors increase, it is more likely that $p(\mu_{v_n}|c_i) < p(\mu_{v_n}|\neg c_i)$. Def.16 implies that if $i_1, i_2 \notin \hat{V}(v_n)$, $i_3 \in \hat{V}(v_n)$ and $RT(i_t) \notin \hat{V}(v_n)$, $p(\mu_{v_n}|c_{i_1}) = p(\mu_{v_n}|c_{i_2})$ and $p(\mu_{v_n}|c_{i_3}) < p(\mu_{v_n}|c_{i_t})$. Also, if $RT(i_t) \in \hat{V}(v_n)$, $p(\mu_{v_n}|c_{RT(i_t)}) > p(\mu_{v_n}|c_j) \forall j \neq RT(i_t)$ (see Sec. 4.2).

Theorem 2. (Detection Tradeoff) Assume $i_t \in \hat{V}(v_n)$, $i \in \hat{V}(v_n)$. Assume a uniform target map prior and good single-view recognition. Let $X_{i_t}^{(n)}$, $Y_{i_t}^{(n)}$ denote Bernoulli random variables with probability of success $p(c_i|\mu_{v_n}, ..., \mu_{v_1})$, $p(c_i|\mu_{v_n})$ respectively.

Detection at step $n$ is based on max $\max_{j \in \hat{V}(v_n)} E(X_{i_t}^{(n)})$ or $\max_{j \in \hat{V}(v_n)} E(Y_{i_t}^{(n)})$ being above a given threshold.

(i) Given $v_n$, $\mu_{v_n}$, single-view detection at step $n$ is independent of dead-reckoning errors.

(ii) If $p(\mu_{v_n}|c_i) \leq p(\mu_{v_n}|\neg c_i)$, $X_{i_t}^{(n)} \leq E(X_{i_t}^{(n)})$.

(iii) If $p(\mu_{v_n}|c_i) \geq p(\mu_{v_n}|\neg c_i)$, $X_{i_t}^{(n)} \geq E(X_{i_t}^{(n)})$.

(iv) If $\hat{i}_t = \hat{j}_t$, $\hat{i}_t = \arg \max_{j \in C} E(Y_{j_t}^{(n)})$ and $\hat{j}_t = \arg \max_{j \in C} E(Y_{j_t}^{(n)})$, $E(X_{i_t}^{(n)}) \geq E(X_{j_t}^{(n)})$.

(v) If $\hat{i}_t \neq \hat{j}_t$, $\hat{i}_t = \arg \max_{j \in C} E(Y_{j_t}^{(n)})$ and $\hat{j}_t = \arg \max_{j \in C} E(Y_{j_t}^{(n)})$, then it is not necessarily the case that $E(X_{i_t}^{(n)}) \geq E(X_{j_t}^{(n)})$.

Case (iv) shows that with good correspondences, detection based on fusing multiple views becomes more reliable than single-view detection (since $i_t, j_t \in \hat{V}(v_n)$). Case (v) shows that under dead-reckoning errors, there is an increased likelihood that fusing multiple-views will lead to more false negative detections (since $\hat{i}_t, \hat{j}_t \notin \hat{V}(v_n)$), and thus, single-view detection (case (i)) might be preferable when dead-reckoning errors occur. Despite the strong assumption of Def.16, correspondence or dead-reckoning errors make the detection problem significantly harder.

Definition 17. (Dual Support) Let $x_i^{(n)} \triangleq p(\mu_{\mu_{v_n}}|\mu_{v_n}, ..., \mu_{v_1})$. A single-view recognition algorithm has dual support at step $n$ if $i \forall i$, $x_i^{(n)} \notin \{\frac{1}{2}, \frac{1}{2}\}$. Equivalently, $\frac{p(\mu_{v_n}|\neg c_i)}{p(\mu_{v_n}|c_i)} > \frac{x_i^{(n-1)}}{1-x_i^{(n-1)}}(e-1)$ or $\frac{p(\mu_{v_n}|\neg c_i)}{p(\mu_{v_n}|c_i)} > \frac{1-x_i^{(n-1)}}{x_i^{(n-1)}}$.

Definition 18. (Flipped Cells) We say that there exist flipped cells at step $n$, if there exist two cells $i_1, i_2$, such that $x_{i_1}^{(n)} < \frac{1}{2}$, $x_{i_2}^{(n)} < \frac{1}{2}$, $x_{i_1}^{(n)} = x_{i_1}^{(n-1)} - x_1 < \frac{1}{2}$, $x_{i_2}^{(n)} = x_{i_2}^{(n-1)} + x_2 > \frac{1}{2}$ for positive $x_1, x_2$.

Under Def.16 and a uniform target map prior, flipped cells can only occur due to correspondence errors.

Definition 19. (Boundary Constraints) We say that the cells in a set $S$ satisfy the boundary constraints at step $n$ if for each $i \in S$, $p(\mu_{v_n}|c_i) < p(\mu_{v_n}|\neg c_i)$ and

$$p(c_i|\mu_{v_n-1}, ..., \mu_{v_1}) < \frac{\sqrt{p(\mu_{v_n}|\neg c_i) - p(\mu_{v_n}|c_i)p(\mu_{v_n}|\neg c_i)}}{p(\mu_{v_n}|\neg c_i) - p(\mu_{v_n}|c_i)},$$

or, $p(\mu_{v_n}|c_i) > p(\mu_{v_n}|\neg c_i)$ and

$$p(c_i|\mu_{v_n-1}, ..., \mu_{v_1}) > \frac{\sqrt{p(\mu_{v_n}|\neg c_i) - p(\mu_{v_n}|c_i)p(\mu_{v_n}|\neg c_i)}}{p(\mu_{v_n}|\neg c_i) - p(\mu_{v_n}|c_i)}.$$

Theorem 3. (Localization Tradeoff) Assume $C$ satisfies the boundary constraints at step $n$. Also assume a uniform prior distribution for the target map. Define $d_i^{(n)} \triangleq x_i^{(n-1)} - x_i^{(n)}$ and $r_{i,k}^{(n)} \triangleq \frac{d_i^{(n)}}{\sum_{j \neq i} d_j^{(n)}}$.

(i) Assume there are no flipped cells at step $n$ and $x_{i_1}^{(n)} < \frac{1}{2}$. Then, there exists a cell $j_1$ for which $x_{i_1}^{(n)} > \frac{1}{2}$. Furthermore, if $x_{i_1}^{(n)} > \prod_{k \neq i_1} x_{i_k}^{(n)}$, the target map entropy at step $n$ is smaller than it is at step $n - 1$.

(ii) If $x_{i_1}^{(n)} > \frac{1}{2}$ for some cell $i_1$, there exists a cell $j_1$, which does not have to equal $i_1$, such that $x_{j_1}^{(n)} > \frac{1}{2}$.

(iii) If there are no flipped cells at step $n$ and there exists a cell $i_1$ satisfying $x_{i_1}^{(n)} > \frac{1}{2}$ and $x_{i_1}^{(n)} > \frac{1}{2}$, then, the target map entropy at step $n$ is smaller than it is at step $n - 1$.

(iv) If there exist flipped cells $i_1, i_2$ at step $n$, the condition $x_1, x_2 > x_{i_1}^{(n)} - x_{i_2}^{(n)}$ (see Def.18) and single-view recognition with dual support, guarantees that the target map entropy at step $n$ is smaller than it is at step $n - 1$. 907
Any termination condition based on probability thresholding (e.g., Def.7), requires a decreasing target map entropy. The above theorem quantifies a set of sufficient properties of the recognition algorithm, under which, multiple-view localization leads to a decreasing entropy and therefore, after a certain number of steps, a smaller target map entropy than that of a single-view. Theorem 3 lists all possible target map behaviours under the boundary constraints. If we also assume good single-view recognition and that no correspondence errors exist, Theorem 3 defines a set of sufficient properties of the single-view recognition algorithm so that multiple-view recognition leads to a decreasing target map entropy and a smaller bias and variance in the target’s localization at each step. Without the boundary constraints, we have no guaranty of a decreasing entropy. Under good single-view recognition and a uniform target map distribution, we have no guaranty of a decreasing entropy. Under good single-view recognition and a uniform target map distribution, we have no guaranty of a decreasing entropy. Let \( g(x, \alpha, \beta) = \frac{\alpha}{\alpha x + \beta (1 - x)} \) with \( 0 \leq \alpha, \beta, x \leq 1 \) such that \( \alpha x + \beta (1 - x) \neq 0 \). Then \( g(x, \alpha, \beta) \leq 1 \) if \( \beta > \alpha \) or \( x = 1 \) or \( \alpha = \beta \).

**Proof.** Notice that \( g(x, \alpha, \beta) \leq 1 \) if \( \beta > \alpha \) or \( x = 1 \) or \( \alpha = \beta \). If \( \alpha < \beta \), then \( g(x, \alpha, \beta) > 1 \) iff \( x = 1 \).

Let \( \mu_{v_1} \) denote a Bernoulli random variable with probability of success \( x_i^n \). By Lemma 3, the boundary constraint assumption of Theorem 3 is equivalent to \( \text{Var}(X_i^n) < \text{Var}(X(i) = 1) < C \). Since the variance of a Bernoulli(\( p \)) random variable is equal to \( p(1 - p) \), it is maximized at \( p = \frac{1}{2} \) and it is also symmetric around \( p = \frac{1}{2} \), which implies that when the variance of \( X_i^n \) has decreased, \( x_i^n - \frac{1}{2} > |x_i^n - \frac{1}{2}| \). Since \( \text{Var}(X_i^n) < \text{Var}(X(i) = 1) \) for all cells \( i \), there exists exactly one cell \( i_1 \) at step \( n \) with \( x_i^n > \frac{1}{2} \), since otherwise, the variance of all cells could not have decreased and maintained a sum of one across all target map cells. This proves the first half of Theorem 3(i).

One of the following conditions must hold at each step \( n \):

1. \( i_1 \): \( x_i^{(n-1)} \leq \frac{1}{2} \) and there exists exactly one cell \( i_1 \) such that \( x_i^{(n)} > \frac{1}{2} \).

2. There exist two cells \( i_1, i_2 \) such that \( x_i^{(n)} > \frac{1}{2} \), \( x_i^{(n)} < \frac{1}{2} \), \( x_i^{(n)} > \frac{1}{2} \), \( x_i^{(n)} < \frac{1}{2} \).

3. \( x_i^{(n)} > \frac{1}{2}, x_i^{(n)} < \frac{1}{2}, x_i^{(n)} > \frac{1}{2}, x_i^{(n)} < \frac{1}{2} \).

Assume condition (1) applies. We now prove the second half of Theorem 3(i). For notational simplicity we index the \( |C| - 1 \) cells that are not equal to \( i_1 \) by the set \( \{1, 2, \ldots, |C| - 1\} \). Let \( g(p) = -p \log(p) \). We want to show that \( g(x_i^{(n-1)}) + \)
Notice that because of the boundary constraint and Lemma 3, $d_i^{(n)} > 0$ for $i \neq t_1$ and $x_i^{(n-1)} = x_i^{(n-1)} + x(n)$ since the target map cells have to sum to one at step $n$. By the Mean Value Theorem, for each $i \in \{1, \ldots, |C| - 1\}$, $\exists z_i \in \left[x_i^{(n-1)} - d_i^{(n)}, x_i^{(n)} \right]$ such that $g(x_i^{(n-1)}) - g(x_i^{(n-1)} - d_i^{(n)}) = d_i^{(n)} g'(z_i)$ and $\exists z \in \left[x_i^{(n-1)}, x_i^{(n)} + x(n) \right]$ such that $g(x_i^{(n-1)}) - g(x_i^{(n-1)}) = x(n) g'(z)$. Notice that $\sum_{i=1}^{\alpha} r_i^{(n)} = 1$ and $g'(p) = -\frac{\log(p)}{\log(2)} - \frac{1}{\log(2)}$. This in turn implies that $\sum_{i=1}^{\alpha} r_i^{(n)} g'(z_i) > g'(z)$ and the entropy decreases if and only if $\prod_{i=1}^{\alpha} z_i^{(n)} r_i^{(n)} < z$.

But since $\prod_{i=1}^{\alpha} z_i^{(n)} r_i^{(n)} \leq \prod_{i=1}^{\alpha} (x_i^{(n-1)})^{r_i^{(n)}}$ and $\prod_{i=1}^{\alpha} z_i^{(n)} r_i^{(n)} \leq z$, a sufficient condition for a decrease in the entropy is $x_i^{(n-1)} > \prod_{i=1}^{\alpha} (x_i^{(n-1)})^{r_i^{(n)}}$. This proves (i).

The proof of part (ii) of the theorem follows, since if $x_i^{(n)} \leq \frac{1}{2}$ for all cells $i \in C$, then the probability of cell $i_1$ has decreased at step $n (x_i^{(n)} \leq \frac{1}{2} < x_i^{(n-1)})$ and for at least one cell $i_2$, $x_i^{(n-1)} < x_i^{(n)} \leq \frac{1}{2}$ so that all cell probabilities sum to one at step $n$. But this contradicts the monotonically decreasing variances implied by Lemma 3, proving (ii).

If condition (2) holds, by a recursive application of Lemma 4 (by setting $\gamma = \frac{1}{2}$), we see that $-\sum_{i \in C} x_i^{(n)} \log(x_i^{(n)}) < -\sum_{i \in C} x_i^{(n-1)} \log(x_i^{(n-1)})$ as desired. This proves part (iii) of the theorem.

For the proof of part (iv) of the theorem, condition (3) applies. Notice that $g(p)$ is monotonically increasing on $(0, \frac{1}{2}]$ and monotonically decreasing on $(\frac{1}{2}, 1]$. Since we have assumed $x_1, x_2 > x_1^{(n-1)} - x_2^{(n-1)}$, it suffices to show that $g(x_1^{(n-1)}) + g(x_2^{(n-1)}) > g(x_1^{(n-1)} + x_2^{(n-1)})$ (since the probabilities of all cells $i \neq i_1, i_2$ have decreased and we assume dual support). Equivalently, we want to show that $g(x_1^{(n-1)}) + g(x_2^{(n-1)}) > g(x_1^{(n-1)} - x_1) + g(x_2^{(n-1)} + x_2)$. But since $x_1^{(n-1)} > x_1^{(n-1)} - x_1, x_2^{(n-1)} < x_2^{(n-1)} + x_2$ and we have assumed dual support $(x_1^{(n-1)} > \frac{1}{2}, x_2^{(n-1)} < \frac{1}{2})$, we have proven part (iv) of the theorem.

**Lemma 3.** \textit{Var}(X_{n}^{(1)}) < \textit{Var}(X_{n}^{(n-1)}) if and only if $p(\mu_{\nu}|c_i) < p(\mu_{\nu}||c_i)$ and} 

$p(c_i|\mu_{\nu_{u-1}}, \ldots, \mu_{\nu_{v}}) > \frac{p(\mu_{\nu}||c_i)}{p(\mu_{\nu}||c_i)} - \frac{p(\mu_{\nu}|c_i)}{p(\mu_{\nu}|c_i)}$ or $p(c_i|\mu_{\nu_{u-1}}, \ldots, \mu_{\nu_{v}}) > \frac{p(\mu_{\nu}|c_i)}{p(\mu_{\nu}|c_i)} - \frac{p(\mu_{\nu}||c_i)}{p(\mu_{\nu}||c_i)}$
\[ g'(\alpha') = \frac{g(\alpha) - g(\alpha - x)}{\frac{\log(p)}{\log(2)} \cdot \log(2)} \] and \[ g'(\beta') = \frac{g(\beta + x) - g(\beta)}{x} \]

But since \[ g'(p) = -\frac{\log(p)}{\log(2)} \] is a decreasing function and \[ \forall \alpha'' \in [\alpha - x, \alpha] \forall \beta'' \in [\beta, \beta + x], g'(\alpha'') > g'(\beta''), \] we have proven the lemma.

5. Discussion

A number of optimization algorithms for navigation, mapping and next-view-planning have been suggested, based on POMDPs [12], Bayesian methods [13, 14], heuristics [5, 8, 19] and greedy algorithms [20] amongst others. The arguments in Sec. 3 suggest what kind of policies would lead to efficient and reliable solutions for the components of active object detection and localization systems that deal with sensor control for recognition. These include next-view-planners that use efficient approximation algorithms, or algorithms based on greedy and dynamic programming solutions to the Knapsack problem [7, 20], suggesting that a mixture of specialized optimizers, rather than a single kind of optimization, could lead to more efficient solutions, without a significant decrease in reliability.

Theorems 2.3 suggest that single-view localization, where the updated target map is only used to guide the where-to-look-next policy, can lead to fewer false positive/negative detections, at the expense of greater localization bias when dead-reckoning errors occur. Alternatively, if we have some prior knowledge about the expected maximum dead reckoning error—typically the main source of correspondence errors—, we can define appropriate dimensions for cell \( i_t \), such that the target’s centroid always falls inside cell \( i_t \). For example, an adaptive multiscale target map approach could be used. At each step, we could adjust the scale of the cells close to the expected target position, based on the expected dead-reckoning errors, in order to guarantee a monotonically decreasing target map entropy. This would make a termination condition based on probability thresholding (e.g., Def.7) more reliable under modest dead-reckoning errors, despite potentially increased target localization bias. At that point, target re-localization could take place within this region, to refine the target position.

6. Conclusions

We have proven that the active object localization problem and a number of its variants, are NP-Hard or NP-Complete. We have studied the tradeoffs of localizing vs. detecting a target object under single-view and multi-view recognition schemes. We have shown that a number of bias/variance/entropy relationships and tradeoffs emerge under single-view and multi-view localization and detection schemes, that depend on the quality of the recognition algorithm and the magnitudes of the correspondence or dead-reckoning errors. The results motivated a set of properties for active detection and localization algorithms.

References