Simultaneous Object Class and Pose Estimation for Mobile Robotic Applications with Minimalistic Recognition

Alper Aydemir, Adrian N. Bishop and Patric Jensfelt

Abstract—In this paper we address the problem of simultaneous object class and pose estimation using nothing more than object class label measurements from a generic object classifier. We detail a method for designing a likelihood function over the robot configuration space. This function provides a likelihood measure of an object being of a certain class given that the robot (from some position) sees and recognizes an object as being of some (possibly different) class. Using this likelihood function in a recursive Bayesian framework allows us to achieve a kind of spatial averaging and determine the object pose (up to certain ambiguities to be made precise). We show how inter-class confusion from certain robot viewpoints can actually increase the ability to determine the object pose. Our approach is motivated by the idea of minimalistic sensing since we use only class label measurements albeit we attempt to estimate the object pose in addition to the class.

I. INTRODUCTION

Object search (or active visual (object) search) is an important component of a mobile robot’s action space [1]. For example, finding, identifying and localizing the pose of objects is a prerequisite for a robot that wishes to interact with objects in the environment. In [2] it is shown that a human understanding of space is significantly based on the objects present in the scene.

In this paper we are interested in the problem of simultaneous object class and pose estimation using a generic object classifier and a spatially dependent measurement likelihood model. One novelty we claim is the ability to estimate both the class and pose of objects in the environment given only measurements of the object class. Indeed, we attempt to push the limits of what information can be estimated given nothing more than a simple object class return and a model of the spatial likelihood for that class return.

Most existing classification and pose estimation algorithms match local geometric features of the object, such as corners, edges, holes and surfaces to a precise geometric model of the object [3]–[7]. Such techniques require extensive storage and training data and are far from minimalistic. In addition, these approaches are sensitive to object occlusions etc where object class measurements are possible but precise geometrical measurements of the object are not possible.

Other techniques [6], [8], [9] use a large number of labeled images taken from different poses and attempt to match specific images in order to determine the pose. The accuracy of these approaches increases with the amount of training and reference images. This technique critically ignores the relative geometry of the sensor and the object and the affect of this relationship on the likelihood of recognizing certain views (or more generally whole objects). In particular, we show how we can achieve more with much less input if we consider this relationship explicitly.

1) Original Contribution: We differ from existing object search methods in the design of our spatial likelihood functions. We will highlight throughout the paper that one novelty of our approach is that it is truly minimalistic in nature. By measuring only the object class label we attempt to extract both the true object class and object pose (orientation and location). Indeed, we generally ignore the notion of object view and assign class labels only to entire objects. We certainly ignore any geometrical aspects of the object and we employ generic classifiers (we ignore the particular features employed by the classifier and indeed in our experiments we use a recognition algorithm from the literature which does not make use of any geometrical model of the object). We can extract certain estimates of the object pose purely from the structure of the likelihood functions which are defined over the robot configuration space and hence are geometrically related to the object pose. In addition, we show how inter-class confusion, e.g. the ability to mistakenly measure multiple class labels for a particular object type from certain views, can be advantageous to the estimation problem (specifically the estimation of object pose). We can achieve a high degree of accuracy in pose estimation using our technique and exploiting inter-class confusion. As far as we are aware our technique is novel and truly minimalistic.

2) Paper Outline: This paper is organized as follows. In the next section we outline the general notation used throughout the paper along with the basics of the robot dynamic model considered. In Section III we formulate the problem and outline the design of the likelihood function. We also provide some intuition regarding the design of the spatial likelihood function through example. Furthermore, in Section III we outline the recursive Bayesian algorithm for computing an objects pose and class and we highlight the algorithm behaviour using a simple toy example. We show how measurements of the class label alone can be used to determine accurately the object pose given a suitable likelihood function defined over the robot configuration space. We then outline an extension of the algorithm in Section IV for object class and orientation estimation over a grid. In Section V we provide the results of a practical experiment over a grid and in Section VI we discuss the results and directions for future work. Our conclusion is given in Section VII.

The authors are with the Centre for Autonomous Systems (CAS) at the Royal Institute of Technology (KTH), Stockholm, Sweden. This work was supported by the Swedish Foundation for Strategic Research (SSF) through CAS and also via the EU FP7 project CogX.
II. PRELIMINARIES

In this section we outline some notational preliminaries and the robot dynamical model considered.

A. Notation

Introduce a global coordinate frame $C$ at some pre-defined time $t_0$. Consider a set of objects $O = \{o_1, \ldots, o_n\}$ with position $x_i \in \mathbb{R}^2$ and orientation $\phi_i \in S^1$. Consider an arbitrary object $o_i$ placed so that the $\mathbb{R}^2$ location of the object’s center hovers over the origin of $C$ at $t_0$. Introduce a local two-dimensional coordinate frame $C_i$ at the center of $o_i$. Then the orientation $\phi_i$ is defined as the relative rotation of $C_i$ with respect to $C$. Each object $o_i$ belongs to a class $\{c_j\}_{j=1}^n$ or the unclassified or non-object class $c_0$.

The position of a single mobile robot is denoted by $s \in \mathbb{R}^2$ with heading $\theta \in S^1$. The distance between the robot and $o_i$ is given by $r_i = |x_i - s|$. The relative direction to $o_i$ from the robot’s heading is given by $\theta_i = \alpha_i - \theta$ where $\alpha_i$ is the azimuthal bearing to $o_i$ in the global coordinate system and $\theta_i \in S^1$. We then define a viewpoint $p_i = [s, \theta_i]^T$.

B. Dynamics

Introduce the matrix Lie group $\mathbb{SE}(2)$ with group element $X(\psi, q) \in \mathbb{SE}(2)$ with $q = [x, y]^T \in \mathbb{R}^2$ and a group (matrix) multiplication operator. An element $X(\theta, r) \in \mathbb{SE}(2)$ acts on a point $p_i \in \mathbb{R}^2$ by mapping it to $(R(\theta)p_i + r) \in \mathbb{R}^2$. Here $(R(\theta)$ is the rotation matrix defined as

$$R(\psi) = \begin{bmatrix} \cos \psi & -\sin \psi \\ \sin \psi & \cos \psi \end{bmatrix}$$

and note that all elements in $\mathbb{SO}(2)$ are congruent to such a matrix. For notational brevity we write the action of $X(\psi, r)$ on $q$ as

$$X(\psi, r) \circ q = R(\psi)q + r$$

which constitutes a left action of $\mathbb{SE}(2)$ on $\mathbb{R}^2$. The inverse $X^{-1}(\theta, r) \in \mathbb{SE}(2)$ maps $p_i$ to $R^\top(\theta)p_i - R^\top(\theta)r$ and the identity is given by $X(0, 0) \in \mathbb{SE}(2)$.

Associated with $\mathbb{SE}(2)$ is the vector space $se(2)$ which is a Lie algebra with respect to the Lie bracket operation. We define the basis of $se(2)$ by $\{E_x, E_y, E_{\psi}\}$ with

$$E_i = \begin{cases} 0 & 0 & 1 \quad (i = x) \\ 0 & 0 & 1 \quad (i = y) \\ 0 & 0 & 0 \end{cases}, \quad E_{\psi} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

with $i \in \{x, y\}$ and where $1(\cdot)$ is an indicator function.

Given translational and angular velocity control inputs $u_1 = v$ and $u_2 = \omega$ we then have

$$\dot{X}(\theta, t) = X(\theta, t)(E_x u_1 + E_{\psi} u_2)$$

which constitutes a left-invariant, drift-free system on the group $\mathbb{SE}(2)$. This model is the Lie group representation of the unicycle model and is our robot kinematic model.

III. PROBLEM FORMULATION

In this section we outline the probabilistic framework within which our estimation problem is formulated.

A. Classification Likelihoods on Lie Groups

For each $p(t)$ the robot takes measurements of the potential class of $o_i$ in the form

$$y_i(t) = [c_j \ldots c_k]^T$$

with $y = [y_1^T \ldots y_n^T]^T$. This means that a measurement of object $o_i$ can return more than a single class value.

We model the likelihood of measuring $c_j$ for $o_i$ as a function of the robot pose. In fact, we model this likelihood as a sum of Gaussian densities on the Lie group $\mathbb{SE}(2)$. Consider an arbitrary normal density in $\mathbb{R}^n$ of the form

$$\gamma(x - \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^\frac{1}{2}} \exp\left(\frac{1}{2}(x - \mu)^\top \Sigma^{-\frac{1}{2}} (x - \mu)\right)$$

where $\Sigma$ is the covariance matrix and $\mu$ is the mean. A Gaussian distribution on the Lie group $\mathbb{SO}(2)$ is given by

$$\chi(x - \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

and if $\sigma^2 < 2\pi$ in $\chi(x - \mu, \sigma^2)$ then $\chi(x - \mu, \sigma^2)$ can be approximated well by the case $k = 0$ in (7). A Gaussian on the product space $\mathbb{SE}(2)$ can then be denoted by $\zeta(x - \mu, \Sigma)$.

We state the following lemma for completeness.

Lemma 1 ([10]): There exists an integer $m$ and constants $w_i > 0$ with $\sum_{i=1}^m w_i = 1$, such that the Gaussian sum

$$p_{\text{approx}}(x) = \sum_{i=1}^m w_i \gamma(x - \mu_i, \Sigma_i)$$

can approximate any density function $p(x)$ as closely as desired in the sense that $\int_{\mathbb{R}^n} |p(x) - p_{\text{approx}}(x)| \, dx$ can be made arbitrarily small.

Recall that the element $X(\psi, r) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ acts on points via left translation denoted by $X(\psi, r) \circ q$. For notational brevity we introduce the following notational definition

$$X(\psi, r) \circ [q_1 q_2 \ldots q_n]^T = [R(\psi)q + r q_2 q_3 \ldots q_n]^T$$

which means that $X(\psi, r) : \mathbb{R}^2 \times A \rightarrow \mathbb{R}^2 \times A$ by acting on the first two dimensions in the standard way and leaving the remaining $n - 2$ dimensions unchanged.

We model the likelihood function by

$$p(c_i, o_j | c_i, o_j, x_j) = P(c_i, o_j | c_i) \sum_{k_i} \frac{w_k}{(2\pi)^{\frac{n}{2}}|\Sigma_k|^{\frac{1}{2}}} \times \exp\left(\frac{1}{2}|\Sigma_k^{-\frac{1}{2}}(p_j - X(\phi_j, x_j) \circ q_{k_i})|^2\right)$$

where $\sum_{k_i} w_k = 1$ with $i = \{1, \ldots, n_c\}$. For an object with position $x_j$ and orientation $\phi_j$ we define $X(\phi_j, x_j)q_{k_i}$.

\footnote{For example, observing a car from the front may yield several positive car model class returns. We assume perfect data association, i.e. we know which objects $o_i$ generate particular class measurements.}
with \( X(\phi_j, x_j) \in \text{SE}(2) \) and \( \mathbf{q}_{bk} \in \mathbb{R}^2 \times \text{SO}(2) \) as the mean of the \( k^{th} \) Gaussian and \( \Sigma_{bk} \) is the covariance\(^2\).

The term \( P(\tilde{c}_i, o_j | c_i) \) specifically deals with the likelihood of \( o_j \) being of class \( c_i \) given the measurement \( \tilde{c}_i \) whereas \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) / P(\tilde{c}_i, o_j | c_i) \) is the likelihood of \( o_j \) being in position \( x_i \) with orientation \( \phi_j \).

The likelihood \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \) is also a probability density function such that for bounded regions of \( A \) of \( \text{SE}(2) \) with positive Lebesgue measure the integral

\[
\int_A p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \, dp
\]

(11)
gives the probability of measuring \( \tilde{c}_i \) for \( o_j \) given that \( o_k \) is of class \( c_i \) and with orientation \( \phi_j \) and position \( x_j \). We state this explicitly since we will require that the so-called confusion densities \( p(\tilde{c}_i, o_k | c_i, \phi_k, x_k) \) with \( i \not= j \) satisfy the inequality

\[
\int_A p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \, dp \leq \int_A p(\tilde{c}_i, o_k | c_i, \phi_k, x_k) \, dp \quad (12)
\]
or \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \leq p(\tilde{c}_i, o_k | c_i, \phi_k, x_k) \) for all bounded subsets \( A \) of \( \text{SE}(2) \) in a defined region of interest \( R \subset \text{SE}(2) \). That is, over any bounded region in \( R \) we want the probability of measuring \( \tilde{c}_i \) for \( o_j \) to be less than (or equal to) the probability of measuring \( c_i \) given that the object is of true class \( c_i \) (and for all object poses).

Of course, this inequality cannot be satisfied over all of \( \text{SE}(2) \) if the confusion likelihood is also required to be a true density function. But to make this definition consistent we note that when viewed as a likelihood function \( p(\tilde{c}_i, o_j | c_k, \phi_j, x_j) \) is valid as long as it is congruent to a probability density function via multiplication by a constant.

We model the likelihood function of false positives by the following Gaussian mixture

\[
p(\tilde{c}_i, o_j | c_k, \phi_j, x_j) = \sum_{k_{ik}} \text{common}(k_{ik}, k_{bk}) \, w_{bk} \times 
\exp \left( \frac{1}{2} \| \Sigma_{bk}^{\frac{1}{2}} (p - X(\phi_j, x_j)) \circ q_{bk} \|_2^2 \right) P(\tilde{c}_i, o_j | c_k)
\]

(13)

where the sum over \( k_{ik} \) has at most \( \min(k_i, k_k) \) terms and (12) must hold in \( R \subset \text{SE}(2) \). The function

\[
\text{common}(k_{ik}, k_{bk}) \in \{0, 1\}
\]

(14)
captures the fact that an object \( o_j \) can be confusingly observed as \( \tilde{c}_i \) and/or \( \tilde{c}_k \) from some robot positions because

\(^2\)We note at this point that the Gaussian parameters \( \mathbf{q}_{bk} \) and \( \Sigma_{bk} \) are defined based on the training scheme of the object classifier and \( p_{bj} \) is the robot position when the relevant class labels are measured. Heuristically, \( \mathbf{q}_{bk} \) is taken to be (one of) the sensor’s position in \( \text{SE}(2) \) at training time relative to the object (which is located at the origin during training with the reference orientation). The variance is (in this paper) tuned to provide a realistic model of the spatial dependence of the recognition algorithm at run time to the trained classifier models. In the next subsection we provide an example further illustrating how the likelihood functions are created. However, we note here that the motivation for these likelihood functions is motivated from experience where we have noticed that often simply by measuring the class label for an entire object (not view point) we most likely restricted one of a small number of points. In reverse, given a known robot position, the object is most likely in one of a small number of locations with one of a small number of orientations.

the underlying true classes \( c_i \) and \( c_k \) share a common indistinguishability from such locations\(^3\).

The class \( c_0 \) is used to model unclassified objects or locations in space where no object exists. The likelihood \( p(\tilde{c}_i, o_j | c_0, \phi_j, x_j) \) where \( i \not= 0 \) is given by

\[
p(\tilde{c}_i, o_j | c_0, \phi_j, x_j) = P(\tilde{c}_i, o_j | c_0)
\]

(15)

which although not a true likelihood function is valid over any bounded region \( R \subset \text{SE}(2) \) since it is congruent to a uniform density over \( R \). For all classes for which it is defined we now require \( \sum_k p(\tilde{c}_i, o_j | c_k) = 1 \).

For much of the space \( \text{SE}(2) \) the object recognizer will not return any class value for \( o_j \). We can (if desired) model the absence of any returns in \( \{c_1, \ldots, c_n \} \) as a measurement of the dummy class \( c_0 \). We would then need to construct the likelihood \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \). We do not explore the design of this likelihood in detail since we will not (in our implementations) incorporate dummy measurements when no class is detected\(^4\).

If we define \( c = [c_0, c_1, \ldots, c_n] \) then the likelihood

\[
p(\tilde{c}_i, o_j | c, \phi_j, x_j) = \sum_k p(\tilde{c}_i, o_j | c_k, \phi_j, x_j)
\]

(16)

is the multi-dimensional likelihood function of the object being in all of the defined classes and all poses given a particular class return. Given a return measurement \( y_i(t) = [\tilde{c}_0, \tilde{c}_1, \ldots, \tilde{c}_2]^\top \) for object \( i \) then the joint likelihood is

\[
p(y_j, o_j | c, \phi_j, x_j) = \prod_{\tilde{c}_k \in y_j} p(\tilde{c}_k, o_j | c, \phi_j, x_j)
\]

under a naive Bayesian assumption, i.e. under the assumption that \( p(\tilde{c}_k, o_j | c, \hat{c}_j) = p(\tilde{c}_k, c | \hat{c}_j) \).

B. Example Likelihood Functions

We now provide some intuition regarding the design of the likelihood functions. These examples are simplified but illustrate the heuristics behind the likelihood structure.

Consider an object \( o_1 \) of class \( c_1 \) located at the origin at time \( t_0 \) with defined orientation \( \phi_1 = 0 \). An object classifier is trained on object \( o_1 \) from a number of relative positions denoted by \( \mathbf{q}_{1k} \) with \( \mathbf{q}_{1k} = [q_{1k}, \phi_{1k}]^\top \). For the \( k^{th} \) training position we define a Gaussian \( \mathcal{N}(p_{1k} - X(\phi_1, x_1), \mathbf{q}_{1k}, \Sigma_{k_1}) \) where \( \Sigma_{k_1} \) is tuned based on the specific classifiers properties\(^5\). We define \( p(\tilde{c}_1, o_1 | c_1, \phi_1, x_1) \) as the sum of such Gaussians as in (10) with \( w_{1k} = 1/4 \) and \( P(\tilde{c}_1, o_1 | c_1) = 1 \). In this example we set \( q_{11} = [10 0 0]^\top, q_{12} = [-10 0 0]^\top, q_{13} = [0 10 0]^\top \) and \( \Sigma_{k_1} = \text{diag}(10, 10, \pi/4) \). Now consider a random object \( o_1 \) at \( x_1 \) =

\(^3\)Such a case happens, for example, if there are object types which look very similar or share a similar internal representation (ambiguous objects or object-data) from certain views.

\(^4\)The reason we do not generate dummy measurements is that \( c_0 \) will, in general, provide little information about the true class \( c_j \) (including \( c_0 \)) for most robot positions \( p \) and objects \( o_j \). Heuristically, over bounded regions the likelihood \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \) would resemble a constant minus the sum (16). Of course this would require some further justification in order for \( p(\tilde{c}_i, o_j | c_i, \phi_j, x_j) \) to be valid as a likelihood function.

\(^5\)We discuss later that an interesting direction for future work is the design of reinforcement learning schemes for tuning such parameters.
which the confusion is likely are not maximal. Consider an

can be aid the pose estimation problem if the views from

specific estimation technique in the next subsection and later

provide examples of the pose accuracy that can be achieved.

However, we now demonstrate how inter-class confusion

can be confusedly recognized as both

and as a result \( \phi_i \) can be determined only up to rotations modulo

\( \pi/2 \) given measurements of the class \( \phi_i \). We outline the

specific estimation technique in the next subsection and later

provide examples of the pose accuracy that can be achieved.

Figure 1 shows from which positions in space relative to

target object \( o_i \) the likelihood of the class of \( o_i \) being

class \( c_1 \) is given that we measure \( \hat{c}_1 \) (and in essence assume

robot relative heading \( \vartheta_i \), invariance - e.g. this will hold

for omni-directional cameras). Note that the likelihood just
demonstrated is symmetric in terms of the orientation \( \vartheta_i \),

and \( \phi_i \) is given in Figure 1.

consider a random object \( o_2 \) located at the origin at time \( t_0 \) with

defined pose \( \phi_2 = 0 \). An object classifier is trained on object

\( o_2 \) from a number of relative positions denoted by \( \mathbf{q}_{2k} \) and

for the \( k^{th} \) position we define a Gaussian \( \zeta_p(\mathbf{p}_2 - \mathbf{x}_2, \Sigma_2) \),

Then we define \( \hat{c}_2, \phi_i, \vartheta_i, x_i \) as the sum of

such Gaussians as in (10) with \( w_{2k} = 1/4 \) and \( P(\hat{c}_2|c_2) = 1 \).

In this example we set \( \mathbf{q}_{2k} = \mathbf{q}_{1i} \) and \( \Sigma_{2k} = \Sigma_{1k} \).

Now consider a random object \( o_i \) at \( x_i = [5, 5] \) with \( \phi_i = 45^\circ \).

The plot of \( p(\hat{c}_2, o_2|c_2, \phi_i, x_i) \) with \( \vartheta_i = 0 \) over \( s \in \mathbb{R}^2 \) is identical in this case to the likelihood shown in Figure 1.

Now suppose from a number of positions we know that

\( \phi_1 \) and \( \phi_2 \) can be confusedly recognized as both \( c_1 \) and \( c_2 \)
in some instances. In this example,

\[
\text{common}(21, 22) = 1
\]

and all other \( \text{common}(\cdot) \) equal to zero. We let \( \Sigma_{k_1, 2} = \Sigma_{k_2} = \Sigma_{k_1} \) and \( w_{k_1, 2} = 1/4 \). Also let \( P(\hat{c}_1|c_1) = P(\hat{c}_2|c_2) = 1/2 \) and now let \( P(\hat{c}_j|c_1) = 1/2 \).

Then \( p(\hat{c}_1, o_2|c_2, \phi_i, x_i) \), for \( j = \{1, 2\} \), with \( \vartheta_i = 0 \) over \( s \in \mathbb{R}^2 \)
is the same shape as in Figure 1 except both likelihoods are

weighted by \( 1/2 \). In Figure 2 we plot \( p(\hat{c}_1, o_1|c_2, \phi_i, x_i) = \)

\( p(\hat{c}_2, o_1|c_2, \phi_i, x_i) \) with \( \vartheta_i = 0 \) over \( s \in \mathbb{R}^2 \) for the same

random object \( o_i \) at \( x_i = [5, 5] \) with \( \phi_i = 45^\circ \).

Given a random object \( o_i \) of class \( c_1 \) or \( c_2 \) we now

gain some intuition about how class confusion can aid in

removing any ambiguity regarding the pose of the object.

For example, if \( o_i \) were viewed from a number of robot

positions around \( \mathbf{X}(\phi_i, x_i) \) \( o\mathbf{q}_{2i} \), i.e. around the confusion

peak, and both \( \hat{c}_1 \) and \( \hat{c}_2 \) were measured at these positions

then the likelihood of the object pose would be (significantly)
dominated by a single mode at the true pose. We will explore

a detailed toy example illustrating this property later. We

will also explore a practical example showing a real-world

experimental result.

C. Maximum A Posterior Probabilities

In terms of Bayes’ rule we know

\[
p(c_i, \phi_j, x_j|\hat{c}_i, o_j) = \frac{p(\hat{c}_i, o_j|c_i, \phi_j, x_j)p(c_i, \phi_j, x_j|o_j)}{p(\hat{c}_i, o_j)} \tag{19}
\]

or in terms of \( y(t) \) we have

\[
p(c_i, \phi_j, x_j|y_j, o_j) = \frac{p(y_j, o_j|c_i, \phi_j, x_j)p(c_i, \phi_j, x_j|o_j)}{p(y_j, o_j)} \tag{20}
\]

where the denominator is given by

\[
p(y_j, o_j) = \int_{\text{SE}(2)} p(y_j, o_j|c, \phi_j, x_j)p(c, \phi_j, x_j|o_j) \, d\phi d\mathbf{x}_j \tag{21}
\]

Note we have neglected illustrating the dependence on

time but the recursion is clear with the prior \( p(c_i, \phi_j, x_j) \)
at time \( t \) equal to the posterior \( p(c_i, \phi_j, x_j|y_j, o_j) \) computed

at some time \( \tau < t \). We also know that

\[
P(c_i|y_j, o_j) = \int_{\text{SE}(2)} p(c_i, \phi_j, x_j|y_j, o_j) \, d\phi d\mathbf{x}_j \tag{22}
\]

is the posterior probability of object \( o_j \) being of class \( c_i \)
given the measurements \( y_j \) and \( \sum_i P(c_i|y_j, o_j) = 1 \) where

\( i = 0 \) can be included naturally. For any \( o_j \) we then have

\[
\sum_i \int_{\text{SE}(2)} p(c_i, \phi_j, x_j|y_j, o_j) \, d\phi d\mathbf{x}_j = 1 \tag{23}
\]

where the sum is taken over all classes \( c_0 \) to \( c_n \).

If we want the maximum a posterior (MAP) class and

object pose estimates via

\[
\{\hat{c}_i, \hat{o}_j, \hat{x}_j\} = \arg \max_i \{p(c_i, \phi_j, x_j|y_j, o_j)\} \in \{1, \ldots, n_c\} \tag{24}
\]

where \( \hat{c}_i \) is the MAP class estimate for object \( j \) corresponding
to the maximization argument index \( i \).

In general, (24) leads to \( n_c \) maximization problems for

each \( o_j \). Each density is often multi-modal but each
mode can be determined easy via grid-search. If \( \hat{c}_k \notin y_j(t) \) and \( p(\hat{c}_k, o_j| c_i, o_j, x_j) = 0 \) for all \( i \neq k \) then \( p(c_i, o_j, x_j| y_j, o_j) = 0 \) at time \( t \) and at least one maximization problem is avoided.

D. Bringing it All Together with a Toy Example

A toy example is now examined in order to further develop an intuition regarding the approach outlined in this paper. A more detailed practical experiment is given later in the paper.

The fact we can localize the pose of the object accurately (even up to an ambiguity determined by the number of Gaussians in the likelihood function) is quite novel given we only use class label measurements. However, we go further then this and show how class confusions (from certain view points) can even reduce the number of ambiguities.

Consider an object \( o_1 \) located at \( x_1 = [5, 5]^\top \) with true orientation \( \phi_1 = 45^\circ \). Consider two potential object classes \( c_1 \) and \( c_2 \) with defined likelihood functions

\[
p(\hat{c}_i, o_1| c_i, \phi_1, x_1) = \frac{0.2495}{(2\pi)^2|\Sigma_{i,1}|^{1/2}} \exp\left( -\frac{1}{2} \|\Sigma_{i,1}^{-1/2}(p - X(\phi_1, x_1)q_{1,i})\|^2 \right) + \frac{0.2495}{(2\pi)^2|\Sigma_{i,2}|^{1/2}} \exp\left( -\frac{1}{2} \|\Sigma_{i,2}^{-1/2}(p - X(\phi_1, x_1)q_{2,i})\|^2 \right)
\]

for both \( i = 1 \) and \( i = 2 \) (with \( P(\hat{c}_i| c_i) = 1/2 - 0.001 \) as a consequence). The mean parameters are given by \( q_{1,i} = [0, 0, 100]^\top \) and \( q_{2,i} = [0, 0, -100]^\top \). The false positive likelihoods are given by

\[
p(\hat{c}_i, o_1| c_j, \phi_1, x_1) = \frac{0.2495}{(2\pi)^2|\Sigma_{i,j}|^{1/2}} \exp\left( -\frac{1}{2} \|\Sigma_{i,j}^{-1/2}(p - X(\phi_1, x_1)q_{1,i})\|^2 \right)
\]

with \( i \neq j \in \{1, 2\} \) and \( q_{1,j} = [0, 0, 100]^\top \) (and \( P(\hat{c}_j| c_i) = 1/2 - 0.001 \)). The variance is given by \( \Sigma_{i,j} = \text{diag}(10, 10, \pi/4) \) for all combinations of \( i \) and \( j \). Now consider the class \( c_0 \) with

\[
p(\hat{c}_i, o_1| c_0, \phi_1, x_1) = P(\hat{c}| c_0) = 0.001
\]

for \( i \in \{1, 2\} \). The recognition system can return class measurements \( c_1 \) and \( c_2 \).

We plot \( p(\hat{c}_i, o_1| c_1, \phi_1, x_1) \) and \( p(\hat{c}_i, o_1| c_2, \phi_1, x_1) \) with \( i \neq j \in \{1, 2\} \) evaluated at \( x_1 = [5, 5]^\top \) and \( \phi_1 = 45^\circ \). This shows the relationship between the robot position and the likelihoods.

In this example we assume \( x_1 \) is known but the true object class and orientation \( \phi_1 \) is unknown. This is a reasonable approximation in many active object search problems. In the next section, we consider a grid-based object search and orientation estimation problem where this assumption is explicitly realized.

The initial priors are thus \( p(c_1, \phi_1, x_1|o_1) = 1/(2\pi) \). We simulate measurements at a number of positions in space in order to examine their affect on the posterior densities.

Time 1

The robot position is given by \( p = X(\phi_1, x_1)\circ [2 - 100]^\top \). The measurements are given by \( y_1(1) = [c_1]^\top \). The posterior density functions are shown in Figure 4.

We know \( P(c_i| y_j, o_j) = \int_{SO(2)} p(c_i, \phi_j, x_j| y_j, o_j) \, d\phi_j \) and we can compute \( P(c_0| y_1, o_1) = 0.4231, P(c_1| y_1, o_1) = 0.4244 \) and \( P(c_2| y_1, o_1) = 0.1524 \) all at time 1. Then for the maximum class posterior estimate \( \hat{c}_1 \) we can compute the maximum \( \arg\max_\phi p(c_1, \phi, x_1| y_1(1), o_1) \) which is clearly (up to numerical tolerance) ambiguous with \( \phi_1 \approx 55^\circ \) and \( \phi_1 \approx 235^\circ \). The estimate of \( \hat{c}_1 \) is not overwhelmingly probable and the orientation estimate \( \hat{\phi}_1 \) is not exceedingly accurate since we have only employed a single measurement.

Time 2

The robot is at \( p = X(\phi_1, x_1)\circ [-2 - 100]^\top \). The measurements are given by \( y_1(2) = [c_1]^\top \). The posterior density functions are shown in Figure 5.

We compute \( P(c_0| y_1, o_1) = 0.3019, P(c_1| y_1, o_1) = 0.5735 \) and \( P(c_2| y_1, o_1) = 0.1247 \) at time 2. Then for the maximum class posterior estimate \( \hat{c}_1 \) we compute the maximum \( \arg\max_\phi p(c_1, \phi, x_1| y_1(1), o_1) \) which is again (up to numerical tolerance) ambiguous with \( \phi_1 \approx 45^\circ \) and \( \phi_1 \approx 225^\circ \). However, now the orientation estimate is accurate up to the ambiguity. The increased accuracy in the orientation (neglecting the ambiguity) is a result of the spatial averaging that occurs when observing the object from different robot positions (and this accuracy is quite interesting given we only physically measure the class label).
In the next time step we move to the other side of the object and show how confusion aids in removing the ambiguity.

Time 3

The robot position is given by $p = X(\phi_1, x_1) \in [0, 10, 0]^T$. The measurements are given by $y_1(3) = [\tilde{c}_1, \tilde{c}_2]^T$. The posterior density functions are shown in Figure 6.

![Figure 5. The posteriors densities after the second measurement.](image)

![Figure 6. The posteriors densities after the third measurement.](image)

We compute $P(o_0|y_1, o_1) = 0.1502$, $P(o_1|y_1, o_1) = 0.8498$ and $P(o_2|y_1, o_1) = 1.8 \times 10^{-5}$ at time 3. Then for the maximum class posterior estimate $\hat{c}_1$ we compute the maximum $\arg\max_{\phi_1} p(c_1, \phi_1, x_1|y_1(1), o_1)$ which is now unique $\phi_1 \approx 45^\circ$. The orientation estimate is non-ambiguous in this case since we exploited inter-class confusion.

Note that we have estimated the orientation quite accurately using only measurements of the object class label and a pre-defined heuristic spatial likelihood function. We believe this is a novel result in the sense of minimalistic sensing.

IV. GRID-BASED OBJECT CLASSIFICATION AND ORIENTATION ESTIMATION

Consider a grid on $\mathbb{R}^2$ denoted by $\mathcal{G}$. For simplicity, we assume the grid $\mathcal{G}$ consists of $n_g$ grid squares of uniform size (the generalization to nonuniform grid cells is straightforward). Each grid square is denoted by $g_i \in \mathcal{G}$ and can be characterized by the center point $g_i \in \mathbb{R}^2$. We are interested in assigning to each cell $g_i$ the probability density $p(c_i, \phi_i, x_i|y_j, g_j)$ from which we can determine the probability $P(c_i|y_j, g_j)$ via marginalization. In fact, for each cell we assign $n_c$ such probability densities - one for each class. Then $\sum_{i} P(c_i|y_j, g_j) = 1$ where $i = \{0, \ldots, n_c\}$ for each cell. In practice a lot of the cells will be dominated by the probability value $P(c_0|y_j, g_j)$.

In this scenario, $x_j$ is the location of the $j$th cell $g_j$ and is known. If we imagine a robot located at $s$ with $\theta_j$ the direction to $o_j$ defines a ray which we limit to the length $d$. We update the set of cells $\{g_j\}$ that intersect such a ray using the posterior density formula given in the previous sections. The value $x_j$ is taken as the cell center $g_j \in \mathbb{R}^2$ and thus cells close or far from the robot (along the ray) are likely to be estimated as $c_0$. We could also define a conic region, e.g. by defining two rays using the bounding box of the object in the image, and then update the cells which intersect the conic region, e.g. see Figure 7.

![Figure 7. An example grid environment.](image)

The grid-based estimation problem follows closely the examples given in the last subsection where the location of each cell is known and is analogous to the location of an object position. For simplicity we have assumed cell independence. It is possible to relax this assumption but there are difficulties in doing so that are beyond the scope of this paper. For the grid-based scenario we will examine a practical experiment which is outlined in the next section.

V. EXPERIMENTAL RESULTS

We conduct our experiments on the EU FP7 project CogX robot platform. The robot is equipped with a Point Grey Flea stereo camera (only one camera used in the experiment) on top of a Pioneer P3X robot base; see Figure 8.

![Figure 8. All objects have the same lab logo on one side.](image)

We use FERNS as an object class detector [11], [12]. The robot position is computed from only odometry and the grid organization is known (each grid cell is 2 square decimeters).

The robot is in a room with three objects, $o_1$ is a box containing physics books and $o_2$ or $o_3$ are identical boxes containing robot parts. The boxes are located as shown in Figure 8. All objects have the same CAS lab logo on one of their sides and cannot be differentiated based on the class returns when viewed from this side. We call this the confusion side of the object. On the polar opposite side, $o_1$ exhibits a label indicating the box contains physics books whereas both $o_2$ and $o_3$ contain identical labels indicating
they contain robot parts. Views of this distinguishing box label are said to be of the non-confusion side of the object. The true orientations for \( o_1, o_2 \) and \( o_3 \) are \( \phi_1 = 255^\circ, \phi_2 = 315^\circ \) and \( \phi_3 = 180^\circ \).

The robot starts at \((0,0)\) and follows the trajectory shown in Figure 8. Numbered positions in Figure 8 are where the robot takes a class label measurement. The non-object class, physics books box and robot parts box are labeled as \( c_0, c_1 \) and \( c_2 \) respectively. In this scenario the units are decimeters. As such, the likelihood functions \( p(\tilde{c},o_i|\phi_j,\phi_k,x_j) \) for \( i \in \{1,2\} \) and \( j \in \{1,2,3\} \) are identical to those defined in (25) in the simulated example problem. Similarly, \( p(\tilde{c},o_i|c_k,\phi_j,x_j) \) for \( i \neq k \in \{1,2\} \) are identical to the likelihood functions defined in (26). Finally, \( p(\tilde{c},o_i|c_0,\phi_1,x_1) \) for \( i \in \{1,2\} \) is identical to the function defined in (27). The recognition system can of course return class measurements \( \tilde{c}_1 \) and \( \tilde{c}_2 \).

A. Orientation Estimation at the Correct Grid Cell

The estimation algorithm in this section is run over a grid as discussed in the last section. However, to visualize the orientation estimate’s density we need to essentially look at an individual cell. Thus, in this subsection we examine the orientation estimate in the practical experiment at the true object grid cell. Later we examine the grid map for the environment and show the distribution of the class label probabilities over a number of cells.

At point 1, the robot detects \( o_2 \) on its confusion side, i.e. both \( \tilde{c}_1 \) and \( \tilde{c}_2 \) are measured. The resulting orientation estimates for each class are shown in Figure 10 part (a). Since both \( \tilde{c}_1 \) and \( \tilde{c}_2 \) are detected and no further information is available, the probability estimates for both classes are equal but the maximum a posterior orientation estimate is non-ambiguous. The orientation estimate \( \phi_2 \approx 317^\circ \) which is relatively close to the true orientation estimate.

At point 2, the robot detects \( o_3 \) from a non-confusion side; see Figure 9. The class measurement is only \( \tilde{c}_2 \) since the observed side is a discriminative one. However notice that the orientation estimate is multi-modal with \( \tilde{\phi}_3 \approx 3^\circ \) and \( \tilde{\phi}_3 \approx 183^\circ \). Since this box is on a shelf against a wall, it is not possible to observe it from other sides. We will show in the following how observing the confusion side improves the overall orientation estimate.

At point 3, the robot observes a non-confusion side of \( o_2 \), i.e. only \( \tilde{c}_2 \) is measured which is the true class of \( o_2 \); see Figure 10 part (b). Notice that the probability over \( \phi_2 \) for the class \( c_1 \) has dropped and will continue to do so as more measurements of \( \tilde{c}_2 \) are acquired.

At point 4, the robot detects the \( o_1 \) from its non-confusion side, i.e. only \( \tilde{c}_1 \) is measured which is the true class of \( o_1 \). As with the first measurement of \( o_1 \), we have two peaks for the detected class shown in Figure 11. The orientation estimate is given by \( \tilde{\phi}_1 \approx 57^\circ \) and \( \tilde{\phi}_1 \approx 237^\circ \). The maximum class estimate is \( \tilde{c}_1 \).

At point 5, the robot observes \( o_1 \) from its confusion side. In this case since the object has been detected once from its non-confusion side, the probability of \( o_1 \) being of class \( c_1 \) is now much higher and the orientation estimate is now non-ambiguous with \( \tilde{\phi}_1 \approx 258^\circ \) as shown in Figure 11. We now see that the confusion side helps to eliminate one of the peaks in the orientation estimate and the spatial likelihood function has helped the estimate converge to an accurate value.
Fig. 11. [Top] The robot first measures $c_1$ and then [Bottom] both $c_1$ and $c_2$. Notice even though the confusion of the second measurement improves the orientation estimation.

B. Class Probability Estimation over the Grid Map

As described previously, the estimation algorithm is executed over a grid where the detected object rays define a set of cells updated after each measurement. Each cell is associated with an orientation and class density (for all possible classes). As an example, the marginalized probabilities for $c_0$, $c_1$ and $c_2$ are visualized in Figure 12 for class measurements of $o_2$ (recall we have assumed that particular class measurements can be assigned to the correct object).

In this particular snapshot, $o_2$, which is of true class $c_2$, is seen from two positions (points 3 and 4 in Figure 8) and two rays are cast. The gray shading in each picture along the rays represents the probability $P(c_1, o_2 | y_2)$ for each respective class. We have normalized the shading so $P(c_1, o_2 | y_2) = 0$ is pure white while $P(c_1, o_2 | y_2) = 1$ is pure black. The orange background is used here to simplify visualization but can be thought of as the initial prior class probabilities for all cells (i.e. equal priors for all classes) and remains valid since these cells are not updated given only these measurements.

In Figure 12 we can note the probability $P(c_2, o_2 | y_2)$ along the rays increases in magnitude up until the grid cells located at the approximate object location, i.e. the intersection point of the rays. It then decreases as expected. Similarly, $P(c_0, o_2 | y_2)$ decreases in magnitude along the rays until the intersection where it is almost zero and then begins to increase as expected further away from the object.

VI. CONCLUSION

We have provided a solution to the problem of simultaneous object class and pose estimation using a generic object classifier and a spatially dependent measurement likelihood model. Our novelty is the ability to estimate both the class and pose of the objects in the environment given only measurements of the object class label from a generic classifier.

We believe the heuristics behind the design of the likelihood functions are realistic. However, one practically and theoretically interesting direction for future work includes the development of reinforcement-like learning algorithms for estimating the likelihood function parameters online. Another interesting direction for future work involves the design of control algorithms for actively searching the environment in order to maximize the information gain.

REFERENCES