University of California
Los Angeles

Models and Methods for Sensor-Based Environment Exploration

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Joshua Anthony Hernandez

2015

The dissertation of Joshua Anthony Hernandez is approved.

Luminita Vese

Rafail Ostrovsky

Joseph Teran

Stefano Soatto, Committee Chair

University of California, Los Angeles

2015

Table of Contents

Vita

2012–2015 Graduate Student Researcher, *U*CLA Vision Lab.

- *I*nformation-Driven Exploration of Ising-Modeled Terrain

  Developed a method for efficiently exploring and mapping highly-occluded environments, choosing informative viewpoints using a prior on terrain shape and an efficiently-computed approximation of joint measurement entropy.

- *V*isual-Inertial Navigation

  Implemented efficient numerical methods for the Corvis visual-inertial navigation system. Analyzed the ambiguity inherent in visual-inertial sensor fusion systems, and characterised the set of indistinguishable trajectories.

- *R*epresentation Reduction for Autonomous Agents

  Developed and analyzed algorithms for the lossless compresssion of an autonomous agent's belief state, modulo a given task, with the aim of reducing memory and computational requirements.

Summer 2011 Summer Intern, *I*nstruments Division, Jet Propulsion Lab. Contributed to the onboard image processing system of MSPI (Multiangle SpectroPolarimetric Imager) Satellite. Identified and eliminated an unusual striping artifact characteristic of that system. Advisor: Veljko Jovanovich.

2005–2011 Teaching Assistant, *U*CLA Mathematics. Held twice-weekly discussions for several lower-and upper-division classes, to wit: Precalculus, Calculus 1 & 2, Differential Equations, Complex Analysis, Introductory C++, Algorithms and Data Structures, Advanced A & DS.

Publications

J. Dong, N. Karianakis, D. Davis, *J.* Hernandez, J. Balzer and S. Soatto. Multi-view Feature Engineering and Learning. *P*roc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

J. Hernandez, K. Tsotsos and S. Soatto. Observability, Identifiability and Sensitivity of Vision-Aided Inertial Navigation. *P*roc. International Conference on Robotics and Automation (ICRA), May 2015. *B*est Paper, ICRA 2015

J. Helton, D. McAllaster, *J.* Hernandez. Non-Commutative Harmonic and Subharmonic Polynomials. *I*ntegral Equations and Operator Theory, May 2008, Vol. 61, Iss. 1, pp 77-102.

CHAPTER 1

Representations of the Environment for Robotics Applications

The work in the following chapters addresses different aspects of the problem of autonomous interaction of a robot with physical environment. The need for such autonomous capabilities is evident - autonomous vehicles, drones, and robots, ar safer than humans to drive, can replace humans where tasks are dangerous, boring, etc. There are great difficulties in this work, however. We know how to build robots, sensors, control etc., but we do not know how to endow a robot with a "sense" of the surrounding environment. This sense could be a sense of geometry (where is point A? relative to which reference frame? what is a path to A, are there "obstacles"? what are "obstacles"), localization (where are you in relation to obstacles?), topology (what paths are traversable?), photometry (what is an "object"? How do I find it?). Instead of tackling the problem of autonomous interaction head-on, we have focused on a few subproblems.

1.1 Summary of core results

1. First: localization. To interact with the environment, first you need to know where you are relative to it. Fundamental problem studied for many years; Observability of the undelying dynamical model is a necessary condition for ANY algorithm to work, in the sense of yielding a unique point estimate. Our contribution is to show that all existing analysis of observability was

flawed, and propose new analysis that shows that, contrary to popular belief, pose is not observable from visual and inertial sensors. However, we show that the ambiguous set is bounded, and compute it analytically.

2. Second: Once I know where I am, I need to know what is around me. This is a problem called "mapping". Building geometric maps (point clouds) well explored problem. However, to interact intelligently need more than point cloud, need topology. How is the world around us divided into "objects"? Chapter 4 talks about a way of organizing points into surfaces and then connected components of surfaces, that can be considered "objects" for the purpose of interaction, from video.

3. Once I know where I am and have a model of the (Visible) environment, with respect to which I know the location of an object of interest (point A), I need to know how to get to point A, which may not be visible. This requires exploration. Chapter 1 deals with this problem. The contribution is an efficient algorithm with provable bounds on the exploration time and amenable to be extended to non-compact domains (relevant in vision because one can see to infinity).

4. To explore the boundaries of this problem set, we also ask whether a representation is needed at all, at least for simple problems like going to point A. To this end, we explore the possibility of directly encoding/representing/optimizing the map from sensory data to control action, designed so as to achieve the goal (of getting to point A).

1.2 Overview of the thesis

This work is organized as follows: Chapter 2 analyzes the problem of localization using monocular video and inertial sensors. Chapter 3 deals with the problem of

segmenting 3D pointclouds into task-relevant "objects". Chapter 4 explores the problem of efficient autonomous mapping using range sensors. Chapter 5 looks into schemes for reducing the computational overhead required to maintain an agent's awareness relative to a task.

CHAPTER 2

Observability of Visual-Inertial Navigation

2.1 Introduction

Visually-aided navigation (bearing), and range-aided navigation (radar) can be framed as a filtering problem. The model is non-linear, has unknown parameters, and unknown inputs (*e.g.,* accelerometer and gyrometer bias derivative), typically treated as driving noise in a random walk model. Observability is a necessary condition for *any* filter/observer to operate, hence a literature on observability analysis of visually-aided navigation [38, 58, 35]. Relatively little on range-aided. Unknown parameters are typically included in the state, thus transforming an identification problem into a filtering one, and their identifiability analysis lumped in the observability analysis of the resulting (augmented) model. Noise does not affect the observability of a model, so for the purpose of observability analysis, they are set to zero. This is because, by assumption, noise is "uninformative:" It is typically modeled as a realization of a white zero-mean, homoscedastic process, independent of the state of the model. However, the driving input to the random walk model of accelerometer and gyro bias is typically small but *not* independent of the state. In fact, far from being uninformative, it is strongly correlated with it, as it is its temporal derivative. Thus, it should be treated as an *unknown input*, rather than a "noise." As such, it should be included in the observability/identifiability analysis. Our first contribution is to show that while (a prototypical model of) assisted navigation and auto-calibration is *observable* in the absence of unknown input, it is *not* observable when unknown inputs are taken into account. This

exposes a methodological flaw with the observability analysis of assisted navigation in the existing literature. Our second contribution is to reframe observability as a *sensitivity* analysis, and to show that while the set of indistinguishable trajectories is *not* a singleton (as it would be if the model was observable), but it is nevertheless bounded to a set. We explicitly characterize this set and show that, interestingly, it may not contain the "true" state trajectory. Finally, we provide bounds on the volume of this subset as a function of the characteristics of the unknown inputs. We do so for bearing-only augmentation, range-only augmentation, and combined augmentation. Rather than study observability of linearized system, or algebraically checking the rank conditions, that offers no insight on the structure of the indistinguishable states, we characterize observability directly in terms of indistinguishable sets.

2.1.1 Notation

A reference frame is represented by an orthogonal 3×3 positive-determinant (rotation) matrix $R \in \mathrm{SO}(3) \doteq \{R \in \mathbb{R}^{3\times 3} \mid R^T R = RR^T = I,\ \det(R) = +1\}$ and a translation vector $T \in \mathbb{R}^3$. They are collectively indicated by $g = (R, T) \in \mathrm{SE}(3)$. When g represents the change of coordinates from a reference frame "a" to another ("b"), it is indicated by g_{ba}. Then the columns of R_{ba} are the coordinate axes of a relative to the reference frame b, and T_{ba} is the origin of a in the reference frame b. If p_a is a point relative to the reference frame a, then its representation relative to b is $p_b = g_{ba}p_a$. In coordinates, if X_a are the coordinates of p_a, then $X_b = R_{ba}X_a + T_{ba}$ are the coordinates of p_b.

A time-varying pose is indicated with $g(t) = (R(t), T(t))$ or $g_t = (R_t, T_t)$, and the entire trajectory from an initial time t_i and a final time t_f $\{g(t)\}_{t=t_i}^{t_f}$ is indicated in short-hand notation with $g_{t_i}^{t_f}$; when the initial time is $t_0 = 0$, we omit the subscript and call g^t the trajectory "up to time t". The time-index is sometimes omitted for simplicity of notation when it is clear from the context.

We indicate with $\widehat{V} = (\widehat{\omega}, v) \in \mathfrak{se}(3)$ the (generalized) velocity or "twist", where $\widehat{\omega}$ is a skew-symmetric matrix $\widehat{\omega} \in \mathfrak{so}(3) \doteq \{S \in \mathbb{R}^{3\times 3} \mid S^T = -S\}$ corresponding to the cross product with the vector $\omega \in \mathbb{R}^3$, so that $\widehat{\omega}v = \omega \times v$ for any vector $v \in \mathbb{R}^3$. We indicate the generalized velocity with $V = (\omega, v)$. We indicate the group composition $g_1 \circ g_2$ simply as $g_1 g_2$. In homogeneous coordinates, $\bar{X}_b = G_{ba}\bar{X}_a$ where $\bar{X}^T = [X^T \ 1]$ and

$$G \doteq \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \quad \hat{V} \doteq \begin{bmatrix} \widehat{\omega} & v \\ 0 & 0 \end{bmatrix}. \tag{2.1}$$

Composition of rigid motions is then represented by matrix product.

2.1.2 Mechanization Equations

The motion of a sensor platform is represented as the time-varying pose g_{sb} of the body relative to the spatial frame. To relate this to measurements of an inertial measurement unit (IMU) we compute the temporal derivatives of g_{sb}, which yield the (generalized) body velocity V^b_{sb}, defined by $\dot{g}_{sb}(t) = g_{sb}(t)\widehat{V}^b_{sb}(t)$, which can be broken down into the rotational and translational components $\dot{R}_{sb}(t) = R_{sb}(t)\widehat{\omega}^b_{sb}(t)$ and $\dot{T}_{sb}(t) = R_{sb}(t)v^b_{sb}(t)$. An ideal gyrometer (gyro) would measure $\omega_{\text{imu}} = \omega^b_{sb}$. The translational component of body velocity, v^b_{sb}, can be obtained from the last column of the matrix $\frac{d}{dt}\widehat{V}^b_{sb}(t)$. That is, $\dot{v}^b_{sb} = \dot{R}^T_{sb}\dot{T}_{sb} + R^T_{sb}\ddot{T}_{sb} = -\widehat{\omega}^b_{sb}v^b_{sb} + R^T_{sb}\ddot{T}_{sb} \doteq -\widehat{\omega}^b_{sb}v^b_{sb} + \alpha^b_{sb}$, which serves to define $\alpha^b_{sb} \doteq R^T_{sb}\ddot{T}_{sb}$. These equations can be simplified by defining a new linear velocity, v_{sb}, which is neither the body velocity v^b_{sb} nor the spatial velocity v^s_{sb}, but instead $v_{sb} \doteq R_{sb}v^b_{sb}$. Consequently, we have that $\dot{T}_{sb}(t) = v_{sb}(t)$ and $\dot{v}_{sb}(t) = \dot{R}_{sb}v^b_{sb} + R_{sb}\dot{v}^b_{sb} = \ddot{T}_{sb} \doteq \alpha_{sb}(t)$ where the last equation serves to define the new linear acceleration α_{sb}; as one can easily verify we have that $\alpha_{sb} = R_{sb}\alpha^b_{sb}$. An ideal accelerometer (accel) would then measure $\alpha_{\text{imu}} = R^T_{sb}(t)(\alpha_{sb}(t) - \gamma)$.

There are several reference frames to be considered in an aided navigation scenario. The *spatial frame* s, typically attached to Earth and oriented so that

gravity γ takes the form $\gamma^T = [0\ 0\ 1]^T \|\gamma\|$ where $\|\gamma\|$ can be read from tabulates based on location and is typically around $9.8 m/s^2$. The *body frame* b is attached to the IMU.[1] The *camera frame* c, relative to which image measurements are captured, is also unknown, although we will assume that *intrinsic calibration* has ben performed, so that measurements on the image plane are provided in metric units. Finally, the *radar frame*, or range frame r, is that of the antenna relative to which range measurements are provided.

The equations of motion (known as mechanization equations) are usually described in terms of the body frame at time t relative to the spatial frame $g_{sb}(t)$. Since the spatial frame is arbitrary (other than for being aligned to gravity), it is often chosen to be co-located with the body frame at time $t = 0$. To simplify the notation, we indicate this time-varying frame $g_{sb}(t)$ simply as g, and so for $R_{sb}, T_{sb}, \omega_{sb}, v_{sb}$, thus effectively omitting the subscript sb everywhere it appears. This yields

$$\begin{cases} \dot{T} &= V \\ \dot{R} &= R\widehat{\omega} \\ \dot{V} &= \alpha \\ \dot{\omega} &= w \\ \dot{\alpha} &= \xi \end{cases} \tag{2.2}$$

where $w \in \mathbb{R}^3$ is the rotational acceleration, and $\xi \in \mathbb{R}^3$ the translational jerk (derivative of acceleration). Although α corresponds to neither body nor spatial acceleration, it can be easily related to accel measurements:

$$\boxed{\alpha_{\text{imu}}(t) = R^T(t)(\alpha(t) - \gamma) + \underbrace{\alpha_b(t) + n_\alpha(t)}} \tag{2.3}$$

[1] In practice, the IMU has several different frames due to the fact that the gyro and accel are not co-located and aligned, and even each sensor (gyro or accel) is composed of multiple sensors, each of which can have its own reference frame. Here we will assume that the IMU has ben pre-calibrated so that accel and gyro yield measurements relative to a common reference frame, the *body frame*. In reality, it may be necessary to calibrate the alignment between the multiple-axes sensors (non-orthogonality), as well as the gains (scale factors) of each axis.

where the measurement error in bracket includes a slowly-varying mean ("bias") $\alpha_b(t)$ and a residual term n_α that is commonly modeled as a zero-mean (its mean is captured by the bias), white, homoscedastic and Gaussian noise process. In other words, it is assumed that n_α is independent of α, hence uninformative. Here γ is the gravity vector expressed in the spatial frame.[2] Measurements from a gyro can be similarly modeled as

$$\boxed{\omega_{\text{imu}}(t) = \omega(t) + \underbrace{\omega_b(t) + n_\omega(t)}} \tag{2.4}$$

where the measurement error in bracket includes a slowly-varying bias $\omega_b(t)$ and a residual "noise" n_ω also assumed zero-mean, white, homoscedastic and Gaussian, independent of ω.

Other than the fact that the biases α_b, ω_b change *slowly*, they can change arbitrarily. One can therefore consider them an *unknown input* to the model, or a *state* in the model, in which case one has to hypothesize a dynamical model for them. For instance instance

$$\dot{\omega}_b(t) = v_b(t), \quad \dot{\alpha}_b(t) = v_\alpha(t) \tag{2.5}$$

for some unknown input v_b, v_α. While it is safe to assume that v_b, v_α are *small*, they certainly are not (white, zero-mean and, most importantly) uninformative. Nevertheless, it is common to consider v_b, v_α, to be realizations of a Brownian motion that is *independent* of ω_b, α_b. This is done for convenience as one can then consider all unknown inputs as "noise." Unfortunately, however, this has repercussion on the analysis of the observability and identifiability of the resulting model (Sect. 2.2).

[2]The orientation of the body frame relative to gravity, R_0, is unknown, but can be approximated by keeping the IMU still (so $R^T(t) = R_0$) and averaging the accel measurements, so that $\frac{1}{T}\sum_{t=0}^{T} \alpha_{\text{imu}}(t) \simeq -R_0^T\gamma + \alpha_b$. Assuming the bias to be small (zero), this equation defines R_0 up to a rotation around gravity, which is arbitrary. Note that if $\alpha_b \neq 0$, the initial bias will affect the initial orientation estimate.

2.1.3 Standard and reduced models

The mechanization equations above define a dynamical model having as output the IMU measurements. Including the initial conditions and biases, we have

$$\begin{cases} \dot{T} = V \qquad T(0) = 0 \\ \dot{R} = R\widehat{\omega} \quad R(0) = R_0 \\ \dot{V} = \alpha \\ \dot{\omega} = w \\ \dot{\alpha} = \xi \\ \dot{\omega}_b = n_{\omega_b} \\ \dot{\alpha}_b = n_{\alpha_b} \\ \dot{\gamma} = 0 \\ \omega_{\text{imu}}(t) = \omega(t) + \omega_b(t) + n_\omega(t) \\ \alpha_{\text{imu}}(t) = R^T(t)(\alpha(t) - \gamma) + \alpha_b(t) + n_\alpha(t) \end{cases} \tag{2.6}$$

In this standard model, data from the IMU are considered as (output) *measurements*. However, it is customary to treat them as (known) *input* to the system, by writing ω in terms of ω_{imu} and α in terms of α_{imu}:

$$\boxed{\omega = \omega_{\text{imu}} - \omega_b + \underbrace{n_R}_{-n_\omega}} \qquad \boxed{\alpha = R(\alpha_{\text{imu}} - \alpha_b) + \gamma + \underbrace{n_V}_{-Rn_\alpha}} \tag{2.7}$$

This equality is valid for *samples* (realizations) of the stochastic processes involved, but it can be misleading as, if considered as stochastic processes, the noises above are *not* independent of the states. Such a dependency, is nevertheless typically

neglected. The resulting mechanization model is

$$\boxed{\begin{cases} \dot{T} = V & T(0) = 0 \\ \dot{R} = R(\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b) + n_R & R(0) = R_0 \\ \dot{V} = R(\alpha_{\text{imu}} - \alpha_b) + \gamma + n_V & \\ \dot{\omega}_b = n_{\omega_b} & \\ \dot{\alpha}_b = n_{\alpha_b}. & \end{cases}} \tag{2.8}$$

Next we will consider augmenting the models above with measurement equations coming either from *range* or *bearing* measurements for a finite set N of isolated points with coordinates $X^i \in \mathbb{R}^3,\ i = 1, \ldots, N$.

2.1.4 Bearing augmentation (vision)

Initially we assume there is a collection of points $X^i,\ i = 1, \ldots, N$, visible from time $t = 0$ to the current time t. If $\pi : \mathbb{R}^3 \to \mathbb{R}^2; X \mapsto [X_1/X_3,\ X_2/X_3]$ is a canonical central (perspective) projection, assuming that the camera is *calibrated*,[3] *aligned*,[4] and that the spatial frame coincides with the body frame at time 0, we have

$$\boxed{y^i(t) = \frac{R_{1:2}^T(t)(X^i - T_{1:2}(t))}{R_3^T(t)(X^i - T_3(t))} \doteq \pi(g^{-1}(t)X^i) + n^i(t), \quad t \geq 0.} \tag{2.9}$$

If the feature first appears at time $t = 0$ *and if the camera reference frame is chosen to be the origin the world reference frame* so that $T(0) = 0; R(0) = I$, then we have that $y^i(0) = \pi(X^i) + n^i(0)$, and therefore

$$\boxed{X^i = \bar{y}^i(0)Z^i + \tilde{n}^i} \tag{2.10}$$

where $\bar{y}$ is the homogeneous coordinate of y, $\bar{y} = [y^T\ 1]^T$, and $\tilde{n}^i = [n^{iT}(0)Z^i\ \ 0]^T$. Here Z^i is the (unknown, scalar) depth of the point at time $t = 0$. With an

[3]Intrinsic calibration parameters are known and compensated for.

[4]The pose of the camera relative to the IMU is known and compensated for.

abuse of notation, we write the map that collectively projects all points to their corresponding locations on the image plane as:

$$y(t) \doteq \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{bmatrix}(t) = \begin{bmatrix} \pi(R^T(X^1 - T)) \\ \pi(R^T(X^2 - T)) \\ \vdots \\ \pi(R^T(X^N - T)) \end{bmatrix} + \begin{bmatrix} n^1(t) \\ n^2(t) \\ \vdots \\ n^N(t) \end{bmatrix} \tag{2.11}$$

2.1.5 Alignment (calibration)

Consider the model (2.8) with measurements $y^i(t)$ can representing either the range of a number of sparse reflectors or the position on the image plane of a sparse collection of point features. In the former case, the range is measured in the reference frame of the radar, and therefore we have

$$y^i(t) = \pi\left(g_{rb}g^{-1}(t)X_s^i\right) + n^i(t) \in \mathbb{R} \tag{2.12}$$

where $\pi(X) = \|X\|$ and g_{rb} is the transformation from the body frame to the radar. In the latter we have

$$\boxed{y^i(t) = \pi\left(g_{cb}g^{-1}(t)X_s^i\right) + n^i(t) \in \mathbb{R}^2} \tag{2.13}$$

where $\pi(X) = [X_1/X_3,\ X_2/X_3]^T$, and g_{cb} is the transformation from the body frame to the camera. The *"alignment"* transformations g_{cb}, g_{rb} are typically not known and should be inferred. We can then, as done for the points X^i, add them to the state with trivial dynamics $\dot{g}_{cb} = \dot{g}_{rb} = 0$.

2.1.6 Groups (occlusions)

It may convenient in some cases to represent the points X_s^i in the reference frame where they first appear, say at time t_i, rather than in the spatial frame. This is because the uncertainty is highly structured in the frame where they first appear. Consider $X^i(t_i) = \bar{y}^i(t_i)Z^i(t_i)$, then $y^i(t_i)$ has the same uncertainty of the feature

detector (small and isotropic on the image plane) and Z^i has a large uncertainty, but it is constrained to be positive.

However, to relate $X^i(t_i)$ to the state, we must bring it to the spatial frame, via $g(t_i)$, which is unknown. Although we may have a good approximation of it, the current estimate of the state $\hat{g}(t_i)$, the pose when the point first appears should be estimated along with the coordinates of the points. Therefore, we can represent X^i using $y^i(t_i)$, $Z^i(t_i)$ *and* $g(t_i)$:

$$X^i_s = X^i_s(g_{t_i}, y_{t_i}, Z_{t_i}) = g_{t_i}\bar{y}_{t_i}Z_{t_i} \tag{2.14}$$

Clearly this is an over-parametrization, since each point is now represented by $3+6$ parameters instead of 3. However, the pose g_{t_i} can be pooled among all points that appear at time t_i, considered therefore as a *group*. At each time, there may be a number $j = 1, \dots, K(t)$ groups, each of which has a number $i = 1, \dots, N_j(t)$ points. We indicate the group index with j and the point index with $i = i(j)$, omitting the dependency on j for simplicity. The representation of X^i_s then evolves according to

$$\begin{cases} \dot{y}^i_{t_i} = 0, \quad i = 1, \dots, N(j) \\ \dot{Z}^i_{t_i} = 0 \\ \dot{g}_j = 0, \quad j = 1, \dots, K(t). \end{cases} \tag{2.15}$$

For the case of range, this is not relevant as there is no reference frame that offers a preferential treatment of uncertainty.

2.1.7 Compact notation

If we call the "state" $x = \{T, R, V, \alpha_b, \omega_b, X\} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ the "known input" $u = \{\omega_{\text{imu}}, \alpha_{\text{imu}}\} = \{u_1, u_2\}$, the *unknown input* $v = \{n_{\omega_b}, n_{\alpha_b}\} = \{v_1, v_2\}$, we can write the mechanization equations (2.8) as

$$\dot{x} = f(x) + c(x)u + Dv \tag{2.16}$$

where

$$f(x) \doteq \begin{bmatrix} x_3 \\ -x_2 x_4 \\ -x_2 x_5 + \gamma \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad c(x) \doteq \begin{bmatrix} 0 \\ R \\ R \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad D \doteq \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \tag{2.17}$$

and the measurement equation (2.11) as

$$y = h(x) + n \tag{2.18}$$

where

$$h(x) \doteq \begin{bmatrix} \vdots \\ \pi(x_2'(x_6^i - x_1)) \\ \vdots \end{bmatrix} \tag{2.19}$$

Putting together (2.8)-(2.11) we have a model of the form

$$\boxed{\begin{cases} \dot{x} = f(x) + c(x)u + Dv \\ y = h(x) + n. \end{cases}} \tag{2.20}$$

2.1.8 Definitions

We call $y^t = \{y(\tau)\}_{\tau=0}^t$, a collection of output measurements, and $x^t = \{x(\tau)\}_{\tau=0}^t$ a state *trajectory*. Given output measurements y^t and known inputs u^t, we call

$$\mathcal{I}(y^t|u^t; \tilde{x}_0) \doteq \{\tilde{x}^t \mid y^t = h(\tilde{x}^t) \text{ s. } t. \ \dot{\tilde{x}}(t) = f(\tilde{x}) + c(\tilde{x})u(t), \ \tilde{x}(0) = \tilde{x}_0 \ \forall \ t\} \tag{2.21}$$

the *indistinguishable set*, or set of *indistinguishable trajectories*, for a given input u^t. If the initial condition $\tilde{x}_0 = x_0$ equals the "true" one, the indistinguishable set contains at least one element, the "true" trajectory x^t. However, if $\tilde{x}_0 \neq x_0$, the true trajectory may not even be part of this set.

If the indistinguishable set is a singleton (it contains only one element, $\tilde{x}^t$, which is a function of the initial condition $\tilde{x}_0$), we say that the model is *observable up to the initial condition*, or simply *observable*.[5] If $\{\tilde{x}^t\}$ is further independent of the initial condition, we say that the model is *strongly observable*: $\mathcal{I}(y^t|u^t;\tilde{x}_0) = \{x^t\}\ \forall\ \tilde{x}_0,\ u^t$.

If the state includes unknown parameters with a trivial dynamic, and there is no unknown input, $v = 0$, then observability of the resulting model implies that the parameters are *identifiable.* That usually requires the input u^t to be *sufficiently exciting* (SE), in order to enable disambiguating the indistinguishable states, [6] as the definition does not require that every input disambiguates states.

In the presence of *unknown inputs* $v \neq 0$, consider the following definition

$$\mathcal{I}_v(y^t|u^t;\tilde{x}_0) \doteq \{\tilde{x}^t \mid \exists\, v^t \text{ s. } t.\ y^t = h(\tilde{x}^t),\ \dot{\tilde{x}}(t) = f(\tilde{x}) + c(\tilde{x})u(t) + Dv(t)\ \forall\, t;\ \tilde{x}(0) = \tilde{x}_0\} \tag{2.22}$$

which is the set of *unknown-input indistinguishable states*. The model $\{f, c, D\}$ is said to be *unknown-input observable* (up to initial conditions) if the unknown-input indistinguishable set is a singleton. If such a singleton is further independent of the initial conditions, the model is strongly observable. The two definitions coincide once the only admissible unknown input is $v^t = 0$ for all t.

It is possible for a model to be observable (the indistinguishable set is a singleton), but not unknown-input observable (the unknown-input indistinguishable set is dense). In that case, the notion of *sensitivity* arises naturally, as one would want to measure the "size" of the unknown-input indistinguishable set as a function of the "size" of the unknown input. For instance, it is possible that if the set of unknown inputs is small in some sense, the resulting set of indistinguishable states

[5] We will assume that the solution of the differential equation $\dot{x} = f(x) + c(x)u$ is unique and continuously dependent on the initial condition, so if we impose $\tilde{x}_0 = x_0$, then $\tilde{x}^t = x^t$.

[6] Sufficient excitation means that the input is *generic*, and does not lie on a thin set. That is, even if we could find a particular input u^t that yields indistinguishable states, there will be another input that is infinitesimally close to it that will disambiguate them.

is also small. If $v \in V$ and for any $\epsilon > 0$ there exists a $\delta > 0$ such that $\mathrm{v}ol(V) \leq \epsilon$ for some measure of volume implies $\mathrm{v}ol(\mathcal{I}_v(y^t|u^t; \tilde{x}_0)) < \delta$ for any $u^t, \tilde{x}_0$, then we say that the model is *bounded-unknown-input/bounded-output observable* (up to the initial condition). If the latter volume is independent of $\tilde{x}_0$ we say that model is strongly bounded-unknown-input/bounded-output observable.

The set of indistinguishable trajectories $\mathcal{I}$ is an equivalence class, and when the model is observable *up to the initial condition*, it is parametrized by $\tilde{x}_0$. Choosing the "true" initial condition $\tilde{x}_0 = x_0$ produces an indistinguishable set consisting of the sole "true" trajectory, otherwise it is a singleton other than the true trajectory. In some cases, the initial condition corresponds to an arbitrary choice of reference frame, and therefore the equivalence class of indistinguishable trajectory are related by a *gauge transformation* (a change of coordinates). As the equivalence class can be represented by any element, enforcing a particular reference for the gauge transformation yields strong observability (although the singleton may not correspond to the true trajectory).

Related work

Unknown-input observability of linear time-invariant systems has ben addressed in [9, 29], for affine systems [30], and non-linear systems in [20, 54, 10]. The literature on robust filtering and robust identification is relevant, if the unknown input is treated as a disturbance. However, the form of the models involved in aided navigation do not fit in the classes treated in the literature above, which motivates our analysis.

2.2 Analysis of Bearing-Augmented Navigation

2.2.1 Preliminary claims

Lemma 2.2.1. Given $S \in \mathrm{SO}(3)$ and $\dot{S} \in T_{\mathrm{SO}(3)}(S)$, and $a \in \mathbb{R}$, the matrix $(aS + \dot{S})$ is nonsingular unless $a = 0$, in which case it has rank 2 or 0.

Proof. The tangent $\dot{S}$ has the form SM, where M is some skew-symmetric matrix. As such, $Mx \perp x$ for any $x \in \mathbb{R}^3$, so

$$\|(aS + \dot{S})x\|_2^2 = \|S(aI + M)x\|_2^2 = \|ax\|_2^2 + \|Mx\|_2^2.$$

The above is zero only if $ax = 0$, so $(aS + \dot{S})$ is nonsingular. For the remaining cases, observe that a 3×3 skew-symmetric matrix has rank 2 or 0. □

Lemma 2.2.2. Let $(R(t), T(t))$ and $(\tilde{R}(t), \tilde{T}(t))$ be differentiable trajectories in SE(3). For each time $t' \in [0, T]$, there exists an open, full-measure subset $\mathcal{A}_{t'} \subset \mathbb{R}^3$ such that:

> For any two static point-clouds $\{X^i\}_{i=1}^N \subset \mathcal{A}_{t'}$ and $\{\tilde{X}^i\}_{i=1}^N \subset \mathbb{R}^3$ that satisfy
>
> $$\pi\big(R^{-1}(t)(X^i - T(t)\big) = \pi\big(\tilde{R}^{-1}(t)(\tilde{X}^i - \tilde{T}(t))\big) \quad \text{for all } i \text{ and } t \tag{2.23}$$
>
> there exist constant scalings $\sigma_{it'} > 0$ and a constant rotation $S_{t'} = \tilde{R}(t')R^{-1}(t')$ such that
>
> $$\sigma_{it'} S_{t'}(X^i - T(t)) = (\tilde{X}^i - \tilde{T}(t)) + O((t - t')^2) \quad \text{for all } i \text{ and } t.$$
>
> Furthermore, if $T(t') \neq 0$, then $\sigma_{it'} = \sigma_{t'}$ for all i.

Proof. Write $S(t) = \tilde{R}(t)R^{-1}(t)$. Equality under the projection π implies that there exists a scaling $\sigma_i(t)$ (possibly varying with X^i and t) such that

$$\sigma_i S\big(X^i - T) = \tilde{X}^i - \tilde{T}. \tag{2.24}$$

For a given time t', we wish to find a suitably large set $\mathcal{A}_{t'}$ such that $\dot{\sigma}_i(t') = \dot{S}(t') = 0$ and $\sigma_i(t')$ is independent of X^i, when $X^i \in \mathcal{A}_{t'}$ Taking time derivatives,

$$\big(\dot{\sigma}_i S + \sigma_i \dot{S}\big)(X^i - T) - \sigma_i S\dot{T} = -\dot{\tilde{T}}$$

or, dividing by σ_i,

$$\big(\tfrac{\dot{\sigma}_i}{\sigma_i} S + \dot{S}\big)(X^i - T) - S\dot{T} = -\tfrac{1}{\sigma_i}\dot{\tilde{T}}. \tag{2.25}$$

Differentiating both sides with respect to X^i,

$$\big(\tfrac{\dot{\sigma}_i}{\sigma_i} S + \dot{S}\big)\delta X^i + \big(\tfrac{d}{dX^i}\big(\tfrac{\dot{\sigma}_i}{\sigma_i}\big)\delta X^i\big) S(X^i - T) = -\big(\tfrac{d}{dX^i}\big(\tfrac{1}{\sigma_i}\big)\delta X^i\big)\dot{\tilde{T}}. \tag{2.26}$$

Observe that $\frac{d}{dX^i}\big(\frac{\dot{\sigma}_i}{\sigma_i}\big)\delta X^i$ and $\frac{d}{dX^i}\big(\frac{1}{\sigma_i}\big)\delta X^i$ are scalars. By Lemma 2.2.1, the LHS has rank 2 or greater (as a linear map on δX^i), unless $\dot{\sigma}_i(t') = 0$. The RHS, however, has rank at most 1. Thus, (2.25) is invalid for almost all X^i, unless $\dot{\sigma}_i(t') = 0$ (two maps of different ranks can only agree on a submanifold). Plugging $\dot{\sigma}_i = 0$ into (2.26), we are left with

$$\dot{S}\delta X^i = -\big(\tfrac{d}{dX^i}\big(\tfrac{1}{\sigma_i}\big)\delta X^i\big)\dot{\tilde{T}}. \tag{2.27}$$

Now, the LHS has rank 2 or 0, while the RHS has rank 1 or 0. Again, (2.25) is invalid for almost all X^i, unless $\dot{S}(t') = 0$. Let $\mathcal{A}_{t'} \subset \mathbb{R}^3$ be the open, full-measure subset (being the complement of two submanifolds) on which the latter must hold. If, in addition, $T\dot{(}t') \neq 0$, then $\dot{\tilde{T}}(t') \neq 0$ and $\frac{d\sigma_i}{dX^i}(t') = 0$, we can finally write

$$\sigma_{t'} S_{t'}(X^i - T) = \tilde{X}^i - \tilde{T} + O((t - t')^2).$$

□

Claim 1 (Indistinguishable Trajectories from Bearing Data Sequences)**.** Let $g(t)$ and $\tilde{g}(t)$ be differentiable trajectories in SO(3). There exists an open, full-measure subset $\mathcal{A} \subset \mathbb{R}^3$ such that

Given two static, generic (non-coplanar) point clouds $\{X^i\}_{i=1}^N \subset \mathcal{A}$ and $\{\tilde{X}^i\}_{i=1}^N \subset \mathbb{R}^3$, satisfying

$$\pi(g^{-1}(t)\, X^i) = \pi(\tilde{g}^{-1}(t)\, \tilde{X}^i) \quad \text{for all } i \text{ and } t,$$

there exist constant scalings $\sigma_i > 0$ and a constant transformation $\bar{g} \in \mathrm{SE}(3)$ such that

$$\begin{cases} \tilde{X}^i = \sigma_i(\bar{g}X^i) \\ \tilde{g}(t) = \sigma_i(\bar{g}g(t)) \end{cases} \quad \text{for all } i \text{ and } t. \tag{2.28}$$

Furthermore, if $g(t)$ has a non-constant translational component, then $\sigma_i = \sigma$ for all i.

Proof. Write $g(t) = (R(t), T(t))$ and $\tilde{g}(t) = (\tilde{R}(t), \tilde{T}(t))$. Let $\mathcal{A} = \{X \in \mathbb{R}^3 : X \in \mathcal{A}_{t'} \text{ for almost all } t'\}$, with $\mathcal{A}_{t'}$ defined as in Lemma 2.2.2. By Fubini's theorem, this has full measure in $\mathbb{R}^3$. If $\{X^i\} \subset \mathcal{A}$, then the conditions for Lemma 2.2.2 are satisfied for almost all t, and thus there exist *constant* (being stationary for almost all t) scalings σ_i and rotation $S = \tilde{R}(t)R(t)^{-1} \in \mathrm{SO}(3)$ such that $\tilde{X}^i = \sigma_i S(X^i - T_t) + \tilde{T}_t$.

Define $\bar{g}(t) = (\sigma_i^{-1}\tilde{g}(t))\, g(t)^{-1}$, and observe that

$$\tilde{X}^i = \sigma_i S(X^i - T_t) + \tilde{T}_t = \sigma_i(\tilde{R}_t(g^{-1}X^i) + \sigma_i^{-1}\tilde{T}_t) = \sigma_i\big((\sigma_i^{-1}\tilde{g}(t))\, g(t)^{-1}X^i\big) = \sigma_i(\bar{g}(t)X^i).$$

If this affine relation holds for the generic set $\{X^i\}$, then $\bar{g}(t)$ must be constant. Next,

$$\sigma_i(\bar{g}g(t)) = \sigma_i((\sigma_i^{-1}\tilde{g}(t))\, g(t)^{-1}g(t)) = \sigma_i(\sigma_i^{-1}\tilde{g}(t)) = \tilde{g}(t).$$

Finally, if $T(t') = 0$ for some t', then $\sigma_i = \sigma_i(t') = \sigma(t') = \sigma$ for all i. □

In what follows, we will avoid the cumbersome discussion of sets such as $\mathcal{A} \subset \mathbb{R}^3$, defined by a given trajectory, and will instead speak of *sufficiently exciting* trajectories, for which a given point cloud is suitable for tracking.

Definition 1 (Sufficiently Exciting Motion)**.** A trajectory $g(t)$ is **sufficiently exciting** relative to a point-cloud $\{X^i\}_{i=1}^N \subset \mathbb{R}^3$ if, for all $\{\tilde{X}^i\}_{i=1}^N \subset \mathbb{R}^3$ and $\tilde{g}(t)$ in SE(3),

$$\pi(g(t)^{-1}(t)X^i) = \pi(\tilde{g}(t)^{-1}\tilde{X}^i) \quad \text{for all } i \text{ and } t \iff \tag{2.29}$$
$$\left(\begin{array}{c} \tilde{X}^i = \sigma(\bar{g}X^i) \\ \tilde{g}(t) = \sigma(\bar{g}g(t)) \end{array} \text{ for all } i \text{ and } t \right) \text{ for some constant } \sigma > 0 \text{ and } \bar{g} \in \mathrm{SE}(3).$$

That is, if the projection map $\pi(g(t)X^i)$ defines $g(t)$ and $\{X^i\}$ up to a constant rotation and mapping.

Observe that the right-to-left implication is always true: if the RHS holds, then

$$\pi(\tilde{g}(t)^{-1}\tilde{X}^i) = \pi((\sigma\bar{g}g(t))^{-1}\sigma(\bar{g}X^i))\pi(g(t)^{-1}\bar{g}^{-1}\sigma^{-1}\sigma\bar{g}X^i) = \pi(g(t)^{-1}X^i).$$

We will see that the sufficient excitation condition is very easily satisfied.

Claim 2. Given trajectories $g(t)$ and $\tilde{g}(t)$ in SE(3) with non-constant translation, and a set $\{X^i\}_{i=1}^N$ of $N \geq 4$ points sampled i.i.d. from a non-singular distribution over $\mathbb{R}^3$, the trajectory $g(t)$ is a.s. sufficiently exciting relative to $\{X^i\}$.

Proof. Fix $g(t)$. By Claim 1, there exists a full-measure $\mathcal{A} \subset \mathbb{R}^3$ such that (2.29) holds for any static, generic point clouds $\{X^i\}_{i=1}^N \subset \mathcal{A}$ and $\{\tilde{X}^i\}_{i=1}^N \subset \mathbb{R}^3$. If $\{X^i\}$ is sampled i.i.d. from a non-singular distribution over $\mathbb{R}^3$, then $\{X^i\} \subset \mathcal{A}$ almost surely. □

Equation (2.28) establishes the fact that the indistinguishable trajectories are an equivalence class parameterized by a group $\sigma(\bar{g})$, called a *gauge transformation*. We now include a constant reference frame g_a. We then have the following claim.

Claim 3 (Indistinguishable Alignments)**.** For a point cloud $\{X^i\}_{i=1}^{N(t)}$, $N(t) > 3$, in general position (non-coplanar), and sufficiently exciting motion,

$$\pi(g_a g^{-1}(t)X^i) = \pi(\tilde{g}_a\tilde{g}^{-1}(t)\tilde{X}^i) \tag{2.30}$$

if and only if there exist constants $\sigma > 0$, g_A and $g_B \in \mathrm{SE}(3)$ such that

$$\begin{cases} \tilde{X}^i = \sigma(g_B X^i) \\ \tilde{g}(t) = \sigma(g_B g(t) g_A) \\ \tilde{g}_a = \sigma(g_a g_A). \end{cases} \tag{2.31}$$

Proof. From Claim 1 we get constant $g_B \in \mathrm{SE}(3)$ and $\sigma > 0$ such that $\tilde{X}^i = \sigma(g_B X^i)$ and

$$\tilde{g}(t)\tilde{g}_a^{-1} = \sigma(g_B g(t) g_a^{-1}) \tag{2.32}$$

Let $g_A = g_a^{-1}\sigma^{-1}(\tilde{g}_a)$. Then $\tilde{g}_a = \sigma(g_a g_A)$ and

$$\tilde{g}(t) = \sigma(g_B g(t) g_A).$$

□

We now include groups of points, each with its own reference frame.

Claim 4 (Indistinguishable Groups)**.** For a number $i = 1, \ldots, K$ of groups each with a number $j = 1, \ldots, N_i \geq 3$ of points in general position (non-coplanar), and sufficiently exciting motion,

$$\pi(g_a g^{-1}(t) g_i g_a^{-1} X^j) = \pi(\tilde{g}_a \tilde{g}^{-1}(t) \tilde{g}_i \tilde{g}_a^{-1} \tilde{X}^j) \tag{2.33}$$

if and only if there exist constants $\sigma > 0, g_A, g_B, \bar{g}_i \in \mathrm{SE}(3)$ such that

$$\begin{cases} \tilde{X}^j = \sigma(g_a \bar{g}_i^{-1} g_i g_a^{-1} X^j) \\ \tilde{g}(t) = \sigma(g_B g(t) g_A) \\ \tilde{g}_i = \sigma(g_B \bar{g}_i g_A) \\ \tilde{g}_a = \sigma(g_a g_A) \end{cases} \tag{2.34}$$

Proof. From Claim 1, we get constant $g_C \in SE(3)$ and $\sigma > 0$ such that

$$\tilde{X}^i = \sigma(g_C X^i), \tag{2.35}$$

$$\tilde{g}_a \tilde{g}_i^{-1} \tilde{g}(t) \tilde{g}_a^{-1} = \sigma(g_C g_a g_i^{-1} g(t) g_a^{-1}). \tag{2.36}$$

Define

$$g_A := g_a^{-1}\sigma^{-1}(\tilde{g}_a), \qquad g_B := \sigma^{-1}(\tilde{g}_i g_a^{-1}) g_C g_a g_i^{-1}, \qquad \bar{g}_i := g_i g_a^{-1} g_C^{-1} g_a.$$

Then, applying the definition of $\bar{g}_i$ to (2.35),

$$\tilde{X}^j = \sigma(g_C X^j) = \sigma((g_a \bar{g}_i^{-1} g_i g_a^{-1}) X^j).$$

Applying the definitions of g_A and g_B to (2.36),

$$\tilde{g}(t) = \tilde{g}_i \tilde{g}_a^{-1} \sigma(g_C g_a g_i^{-1} g(t) g_a^{-1}) \tilde{g}_a = \sigma\big(\underbrace{\sigma^{-1}(g_i \tilde{g}_a^{-1}) g_C g_a g_i^{-1}}_{g_B}\, g(t)\, \underbrace{g_a^{-1}\sigma^{-1}(\tilde{g}_a)}_{g_A}\big) = \sigma(g_B g(t) g_A).$$

Rearranging the definitions of g_A, g_B and $\bar{g}_i$,

$$\tilde{g}_i = \sigma(g_B g_i g_a^{-1} g_C^{-1}) \tilde{g}_a = \sigma\left(g_B g_i g_a^{-1} g_C^{-1} \sigma(\tilde{g}_a)\right) = \sigma\big(g_B \underbrace{g_i g_a^{-1} g_C^{-1} g_a}_{\bar{g}_i} \underbrace{g_a^{-1}\sigma(\tilde{g}_a)}_{g_A}\big) = \sigma(g_B \bar{g}_i g_A).$$

Finally, rearrange the definition of g_A to get

$$\tilde{g}_a = \sigma(g_a g_A).$$

□

Eq. (2.34) describes the ambiguous state trajectories if only bearing measurement time series are given. In that case, there is no alignment to other sensor, so we can assume without loss of generality that $g_a = Id$ and so for $\tilde{g}_a$, which in turn implies $g_A = Id$. The resulting ambiguity is well-known [79] and shows that scale σ is constant but arbitrary, that the global reference frame is arbitrary (since g_B is), and that the reference frame of each group is also arbitrary (since $\bar{g}_i$ is). To lock these ambiguities, we can fix three directions for each group (thus fixing

$\bar{g}_i$) and, in addition, for one of the groups fix the pose (thus fixing g_B); finally, we can impose that the centroid of the points in that one group (the "reference group") be one, which fixes σ. Thus, an observer designed based on the standard model, where 3 directions within each group are saturated, and where the pose of one group is fixed, and the centroid of the group is at distance one, is observable, and under the usual assumptions it should converge to a state trajectory that is related to the true one by an arbitrary unknown scaling, and global reference frame.

Now, when inertial measurements are present, of all the possible trajectories that are indistinguishable from the measurements, we are interested *only* in those that are compatible with the dynamical model driven by IMU measurements. Since the fact that X^j and g_a are constant has already ben enforced, the model will impose no constraints on $\tilde{X}^j, \tilde{g}_i$ and $\tilde{g}_a$. However, it will offer constraints on $\tilde{g}(t)$, that depends on the arbitrary constants σ, g_A, g_B.

2.2.2 Indistinguishable trajectories in bearing augmentation

Definition 2. For an $\mathbb{R}^3$-valued trajectory $f : \mathbb{R} \to \mathbb{R}^3$ and interval $\mathcal{I} \subset \mathbb{R}^+$, define

$$
\begin{aligned}
m(f:\mathcal{I}) &:= \inf_{\|x\|=1} \Big(\sup_{t\in\mathcal{I}} |f(t)\cdot x| \Big) = \inf_{\|x\|=1} \Big(\sup_{t\in\mathcal{I}} \|f(t)\times x\| \Big), \\
M(f:\mathcal{I}) &:= \sup_{\|x\|=1} \Big(\sup_{t\in\mathcal{I}} |f(t)\cdot x| \Big) = \sup_{t\in\mathcal{I}} \|f(t)\|, \quad \text{and} \\
\bar{m}(f:\mathcal{I}) &:= \sqrt{\max\{0,\, 2m(f:\mathcal{I})^2 - M(f:\mathcal{I})^2\}}.
\end{aligned}
$$

Observe that $M(f:\mathcal{I}) \geq m(f:\mathcal{I}) \geq \bar{m}(f:\mathcal{I})$, and that the inequalities are strict unless $\{\pm f(t) | \, t \in \mathcal{I}\}$ is dense on the sphere of radius $M(f:\mathcal{I})$. We use these "minimum-excitation" bounds in order to prove a partial converse of the Cauchy-Schwarz inequality:

Lemma 2.2.3. Let $A = c_1 I + c_2 R$, for some rotation $R \in \mathrm{SO}(3)$ and scalars c_1

and c_2. Then, for any trajectory $f : \mathbb{R}^+ \to \mathbb{R}^3$ and set of times $\mathcal{I} \subset \mathbb{R}^+$,

$$\sup_{t\in\mathcal{I}} \|Af(t)\| \;\geq\; \|A\|\, \bar{m}(f{:}\mathcal{I}).$$

Proof. First, observe that A is normal:

$$AA^T = (c_1 I + c_2 R)(c_1 I + c_2 R^T) = 2c_1c_2 I + c_1c_2(R + R^T) = A^T A.$$

Let $\{(\lambda_i, v_i)\}_{i=1}^3$ be orthonormal eigenvalue/eigenvector pairs of A, with $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

$$\begin{aligned}
\|Af(t)\|^2 &= \lambda_1^2(v_1 \cdot f(t))^2 + \lambda_2^2(v_2 \cdot f(t)^2 + \lambda_3^2(v_3 \cdot f(t))^2 \\
&\geq \lambda_1^2\big((v_1 \cdot f(t))^2 - (v_2 \cdot f(t)^2 - (v_3 \cdot f(t))^2\big) \\
&= \|A\|^2\big(2(v_1 \cdot f(t))^2 - \|f(t)\|^2\big).
\end{aligned}$$

Taking the supremum over $\mathcal{I}$,

$$\begin{aligned}
\sup_{t\in\mathcal{I}} \|Af(t)\|^2 &\geq \|A\|^2 \sup_{t\in\mathcal{I}}\big(2(v_1 \cdot f(t))^2 - \|f(t)\|^2\big) \\
&\geq \|A\|^2\big(2\sup_{t\in\mathcal{I}}(v_1 \cdot f(t))^2 - \sup_{t\in\mathcal{I}}\|f(t)\|^2\big) \\
&\geq \|A\|^2\big(2m(f{:}\mathcal{I})^2 - M(f{:}\mathcal{I})^2\big)
\end{aligned}$$

□

Lemma 2.2.4. Let $A = I - R$, for some rotation $R \in \mathrm{SO}(3)$. Then, for trajectory $f : \mathbb{R}^+ \to \mathbb{R}^3$ and $\mathcal{I} \subset \mathbb{R}^+$,

$$\sup_{t\in\mathcal{I}} \|Af(t)\| \;\geq\; \|A\|\, m(f{:}\mathcal{I}).$$

Proof. Let $\{(\lambda, v_1), (\bar{\lambda}, v_2), (1, 0)\}$ be the orthonormal eigenvalue/eigenvector pairs of R. Since R and I commute, $\{(\lambda - 1, v_1),\ (\bar{\lambda} - 1, v_2), (0, u)\}$ are the eigenpairs of A, and $\|A\| = |\lambda - 1| = |\bar{\lambda} - 1|$. Then,

$$\|Af(t)\|^2 = |\lambda - 1|^2(v_1 \cdot f(t))^2 + |\bar{\lambda} - 1|^2(v_2 \cdot f(t))^2 + 0 = \|A\|^2(w \cdot f(t))^2,$$

where

$$w := \frac{(v_1 \cdot f(t))v_1 + (v_2 \cdot f(t))v_2}{\|(v_1 \cdot f(t))v_1 + (v_2 \cdot f(t))v_2\|} = \frac{(v_1 \cdot f(t))v_1 + (v_2 \cdot f(t))v_2}{\sqrt{(v_1 \cdot f(t))^2 + (v_2 \cdot f(t))^2}}.$$

Taking the supremum over $\mathcal{I}$,

$$\sup_{t\in\mathcal{I}} \|Af(t)\|^2 = \|A\|^2 \sup_{t\in\mathcal{I}} \|w \cdot f(t)\|^2 \geq \|A\|^2 m(f\colon\mathcal{I})^2.$$

□

Claim 5 (Indistinguishable Trajectories from IMU Data)**.** Let $g(t) = (R(t), T(t)) \in \mathrm{SE}(3)$ be such that

$$\begin{cases} \dot{R} = R(\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b) \\ \dot{T} = V \\ \dot{V} = R(\alpha_{\text{imu}} - \alpha_b) + \gamma \end{cases} \tag{2.37}$$

for some known constant γ and functions $\alpha_{\text{imu}}(t)$, $\omega_{\text{imu}}(t)$ and for some unknown functions $\alpha_b(t), \omega_b(t)$ that are constrained to have $\|\dot{\alpha}_b(t)\| \leq \epsilon$, $\|\dot{\omega}_b(t)\| \leq \epsilon$, and $\|\ddot{\omega}_b(t)\| \leq \epsilon$ at all t, for some $\epsilon < 1$.

Suppose $\tilde{g}(t) \doteq \sigma(g_B g(t) g_A)$ for some constant $g_A = (R_A, T_A)$, $g_B = (R_B, T_B)$, $\sigma > 0$, with bounds on the configuration space such that $\|T_A\| \leq M_A$ and $|\sigma| \leq M_\sigma$. Then, under sufficient excitation conditions (described in this proof), $\tilde{g}(t)$ satisfies (2.37) if and only if

$$\|I - R_A\| \leq \frac{2\epsilon}{m(\dot{\omega}_{\text{imu}}\colon\mathbb{R}^+)} \tag{2.38}$$

$$|\sigma - 1| \leq \frac{k_{c_1}\epsilon + M_\sigma \|I - R_A\|}{M(\dot{\alpha}_{\text{imu}}\colon\mathcal{I}_{c_1})} \tag{2.39}$$

$$\|T_A\| \leq \frac{\epsilon(k_{c_2} + (2M_\sigma + 1)M_A)}{(1 - |\sigma - 1|)\, m(\ddot{\omega}_{\text{imu}}\colon\mathcal{I}_{c_2})} \tag{2.40}$$

$$\|(1 - R_B^T)\gamma\| \leq \frac{\epsilon(k_{c_3} + M_\sigma M_A) + (|\sigma - 1| + \epsilon)M(\omega_{\text{imu}} - \omega_b\colon\mathcal{I}_{c_3})\|\gamma\|}{m(\omega_{\text{imu}} - \omega_b\colon\mathcal{I}_{c_3})\,(1 - |\sigma - 1|)} \tag{2.41}$$

for $\mathcal{I}_i$ and k_i determined by the sufficient excitation conditions.

Proof.

(2.38) The ambiguous rotation $\tilde{R}$ must satisfy $\dot{\tilde{R}} = \tilde{R}(\widehat{\omega}_{\text{imu}} - \widehat{\tilde{\omega}}_b)$ for some $\tilde{\omega}_b$:

$$\begin{aligned}\dot{\tilde{R}} &= R_B R(\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b)R_A = \tilde{R}R_A^T(\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b)R_A = \tilde{R}(\widehat{R_A^T\omega}_{\text{imu}} - \widehat{R_A^T\omega}_b) \\ &= \tilde{R}(\widehat{\omega}_{\text{imu}} - [\widehat{\omega}_{\text{imu}} + \widehat{R_A^T\omega}_{\text{imu}} - \widehat{R_A^T\omega}_b])\end{aligned}$$

where the quantity in brackets is $-\widehat{\tilde{\omega}}_b$, which defines

$$\tilde{\omega}_b := R_A^T\omega_b + (I - R_A^T)\,\omega_{\text{imu}}. \tag{2.42}$$

Taking derivatives and rearranging,

$$2\epsilon \geq \|\dot{\tilde{\omega}}_b - R_A^T\dot{\omega}_b\| = \|(I - R_A^T)\dot{\omega}_{\text{imu}}\|$$

Since this is true for all $t \in \mathbb{R}$, we can write

$$2\epsilon \geq \sup_{t\in\mathbb{R}} \|(I - R_A^T)\,\dot{\omega}_{\text{imu}}(t)\| \geq \|I - R_A^T\|\, m(\dot{\omega}_{\text{imu}} : \mathbb{R}^+).$$

This rearranges to give (2.38).

(2.39) The ambiguous translation $\tilde{T}$ must satisfy the dynamics in (2.37):

$$\ddot{\tilde{T}} = \dot{\tilde{V}} = \tilde{R}(\alpha_{\text{imu}} - \tilde{\alpha}_b) + \gamma = R_B R R_A(\alpha_{\text{imu}} - \tilde{\alpha}_b) + \gamma.$$

Alternatively, working with $\tilde{T} = \sigma R_B(RT_A + T)$ and applying the dynamics to T,

$$\ddot{\tilde{T}} = \sigma R_B(\ddot{R}T_A + \ddot{T}) = \sigma R_B(\ddot{R}T_A + R(\alpha_{\text{imu}} - \alpha_b) + \gamma).$$

Taking the difference between these two expressions,

$$0 = \sigma R_B\ddot{R}T_A + R_B R(R_A\tilde{\alpha}_b - \sigma\alpha_b) + R_B R(\sigma\alpha_{\text{imu}} - R_A\alpha_{\text{imu}}) + (\sigma R_B - I)\gamma,$$

and multiplying by $R^T R_B^T$,

$$\begin{aligned}0 &= \sigma(R^T\ddot{R})T_A + (R_A\tilde{\alpha}_b - \sigma\alpha_b) + (\sigma\alpha_{\text{imu}} - R_A\alpha_{\text{imu}}) + R^T(\sigma - R_B^T)\gamma \\ &= \sigma((\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b)^2 + (\dot{\widehat{\omega}}_{\text{imu}} - \dot{\widehat{\omega}}_b))T_A + (R_A\tilde{\alpha}_b - \sigma\alpha_b) + (\sigma\alpha_{\text{imu}} - R_A\alpha_{\text{imu}}) + R^T(\sigma - R_B^T)\gamma.\end{aligned}$$

Differentiating again,

$$0 = \sigma(\dot{R}^T\ddot{R} + R^T\dddot{R})T_A \tag{2.43}$$

$$+ ((I - R_A)\sigma + (\sigma - 1)R_A)\dot{\alpha}_{\text{imu}} \tag{2.44}$$

$$+ \dot{R}^T((I - R_B^T)\sigma + (\sigma - 1)R_B^T)\gamma. \tag{2.45}$$

$$+ (R_A\dot{\hat{\alpha}}_b - \sigma\dot{\alpha}_b) \tag{2.46}$$

As a sufficient excitation condition, assume that $\|\dot{R}(t)\| \leq \epsilon$, $\|\ddot{R}(t)\| \leq \epsilon$, and $\|\dddot{R}(t)\| \leq \epsilon$, for $t \in \mathcal{I}_{c_1}$. Under these constraints, (2.44) is bounded by $k_{c_1}\epsilon$, where, e.g. $k_{c_1} := 2M_\sigma M_A + (2M_\sigma + 1)(\|\gamma\| + 1)$. In that case,

$$\begin{aligned} k_{c_1}\epsilon &\geq \max_{t\in\mathcal{I}_{c_1}} \|((I - R_A)\sigma + (\sigma - 1)R_A)\dot{\alpha}_{\text{imu}}(t)\| \\ &\geq |\sigma - 1|M(\dot{\alpha}_{\text{imu}}{:}\mathcal{I}_{c_1}) - M_\sigma\|I - R_A\|. \end{aligned}$$

This rearranges to give (2.39).

(2.40) Now, assume that $\|\dot{R}(t)\| \leq \epsilon$, $\|\ddot{R}(t)\| \leq \epsilon$, and $\|\dddot{T}(t) - \gamma\| \leq \epsilon$, for $t \in \mathcal{I}_{c_2}$. Under these constraints, $\|\dot{\alpha}_{\text{imu}}\| \leq 2\epsilon$, and (2.43) is bounded by $k_{c_2}\epsilon$, where, e.g. $k_{c_2} := (2M_\sigma + 1)(\|\gamma\| + 3)$. In that case,

$$\begin{aligned} k_{c_2}\epsilon &\geq \max_{t\in\mathcal{I}_{c_2}} \|\sigma((\widehat{\omega}_{\text{imu}} - \widehat{\omega}_b)(\dot{\widehat{\omega}}_{\text{imu}} - \dot{\widehat{\omega}}_b) + (\ddot{\widehat{\omega}}_{\text{imu}} - \ddot{\widehat{\omega}}_b))T_A\| \\ &= \max_{t\in\mathcal{I}_{c_2}} \|\sigma((R^T\dot{R})(R^T\ddot{R} - (R^T\dot{R})^2) + (\dddot{\widehat{\omega}}_{\text{imu}} - \dddot{\widehat{\omega}}_b))T_A\| \\ &\geq (1 - |1 - \sigma|)\max_{t\in\mathcal{I}_{c_2}} \|\ddot{\omega}_{\text{imu}}(t) \times T_A\| - (2M_\sigma + 1)M_A\epsilon \\ &\geq (1 - |1 - \sigma|)\,\|T_A\|\,m(\ddot{\omega}_{\text{imu}}{:}\mathcal{I}_{c_2}) - (2M_\sigma + 1)M_A\epsilon. \end{aligned}$$

This rearranges to give (2.40).

(2.41) Finally, assume that $\|\ddot{R}(t)\| \leq \epsilon$, $\|\dddot{R}(t)\| \leq \epsilon$, and $\|\ddot{T}(t) - \gamma\| \leq \epsilon$ for $t \in \mathcal{I}_{c_3}$. As before, $\|\dot{\alpha}_{\text{imu}}\| \leq 2\epsilon$. Then, (2.43) + (2.44) is bounded by $k_{c_3}\epsilon$,

where, e.g. $k_{c_3} = 2M_\sigma + 3$. In that case,

$$\begin{aligned}
k_{c_3}\epsilon &\geq \|\sigma(\dot{R}^T\ddot{R} + R^T\dddot{R})T_A + \dot{R}^T((I - R_B^T)\sigma + (\sigma - 1)R_B^T)\gamma\| \\
&\geq \|\sigma\dot{R}^T(\ddot{R} + (I - R_B^T))\gamma\| - M_\sigma M_A\epsilon - |\sigma - 1|\,\|\dot{R}^T\|\,\|\gamma\| \\
&\geq (1 - |\sigma - 1|)\,\|\dot{R}^T(I - R_B^T)\gamma\| - M_\sigma M_A\epsilon - (|\sigma - 1| + \epsilon)\,\|\dot{R}^T\|\,\|\gamma\| \\
&\geq (1 - |\sigma - 1|)\,m(\dot{R}^T{:}\mathcal{I}_{c_3})\|(1 - R_B^T)\gamma\| - \epsilon(k_{c_3} + M_\sigma M_A) - (|\sigma - 1| + \epsilon)M(\dot{R}^T{:}\mathcal{I}_{c_3})\|\gamma\|
\end{aligned}$$

This rearranges to give (2.41).

□

2.2.3 Gauge transformations

The set of indistinguishable trajectories $\mathcal{I}$ is an equivalence class, and when the model is observable *up to the initial condition*, it is parametrized by $\tilde{x}_0$. Choosing the "true" initial condition $\tilde{x}_0 = x_0$ produces an indistinguishable set consisting of the sole "true" trajectory, otherwise it is a singleton other than the true trajectory. In some cases, the initial condition corresponds to an arbitrary choice of reference frame, and therefore the equivalence class of indistinguishable trajectory are related by a *gauge transformation* (a change of coordinates). As the equivalence class can be represented by any element, enforcing a particular reference for the gauge transformation yields strong observability (although the singleton may not correspond to the true trajectory).

Formally, an arbitrary choice of initial condition is sufficient to fix the gauge reference. For instance, the set of indistinguishable trajectories in the limit where

$\epsilon \to 0$ is parametrized by an arbitrary $T_B \in \mathbb{R}^3$ and $\theta \in \mathbb{R}$,

$$\begin{cases} \tilde{T} = \exp(\widehat{\gamma}\theta)T + T_B \\ \tilde{R} = \exp(\widehat{\gamma}\theta)R \\ \tilde{T}_{t_i} = \exp(\widehat{\gamma}\theta)\bar{T}_{t_i} + T_B \\ \tilde{R}_{t_i} = \exp(\widehat{\gamma}\theta)\bar{R}_{t_i} \qquad\qquad \text{up to } \mathcal{O}\left(\frac{\|\dot{\omega}_b\|}{\|\dot{\omega}_{\text{imu}}\|}, \frac{\|\dot{\alpha}_b\|}{\|\dot{\alpha}_{\text{imu}}\|}, \frac{1}{\|\gamma\|}\right) \\ \tilde{T}_{cb} = T_{cb} \\ \tilde{R}_{cb} = R_{cb} \end{cases} \tag{2.47}$$

If we impose that $T(0) = \tilde{T}(0) = 0$, then $T_B = 0$ is determined; similarly, if we impose the initial pose to be aligned with gravity (so gravity is in the form $[0\ 0\ \|\gamma\|]^T$, then $\theta = 0$. But while we can impose this condition, we cannot *enforce* it, since the initial condition is not a part of the state of the filter, so we cannot relate the measurements at each time t directly to it.

However, if the gauge reference can be associated to *constant parameters* that are part of the state of the model, the gauge ambiguity can be enforced in a consistent manner. For instance, the ambiguous set of points is

$$\tilde{X}^j = g_a \bar{g}_i^{-1} g_i g_a^{-1} X^j. \tag{2.48}$$

If each group i contains at least 3 non-coplanar points, it is possible to fix $\bar{g}_i$ by parametrizing $X^j \doteq \bar{y}^j_{t_i} Z^j$ and imposing thre directions $y^j_{t_i} = \tilde{y}^j_{t_i} = y^j(t_i), j = 1, \ldots, 3$, the measurement of these directions at time t_i when they first appear. This yields $\bar{g}_i = g_i$ and $\tilde{X}^j = X^j$ for that group. Note that it is necessary to impose this constraint in *each group*.

The residual set of indistinguishable trajectories is parameterized by *constants* θ, T_B, that determine a Gauge transformation for the groups, that can be fixed by always fixing the pose of *one* of the groups. This can be done in a number of ways. For instance, if for a certain group i we impose

$$R_{t_i} = \tilde{R}_{t_i} = \hat{R}(t_i) \text{ and } T_{t_i} = \tilde{T}_{t_i} = \hat{T}(t_i) \tag{2.49}$$

by assigning their value to the current best estimate of pose and not including the corresponding variables in the state of the model, then we have that

$$\hat{R}(t_i) = \exp(\widehat{\gamma}\theta)\hat{R}(t_i) \tag{2.50}$$

and therefore $\theta = 0$; similarly,

$$T_B = (I - \exp(\widehat{\gamma}\theta))T(t_i) = 0 \tag{2.51}$$

Therefore, the gauge transformation is enforced explicitly at each instant of time, as each measurement provides a constraint on the states. This suggests the following modeling procedure in the design of a filter/observer for bearing-assisted navigation:

1. Set $T(0) = 0$ with zero model error covariance, and zero initial covariance.
2. Set $R(0) = R_0$ such that $[I_{2\times 2}0]R_0\alpha_{\text{imu}} = 0$, with zero model error and non-zero initial covariance.
3. Fix gravity to $[0,\ 0,\ \|\gamma\|]^T$ from tabulates
4. Initialize at rest, then perform some fast motions before groups of features are added.
5. Add K groups, each with $2N + N$ states, plus their pose for each group but one.
6. Fix 2 directions per group ([34] fixes all directions; this results in a non-zero mean component of the innovation, that in turn results in a small bias in all other states, that have to account for/absorb the mean)
7. Fix the pose of one group (remove its pose from the state)
8. Triage groups before adding them to the state.

After the Gauge Transformation has ben fixed, the model is observable, and therefore a properly designed observer will converge to a solution $\tilde{x}$ that is related to the true one x as follows:

$$\tilde{X}^{\mathrm{ref}} = (1+\tilde{\sigma})\tilde{R}_{cb}e^{\omega_B}e^{\widehat{\gamma\theta}}e^{\omega_A}\tilde{R}_{cb}^T(X^{\mathrm{ref}} - T_A) + (1+\tilde{\sigma})(\tilde{R}_{cb}e^{\omega_A}T_B + \tilde{R}_{cb}T_A + \tilde{T}_{cb}) \quad (2.52)$$

$$\tilde{X}^j = (1+\tilde{\sigma})\tilde{R}_{cb}\bar{R}_i\tilde{R}_{t_i}\tilde{R}_{cb}^T(X^j - T_A) + (1+\tilde{\sigma})(\tilde{R}_{cb}\bar{R}_i\tilde{T}_{t_i} + \tilde{R}_{cb}\bar{T}_i + \tilde{T}_{cb}) \quad (2.53)$$

$$\tilde{T} = e^{\widehat{\gamma\theta}}T + T_B(1+\tilde{\sigma}) + \omega_B e^{\widehat{\gamma\theta}}T + e^{\omega_B}e^{\widehat{\gamma\theta}}RT_A(1+\tilde{\sigma}) \quad (2.54)$$

$$\tilde{R} = e^{\omega_B}e^{\widehat{\gamma\theta}}Re^{\omega_A} \quad (2.55)$$

$$\tilde{T}_{t_i} = e^{\widehat{\gamma\theta}}\bar{T}_i + T_B(1+\tilde{\sigma}) + \omega_B e^{\widehat{\gamma\theta}}\bar{T}_i + e^{\omega_B}e^{\widehat{\gamma\theta}}\bar{R}_iT_A(1+\tilde{\sigma}) \quad (2.56)$$

$$\tilde{R}_{t_i} = e^{\omega_B}e^{\widehat{\gamma\theta}}\bar{R}_ie^{\omega_A} \quad (2.57)$$

$$\tilde{T}_{cb} = T_{cb} + \tilde{\sigma}T_{cb} + R_{cb}T_A(1+\tilde{\sigma}) \quad (2.58)$$

$$\tilde{R}_{cb} = R_{cb}\exp(\omega_A) \quad (2.59)$$

w*here*

$$\|T_A\| \leq \frac{2k\min_t\|\dot{\omega}_b\|}{\max_t\|\ddot{\omega}_{\mathrm{imu}}\|}$$

$$\|\omega_A\| \leq \frac{2\min_t\|\dot{\omega}_b\|}{\max_t\|\dot{\omega}_{\mathrm{imu}}\|}$$

$$\|\omega_B\| \leq \left(\frac{3k\max(\min_t\|\dot{\omega}_b\|, \min_t\|\dot{\alpha}_b\|)}{\min(\max_t\|\dot{\omega}_{\mathrm{imu}}\|, \max_t\|\dot{\alpha}_{\mathrm{imu}}\|, \|\gamma\|)}\right)$$

$$|\tilde{\sigma}| \leq \left(\frac{2k\min_t\|\dot{\alpha}_b\|}{\min(\max_t\|\dot{\omega}_{\mathrm{imu}}\|, \max_t\|\dot{\alpha}_{\mathrm{imu}}\|)}\right)$$

and arbitrary θ, T_B and suitable constant κ. The groups will be defined up to an arbitrary reference frame $(\bar{R}_i, \bar{T}_i)$, except for the reference group where that transformation is fixed. Note that, as the reference group "switches" (when points in the reference group become occluded or otherwise disappear due to failure in the data association mechanism), a small error in pose is accumulated. This error affects the gauge transformation, not the *state* of the system, and therefore is not reflected in the innovation, nor in the covariance of the state estimate, that remains bounded. This is unlike [71], where the covariance of the translation state T_B and the rotation about gravity θ grows unbounded over time, possibly

affecting the numerical aspects of the implementation. Notice that in the limit where $\dot{\omega}_b = \dot{\alpha}_b = 0$, we obtain back Eq. (2.47).

CHAPTER 3

Scene Segmentation by Aggregation of Global Ordering Constraints

3.1 Introduction

In this chapter, we present a method, illustrated in Fig. 3.1, to endow a scene, densely reconstructed from monocular video, with a metric that incorporates geometric homogeneity and image topology through occlusions. While the latter are temporally inconsistent (they change with the video), the way they change is spatially consistent, an observation key to defining affinities that allow us to partition the scene into coherent "objects" at a level of granularity relevant to the viewer. Occlusions inform the scale of the segmentation, allowing the selection of a partitioning of the scene, out of all possible partitions, that respects the occlusions present in the video. For robotic interaction tasks, such as manipulation or obstacle avoidance, the granularity of the scene representation can be critical, and our approach focuses the scene segmentation task on objects that generate

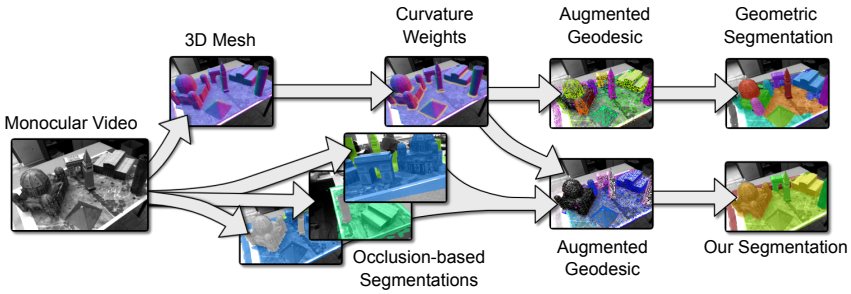


Figure 3.1: Our monocular dense reconstruction and segmentation pipeline

occlusions in the images due the motion of the viewer relative to the scene.

To achieve this, we employ a robust metric on the scene using a combination of curvature-based geodesics on a 3D mesh and back-projected occlusion-constrained image segmentations. The spatial consistency of these segmentations on the scene allows our segmentation method to adapt to the scale informed by the occlusions in the video. While one could employ trained object detectors at the outset to arrive at a semantic segmentation of the scene, we focus on low-level geometric and topological cues first, to segment the scene and images into coherent regions, where one could then deploy object detectors if so desired. Semantic analysis of the scene involves object identities and relations, and knowledge of scene geometry, topology, and putative object regions are key to infer the latter. This is the focus of our work.

The remainder of the chapter is organized as follows: In Sec. 3.2.1 we formulate scene (and video) segmentation as a selection problem on the set of potential nested partitions of the underlying scene based on homogeneity properties, and Sec. 3.2.3.1 presents occlusion-based image segmentation as a scale-selection mechanism on this set of potential partitions. Sec. 3.2.2.2 presents the construction of a curvature-augmented geometry on a 3D mesh used in Sec. 3.2.3.2 to regularize these occlusion-based selections obtained from the image frames through an adaptive geodesic that combines these two components to perform a scene segmentation at the level of granularity relevant to the viewer. Finally, Sec. 3.3 shows a quantitative and qualitative evaluation of our scale-adaptive segmentation scheme on a ground-truthed dataset of monocular dense reconstructions that we have collected.

The work described here is the result of a fifty-fifty collaboration with Konstantine Tsotsos, who designed and assembled the 3D pipeline, made significant contributions to the "object distance" metric, and implemented the segmentation scheme.

3.1.1 Contributions and Related Work

There is a vast body of literature pertaining to semantic segmentation of *images* [22, 85, 57, 21, 99, 83, 72, 92, 51, 43, 62]; our work is particularly related to *joint* segmentation (a.k.a. "co-segmentation") of multiple images, or *video* [50, 100, 41, 74, 52, 14]. However, the goal of such approaches is a partitioning of the spatio-temporal image volume, not of the *scene* that generated it. Seeking a segmentation of the scene allows us to bypass the complex and discontinuous changes in the partitioning of the video due to scale, spatial quantization, and occlusions. Furthermore, occlusions provide local ordering constraints that can be used to partition the image into "layers" [86, 68, 102, 75, 44, 45] by solving a convex optimization problem [41]. In all of these cases, "objects" are collections of pixels that are often *temporally inconsistent* as the local ordering constraints can change over time (think of a merry go round), producing flickering segmentations. However, by by using occlusions as topological cues these segmentations tend to correspond to spatially consistent regions on the scene, even if their image labels are not temporally consistent. Our method relies heavily on this observation to accumulate data-driven cues for the extent of individual objects in the scene.

Our work can be interpreted as an attempt to combine multiple image segmentations (which depend on both the scene and the viewer) into one persistent segmentation of the scene independent of the viewer. Since our method involves the intermediate reconstruction of a dense three-dimensional model of the scene, which we do in real time, our work also relates to multiple-view stereo and structure-from-motion, and in particular real-time dense multi-view stereo [60, 87, 88]. As an alternative, one could use an alternate range sensor [77], for instance based on structured light [61], although those perform poorly in natural illumination and have a fixed scale of interaction.

There are relatively few attempts to generate a dense label field in the *scene*

[36]. While semantic labels have been attached to various forms of 3D reconstruction, these typically are *sparse* (*e.g.,* collections of feature descriptors and their coarse positions [53, 33]).

Our work is also related to [39], in which Regression and Decision Tree Fields are used to segment a 3D scene, and [69], in which SVMs are used to segment point clouds gathered from RGB-D data. Similarly, Bleyer et al. [11] describe a method for labeling that is explicitly compatible with the 3D structure of the scene. Although direct comparison with these algorithms is not possible as neither their code nor their datasets are publicly available, in Sect. 3.3.2 we report experiments on data similar in nature and scale that we intend to release publicly upon completion of the anonymous review process. Other related work includes [101, 80, 6], where the focus is on manipulation. Additionally, Zheng et al. [103] use 3D point clouds to find objects in the scene using geometric and physical cues from RGB-D data. Most closely related to our work are are the recent works of [28, 73, 76, 46]), all of which generate semantic segmentations of the scene using responses from trained detectors as input. We seek to generate segmentations at a similar level of abstraction (albeit without semantic labels) through viewpoint-based topological homogeneity instead of semantic homogeneity (object detector responses). Our method can also be thought of as a generic proposal scheme for regions within which to collect support for semantic categorization based on geometric and viewpoint based contextual information. Partitions based on solely geometric homogeneity have also been explored extensively by the 3D mesh segmentation community ([15, 7, 49]).

Our contributions are (a) a method for scene segmentation leveraging the spatial consistency of temporally inconsistent image cues, (b) an adaptive geodesic distance function on the scene shaped by spatially consistent image cues in the form of occlusions, (c) an object-level scene segmentation scheme that extends a real-time dense reconstruction system based on monocular video. To compare

with standard approaches for segmenting dense geometry, we have (d) collected a calibrated dataset with a variety of objects of different scales and textures, indoor and outdoor, on natural and artificial laboratory scenes. A key assumption for our dense monocular reconstruction pipeline is that the only thing moving in the scene is the viewer. Extension beyond cases where this assumption holds is desirable, but even the static case is relevant to several applications from robotic inspection to autonomous navigation and exploration.

3.2 Methodology

3.2.1 Scene Model

The input to our system is a grayscale video $\{I_t\}_{t=0}^{T}$, with each image I_t mapping from a domain $D \subset \mathbb{R}^2$ to $\mathbb{R}_+$. The desired output is a constant partitioning of a higher-dimensional "scene" that the video observes, from which we can also derive a piecewise-constant, integer-valued function $c_t(x)$ that associates to each pixel $x \in D$ a label. The scene is represented by a (multiply-connected) collection of surfaces $S \subset \mathbb{R}^3$ supporting a reflectance function (albedo) $\rho : S \to \mathbb{R}_+$. Under the Lambert-Ambient model, the image and the scene are related by

$$\begin{cases} I_t(x) = \rho(p) + n_t(x), \ p \in S \\ x = \pi(g_t p) + v_t(x) \end{cases} \tag{3.1}$$

where $g_t \equiv (R(t), T(t)) \in SE(3)$ is the pose of the camera relative to the reference frame of S, and $\pi : \mathbb{R}^3 \to D$ is a canonical central (perspective) projection. The residual $n_t(x)$ accounts for unmodeled photometric phenomena such as changes in illumination (assumed negligible in the short time-span during which the video is captured), deviations from Lambertian reflection, sensor noise etc. The residual $v_t(x)$ accounts for violations of the geometric assumptions (rigid motion, static scene). Estimates $\hat{g}_t, \hat{S}$ are obtained as described in Sec. 3.2.2.1.

We consider that objects in the scene compose a nested covering of sets $\mathfrak{S} = \{S_i\}_{i=1}^{K}$, where $S_i \cap S_j \in \{\emptyset, S_i, S_j\}$ for all i and j, and $\bigcup_{i=1}^{K} S_i = S$. Any segmentation of the scene is a selection $\mathcal{P} = \{S_{\mathcal{P},i}\}_{i=1}^{K_\mathcal{P}}$ of disjoint sets in $\mathfrak{S}$ such that $\bigcup_{i=1}^{K_p} S_{\mathcal{P},i} = S$. For a partitioning $\mathcal{P}$ to be meaningful, typically some homogeneity property must hold on each $S_{\mathcal{P},i}$, be that geometric, photometric, semantic, topological, or some combination.

One can perform a sequence of still-frame (or short-baseline video) segmentations $c_t : D \to \mathbb{Z}^+$ using a subset of these properties which can then be leveraged into a segmentation of the scene. A reasonable such segmentation is one that does not oversegment the scene relative to c_t (distinguish points that have the same label c_t for all or almost all t), or undersegment (fail to distinguish points that tend to have different labels), more than necessary. As these c_t will be temporally inconsistent, they can be regularized by integration on the scene using geometric homogeneity. We consider a particular set of segmentations c_t to induce a selection $\mathcal{P}$ from $\mathfrak{S}$. The use of occlusion-based image segmentation (Sec. 3.2.3.1) to induce a segmentation respecting topological homogeneity as seen by the viewer is a key contribution of our approach.

3.2.2 Curvature Augmented Geometry and Geometric Affinity

Here we discuss the construction of a 3D mesh representation of the scene, and of an augmented geometry for curvature-based affinity computations.

3.2.2.1 Dense Monocular Reconstruction

A estimate of the scene $\hat{S}$ is reconstructed in an on-line fashion as the camera browses the scene. We use the real-time camera tracking system PTAM [40], a fast dense stereo module, and a globally optimal depth map fusion algorithm. The latter component takes as input depth maps and camera poses obtained from the

former components, and computes a dense surface using an implicit volumetric representation via a truncated signed distance function (TSDF), similar to [97, 96, 26]. Dense depth maps are computed using multiview plane-sweeping [16] on a set of images and their camera poses obtained from the tracking module. The main advantage of this approach is that it allows for arbitrary scene topology. It is also closely related to the Kinect Fusion [61], although we do not employ a depth sensor but work solely based on image data. Example surface normal and depth maps extracted from our dense reconstruction are shown in figure 3.2.

For the purposes of computing affinities between points on $\hat{S}$, we construct discretized mesh derived from the regular voxelization of the reconstructed scene. Affinities between regions of the scene can be computed as functions of the nodes of this mesh. Each node q aggregates the spatial information (mean location, mean normal, Sec. 3.2.2.2) and image-based topological cues (Sec. 3.2.3.1) of the surfaces passing through the associated voxel.

3.2.2.2 Computing Geometric Affinity

Affinities (or distances) between regions of the scene can be computed as functions of the nodes of this mesh. Following standard geometric mesh segmentation schemes ([15, 7, 49]) we use surface curvature as a heuristic for partitioning contiguous mesh regions. In particular, we try to cut the mesh at "creases", regions where one of the principal curvature directions dominates the other. At each mesh node, we compute the local principal curvatures $k_1(q) > k_2(q)$, and the corresponding principal directions $v_1(q)$ and $v_2(q)$, by fitting a second-order surface to that node and its neighbors. The scalar field $K(q) := \max\{0, \frac{k_1(q)^2}{k_1(q)+|k_2(q)|}\}$ computed at every mesh node measures the strength and dominance of the most-positive eigenvalue, $k_1(q)$ (see Figure 3.3). The augmented geometric distance between two points q_i and q_j on the mesh is computed as a K-weighted mesh geodesic,

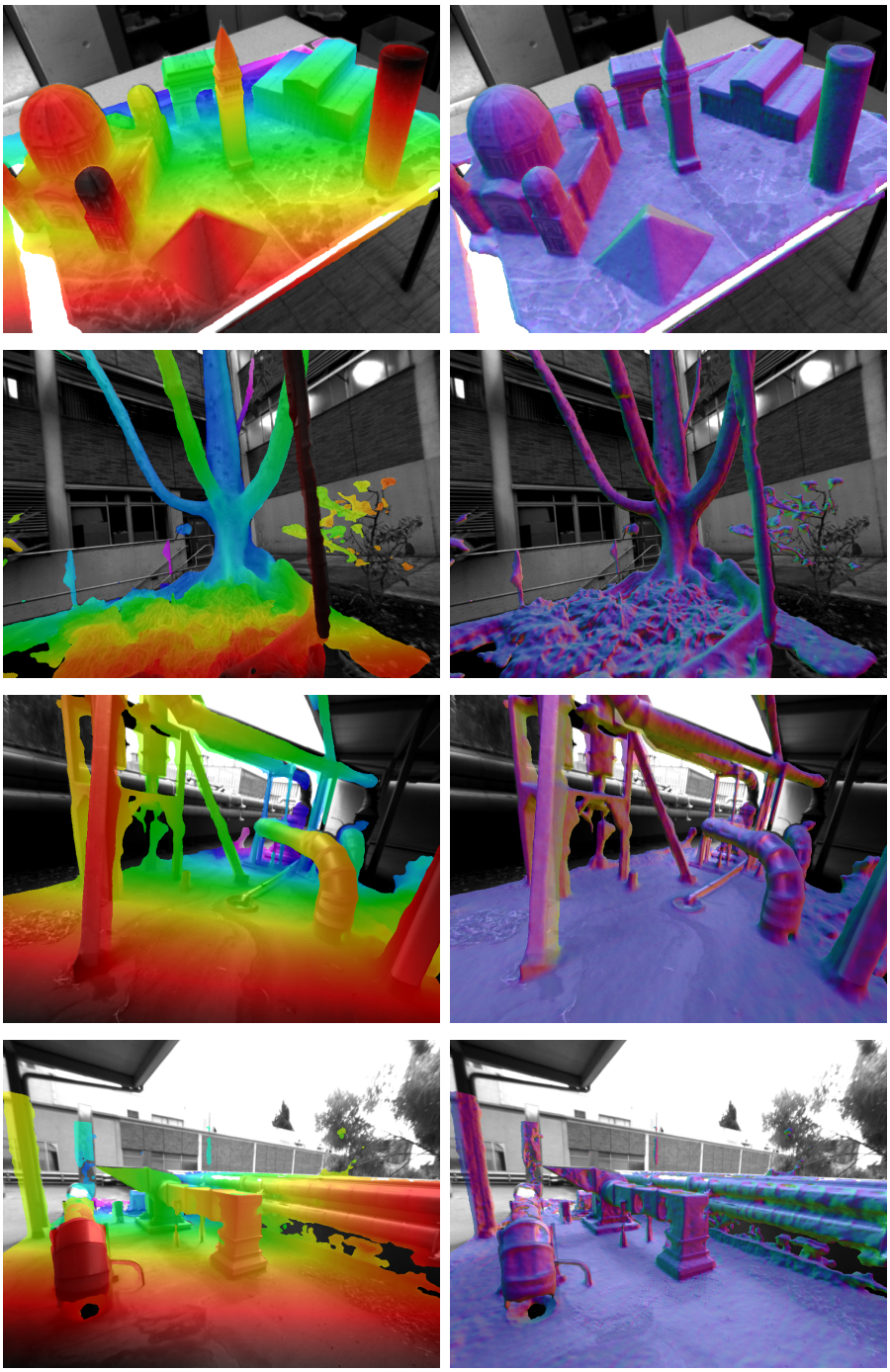


Figure 3.2: Dense reconstructions of indoor and outdoor scenes from monocular video: depth (left column) and normal (right column) maps. From top: *City of Sights*, *Tree*, *Industrial1*, *Industrial2* (described in section 3.3.2).

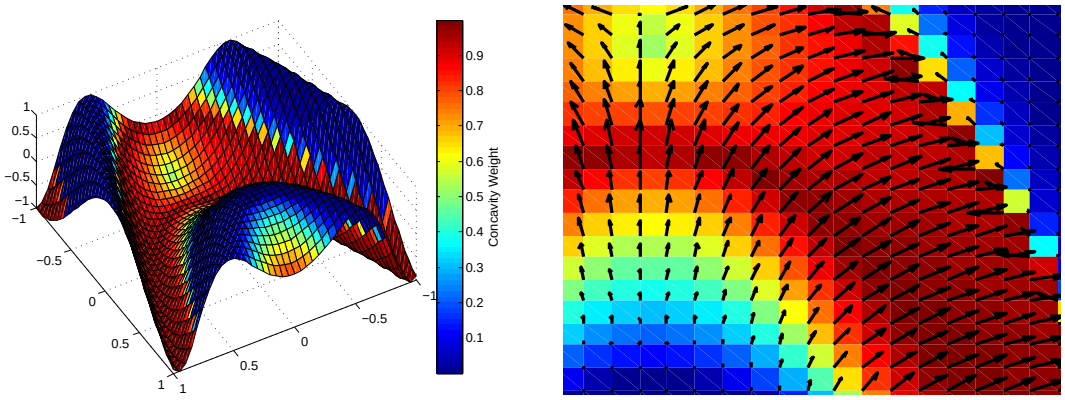


Figure 3.3: Illustration of curvature penalty. Paths are penalized which pass over regions with high concavity, especially in the direction of greatest concavity. At left, we show the scalar field $K(q) = \max\{0, k_1(q)^2/(k_1(q)+|k_2(q)|)\}$, on a sample curved surface, where $k_1(q)$ is the greatest positive principal curvature component, with associated vector field $v_1(q)$. At right, we show the weighted field $K(q)v_1(q)$ at each point.

that is, as the minimum path length $d_G(q_i, q_j) = \min_{\substack{\{s_0 \to \cdots \to s_n \\ q_i = s_0,\ q_j = s_n\}}} d_G(s_0, \ldots, s_n)$, where the $s_0, \ldots, s_n$ are a sequence of connected intervening nodes, $\{A_1, A_2\}$ are scalar weights, and

$$d_G(s_0, \ldots, s_n) := \sum_{i=1}^{n} \Big(\underbrace{\|s_i - s_{i-1}\|_2}_{\text{Path length}} + \underbrace{K(s_i)}_{\text{Concavity weight}} \big(A_1 + A_2 \underbrace{|(s_i - s_{i-1}) \cdot v_1(s_i)|}_{\text{Path-component in direction of greatest concavity}} \big) \Big) \tag{3.2}$$

A segmentation of the scene using these geometric homogeneity cues is a standard approach to 3D mesh segmentation, and is used as a baseline with which to evaluate our scale-aware segmentation in Sec. 3.3. A drawback of a purely geometric approach is that there is no unique scale appropriate for a task-relevant segmentation (such as segmenting potential objects for manipulation), as scenes typically consist of multiple scales of geometric primitives with strong violations of homogeneity between them (think of a coffee mug or a pineapple).

3.2.3 Constructing an Occlusion-Informed Geometry

To upgrade this curvature-augmented geometry to one capable of supporting queries beyond geometric homogeneity, trained detectors are typically used (e.g. [28, 46]) to provide semantic homogeneity cues. In lieu of a battery of trained detectors for a fixed set of object classes, we obtain cues of topological characteristics *as seen by the viewer* through occlusions and use these to adapt the granularity of the distances on the scene to one relevant to the viewer's motion relative to the scene.

3.2.3.1 Single Image Occlusion-Based Segmentation

Salient occlusion boundaries provide a strong topological cue as to the arrangement of surfaces in the scene from a given vantage point and motion. Furthermore, they are derived from the measurements entirely at runtime and not dependent on prior training data. We implement the linear program formulation of [8], which employs occlusion relationships between regions on the image plane as constraints on a depth-ordering of the image based on low level photometric or geometric homogeneity cues. Occlusion boundaries are obtained from salient depth discontinuities using the known geometry of the scene, and the segmentation is performed on a superpixelization of the image derived from the projected areas spanned by voxels associated to nodes of the scene mesh. These nodes are coarsified based on proximity to generate a computationally tractable number of superpixels which respect geometric boundaries in the images. Affinities between neighboring superpixels are found by computing the cost d_G between their corresponding nodes. Sample single image segmentations are shown in section 3.3.2. Since the presence of salient occlusion cues is dependent on the viewpoint (and motion) of the camera, the back-projections of these segmentations give us a homogeneity cue *relevant to the viewer's motion* to combine with the geometric homogeneity cues

of Sec. 3.2.2.2. To use these image segmentations c_t as topological homogeneity cues, we aggregate a history $C(q) = \{C_t(q)\}_{t=1}^{T}$ at each node q. If the node q is visible in the image at time t, then $C_t(q)$ takes the mode of assignments in c_t corresponding to the area its voxel subtends on the image plane. Zeroes in the history $C(q)$ denote frames in which the point p is not visible. A key assumption is that segmented regions in the images will be spatially consistent when back-projected onto the scene. If they consistently have disagreeing labels, then they were typically considered to occupy different depth-layers from the viewer's perspective and are likely not part of the same region. We quantify this by accumulating a penalty d_L along traversals of the scene that cross consistent image segmentation boundaries. This penalty is the normalized total number of frames for which the segmentation assignment changes, at least once, along the path (Eq. 3.3), as illustrated in Fig. 3.4.

$$d_L(s_0, \ldots, s_n) := \frac{1}{T}\underbrace{|\{t \ : \ \exists i, j \in 0, \ldots n, \ 0 \neq C_t(s_j) \neq C_t(s_i) \neq 0\}|}_{\text{Frames in which the layer assignment changes between } s_0 \text{ and } s_n} \tag{3.3}$$

Note that $0 \leq d_L \leq 1$.

3.2.3.2 Occlusion-Constrained Geometric Affinity

Secs. 3.2.2.2 and 3.2.3.1 present two different traversal costs along the nodes of the mesh. d_G models deviations from geometric homogeneity, and d_L models violations of image topology informed by occlusions. Nonparametric segmentation techniques (such as those used in Sec. 3.2.4) are preferred for generic segmentation tasks due to their ability to select the number of segments automatically. As a consequence, any combination of d_G and d_L between two nodes must change the structure of the resulting scene distance matrix at all scales in order to be effective. For example, a cost $d(q_i, q_j) = d_G(q_i, q_j) + d_L(q_i, q_j)$ will amplify geometric distances linearly when $d_L(q_i, q_j) \neq 0$, and have no impact on distances within

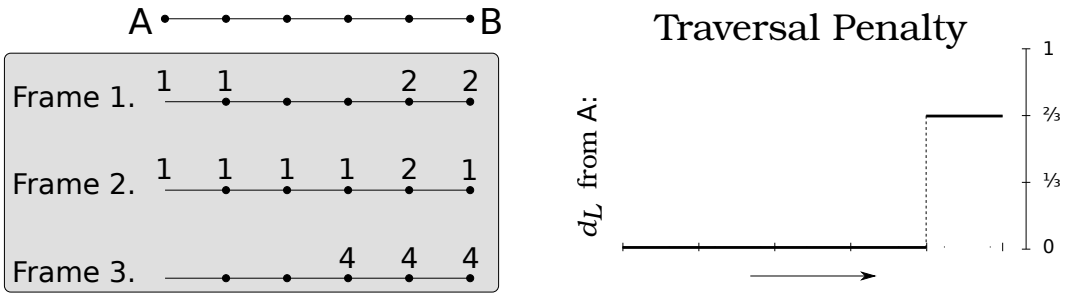


Figure 3.4: At left, example back-projected image segmentation labels c_t from three frames, over a sequence of nodes traversed from 'A' to 'B'. At right, the traversal penalty d_L accumulated over the traversal due to passing through nodes with conflicting image segmentation histories. The fact that some nodes are not visible in some frames means that penalties are not incurred along the same boundaries, depending on the direction of travel.

contiguous regions bounded by occlusions cues in the images. This will likely lead to over-segmentations of the scene in those regions when using non-parametric methods. Therefore a key design criterion for our adaptation of the geometric costs d_G between nodes using d_L is that they be *attenuated* when d_L is small and *amplified* when d_L is close to one.

To achieve this, and serve the dual purpose of providing a natural conversion from distances d_G to affinities for segmentation, we compute the cost of traversing a path between successive adjacent nodes on the scene using the Geman-McClure robust penalty (Eq. 3.4) with a scale parameter $\sigma_{\alpha,\varepsilon}(d_L)$ (Eq. 3.5).

$$d(q_i, q_j) = \min_{\substack{s_0 \to \cdots \to s_n \\ \{q_i = s_0,\, q_j = s_n\}}} \left(1 + \frac{\sigma_{\alpha,\varepsilon}(d_L(s_0, \ldots, s_n))^2}{d_G(s_0, \ldots, s_n)^2}\right)^{-1}, \tag{3.4}$$

$\sigma_{\alpha,\varepsilon}(d_L)$ acts as a *scale shaping* function, the goal of which is to locally adapt the scale of the distance function based on the available evidence for object boundaries. If minimal evidence is present (d_L is small), $\sigma_{\alpha,\varepsilon}$ will be large and attenuate the increase in distance. If d_L is large and a consistent boundary in back-projected image segmentations is present then $\sigma_{\alpha,\varepsilon}(d_L)$ will shrink, accelerating the increase

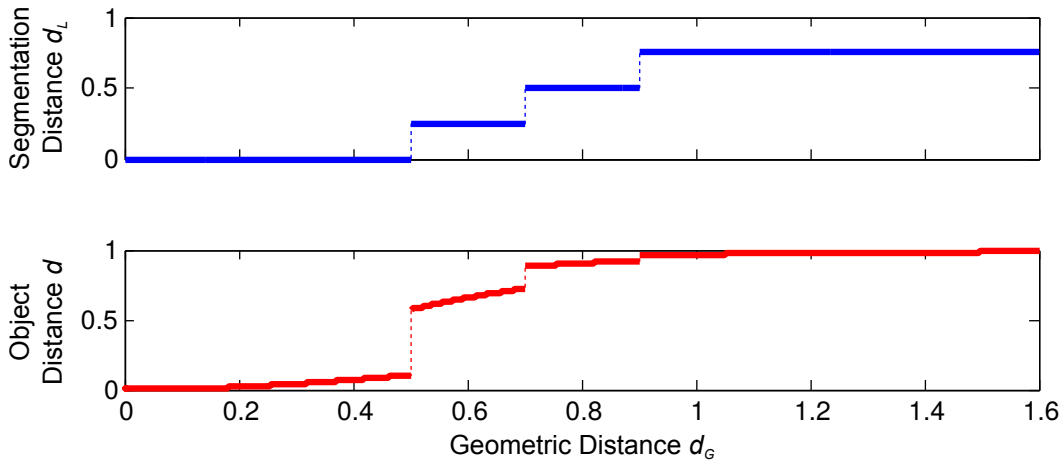


Figure 3.5: Example behavior of the combined traversal cost in Eq. 3.4 using artificial data. Note that as d_L increases, the rate of increase of d as a function of d_G accelerates up until saturation of the robust penalty.

in distance. The parameters α and ε control the rate of decreasing scale and boundary values (at $d_L = 0$ and $d_L = 1$) respectively. Fig. 3.5 shows an example of the behavior of this choice of combined geodesic and adaptive scale using artificial data.

$$\sigma_{\alpha,\varepsilon}(d_L) = \frac{1 - \exp(-\alpha(1 - d_L) - \varepsilon)}{1 - \exp(-d_L - \varepsilon)} \tag{3.5}$$

The set of distances $d_i := \{d_{ij} \ : \ j \in S\}$ is computed in $O(|S|\, n_{\text{frames}})$ time, using a modified Dijkstra's algorithm on the nodes of the scene mesh.

3.2.4 Scene Segmentation

Discrete representations of complex scenes at high resolution can typically consist of many thousands of nodes, making both computation and storage of a pairwise distance matrix of geodesics between all nodes infeasible. To make segmentation based on pairwise distances tractable we generate a subgraph of the scene by

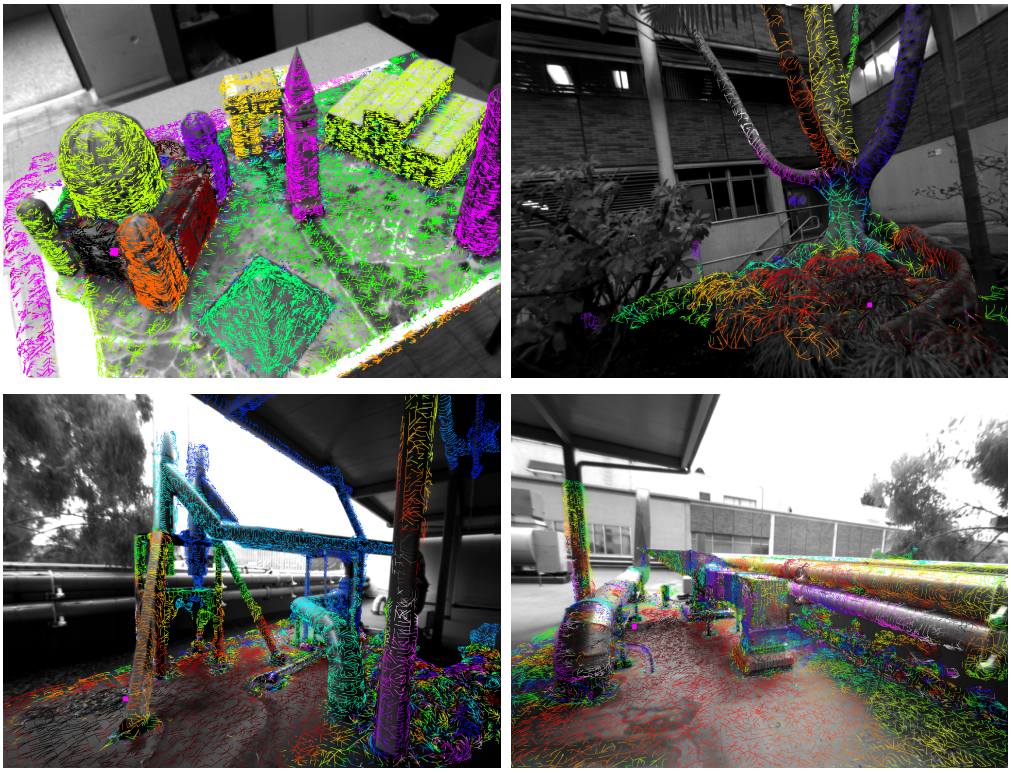


Figure 3.6: Sample geometric traversal costs d_G from a single point to all others across the scene overlaid on images. Colored lines indicate the path of the geodesic on the scene originating at the magenta square in each image, with distance increasing with changing colors, starting from zero (black).

uniformly sampling nodes subject to a minimum Euclidean distance and compute geodesics between them using the full resolution representation. To obtain a sparse segmentation of the scene, we apply a graph-based version of the DP-Means [42] algorithm, a low-variance asymptotic clustering algorithm derived from the Dirichlet process Gaussian mixture model [82]. The DP-Means algorithm was chosen for its nonparametric nature, i.e. its ability to select the number of objects in the scene automatically, and its computational speed. However, the original algorithm is only applicable to clustering data in a normed vector space; thus, we find an initial segmentation of the subgraph by globally optimizing a spectral relaxation [98, 93] of the DP-Means cost, and refine the segmentation via kernelized [19] iterative updates. The partitioned subgraph is projected back to full mesh by performing a Voronoi tessellation of the scene discretization using the previously computed geodesics from each node in the subgraph.

A possible future extension of our segmentation pipeline is to enable adaptation to dynamically changing scales by treating more recent images segmentations preferentially, either through a fixed sliding window or decaying components of $d_L(s_0, \ldots, s_n)$. The Dynamic Means [13] algorithm, a low-variance asymptotic clustering algorithm based on the dependent Dirichlet process Gaussian mixture [55] can be used to make such a segmentation strategy temporally consistent. As in the batch case, this algorithm is ideal due its nonparametric ability to automatically discover the number of objects in the scene, and for its computational speed. However, Dynamic Means suffers from the same limitation as DP-Means; it is only applicable to data in a normed vector space. Therefore, for each video frame, we find an initial segmentation of the single frame alone using spectral clustering, and then enforce temporal consistency in the segmentation by using kernelized refinement iterations based on the Dynamic Means cost.

3.3 Evaluation

3.3.1 Comparison Methodology

We are not aware of benchmarks for evaluating scene segmentation inferred from monocular vision. While several RGB-D datasets for reconstruction and segmentation exist (such as [90, 59, 81, 3]) they tend to have highly variable viewpoint scale, which makes the appropriate scale of a segmentation (based on occlusion cues) vary over time, in addition to having very similar scene content and geometry (indoor office and home scenes). Therefore, we have captured a set of diverse sequences, both indoor and outdoor, on which to test our dense reconstruction and segmentation pipeline, and included one provided by [26]. To generate ground truth for each sequence we manually segment the reconstruction into regions that correspond to objects at a scale appropriate for interaction from the video's perspective. As noted in section 3.2.1, objects compose a nested covering of sets, and the choice of which partition to use for groundtruth is subjective, though we have endeavored to select a fair partition into the dominant objects as visible from the videos. We compare the results of our occlusion-constrained segmentation to a baseline of segmentation using standard geometric homogeneity cues (described in section 3.2.2.2), for which typically a fixed scale parameter must be selected to perform segmentation, chosen to preserve as many of the dominant objects as possible without over-segmenting them. Note that it is infeasible for a single scale to accurately segment the entire scene, necessitating our adaptation of scale using occlusion cues from the viewer's motion. Baseline segmentations are compared numerically to object-level segmentations through F-score and precision-recall metrics. F-scores are computed following standard methodology [63], to determine the agreement between ground-truth segments and computed segments.

Given a correspondence between computed cluster c_i, $i \in I$ and ground-truth regions g_j, $j \in J$, we compute F-scores as follows. Precision P_{ij} and recall R_{ij} are

computed as the average (weighted by cluster size) ground-truth fraction of clusters, $P_{ij} = |c_i \cap g_j|/|c_i|$, which penalizes under-segmentation, and the fraction of the corresponding ground-truth region covered by a cluster, $R_{ij} = |c_i \cap g_j|/|g_j|$, which penalizes over-segmentation. A compromise measure $F_{ij} = 2P_{ij}R_{ij}/(P_{ij} + R_{ij})$ penalizes both. An optimal correspondence $\phi : I \to J$ is found by the Hungarian algorithm, maximizing the the total F-score,

$$F = \max_{\{\phi: I \to J\}} \frac{1}{|I|} \sum\nolimits_{i \in I} F_{i\phi(i)}. \tag{3.6}$$

A precision-recall curve may also be computed by comparing a thresholded affinity matrix $\delta_{M_{ij}>t}$ with the ground truth affinity matrix $\delta_{g_i=g_j}$. Since this is a monotonic function of distance along the scene, the precision-recall curve sampling over affinity thresholds allows us to evaluate the segmentation results across a range of scales.

3.3.2 Geometric and Occlusion-Constrained Segmentation Results

We present results in the form of re-projected segmentations and sample geodesics on four geometrically and topologically complex scenes, three of which were collected outdoors (Park, Industrial1, and Industrial2) and two of which contain notable multi-scale geometry (Park and City of Sights (CoS), made available by [26]). Please refer to our supplemental material for video results on these sequences.

Fig. 3.7 shows sample images and re-projected groundtruth segmentations for each scene, Fig. 3.8 shows sample occlusion-based image segmentation results, and Fig. 3.9 shows sample occlusion-constrained geodesics on the scene built using the occlusion-based image segmentations. Fig. 3.10 shows qualitative examples of our baseline geometric segmentation. Fig. 3.11 shows qualitative examples of our final scene segmentations, with numerical evaluations relative to groundtruth shown in Fig. 3.12.

In the CoS sequence the multi-scale geometry of the domed structure (a single object) makes the selection of a single scale infeasible, however the adaptive geodesic is able to shape distances on the scene based on the available image segmentations, enabling a correct segmentation. The Park sequence shows a scene with complex natural geometry on the ground and smoothly varying geometry on the nearby tree. Highly variable ground geometry makes the selection of a single scale unable to correctly segment both the smooth tree limbs (which occlude each other throughout the sequence) and the rough ground (as a single object). These sequences demonstrate that adapting the geometry using occlusion-based image segmentation enables segmentation at the appropriate viewpoint scale for all dominant objects in the scene. Both Industrial sequences shows scenes of sophisticated topology and geometry, the segmentation of which is improved using our occlusion-informed geometry compared to the over-segmented results using standard geometric approaches.

The quantitative evaluation of Fig. 3.12 demonstrates a consistent improvement over a standard geometric segmentation when compared to our adaptive geodesic. The Precision-Recall curve comparisons serve to provide further support to the claim that no fixed scale treatment of purely geometric distances can correctly segment scenes with complex geometry, as varying affinity scale for both methods typically shows an increase in performance for our adaptive geodesic distances.

Fig. 3.13 shows a summary of timings for the various components of our system. Typical size of meshes used to represent the scene geometry are on the order of fifty thousand, with coarsified meshes for image segmentation on the order of two thousand, and subgraphs used for clustering on the order of several hundred. The entire system runs on a 3.5Ghz desktop machine (dense reconstruction, image, and scene segmentation).

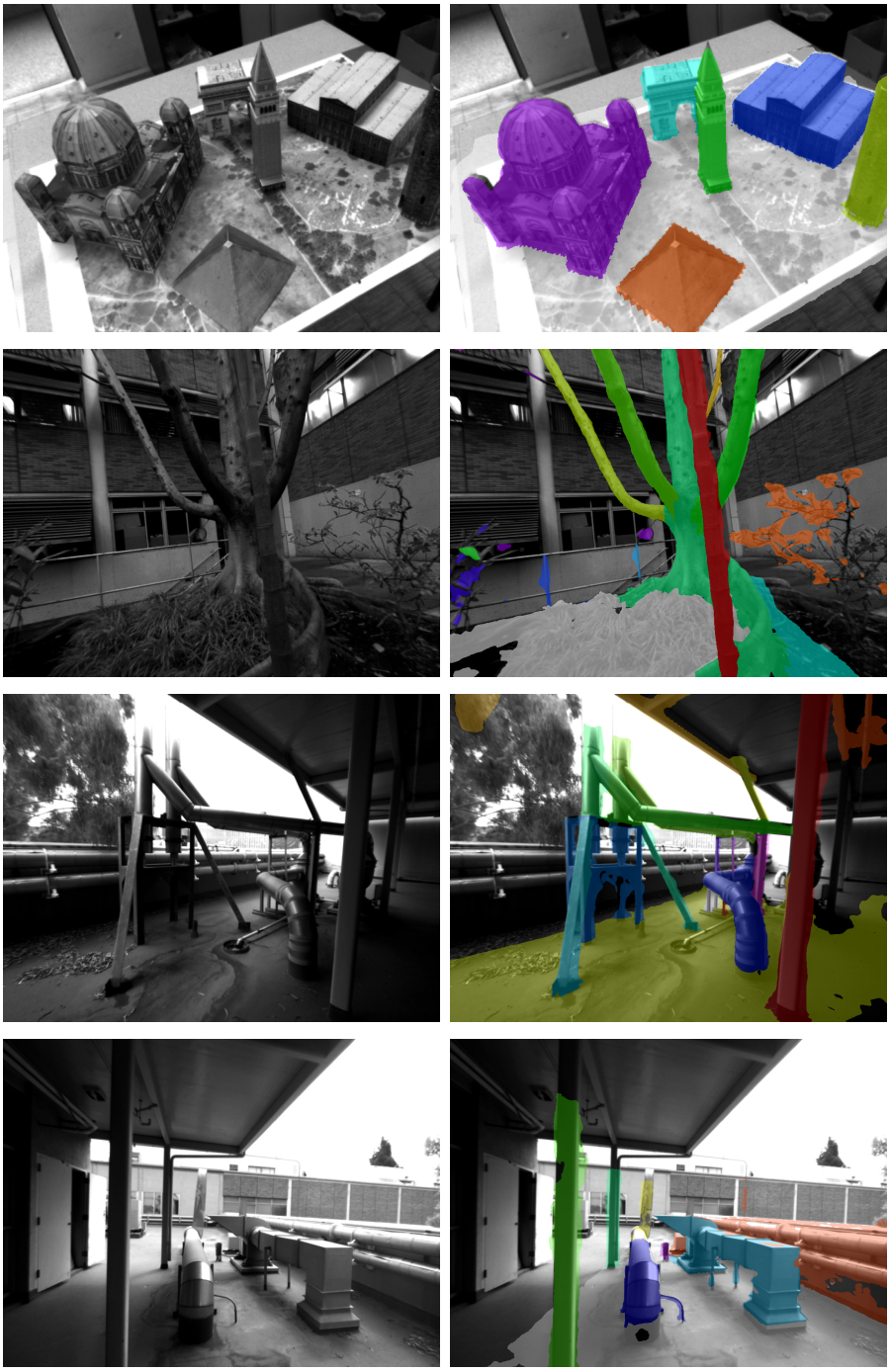


Figure 3.7: Sample frames (left) and re-projected groundtruth segmentations (right) for CoS (top), Park, Industrial1, Industrial2 (bottom). Different colors indicate different segments.

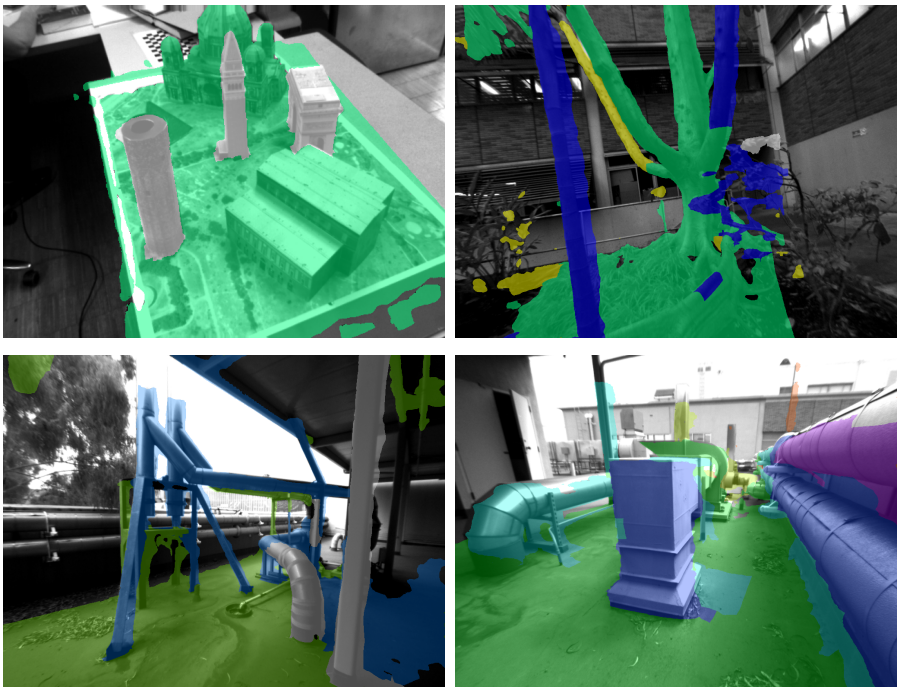


Figure 3.8: Sample single-image segmentation on frames from CoS (top left), Park (top right), Industrial1 (bottom left), Industrial2 (bottom right) inferred as described in Sec. 3.2.3.1. Different colors indicate different segments from occlusion-guided segmentation.

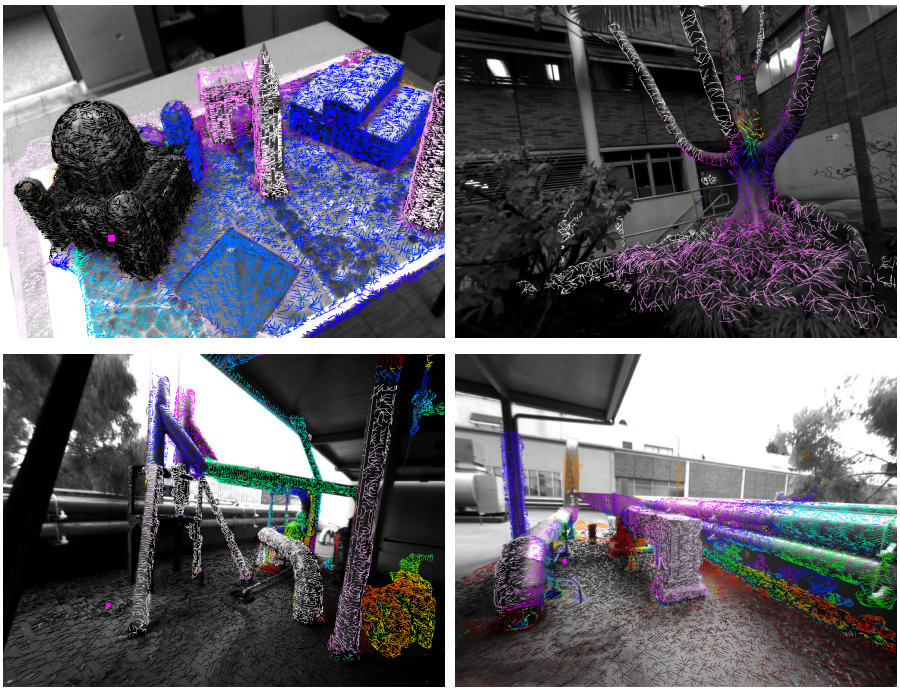


Figure 3.9: Sample occlusion-constrained geodesics on CoS (top left), Park (top right), Industrial1 (bottom left), Industrial2 (bottom right) built as described in Sec. 3.2.3.2. Colored lines indicate the path of the geodesic on the scene originating at the magenta square in each image, with distance coded between zero (black) and one (white) through intervening colors.

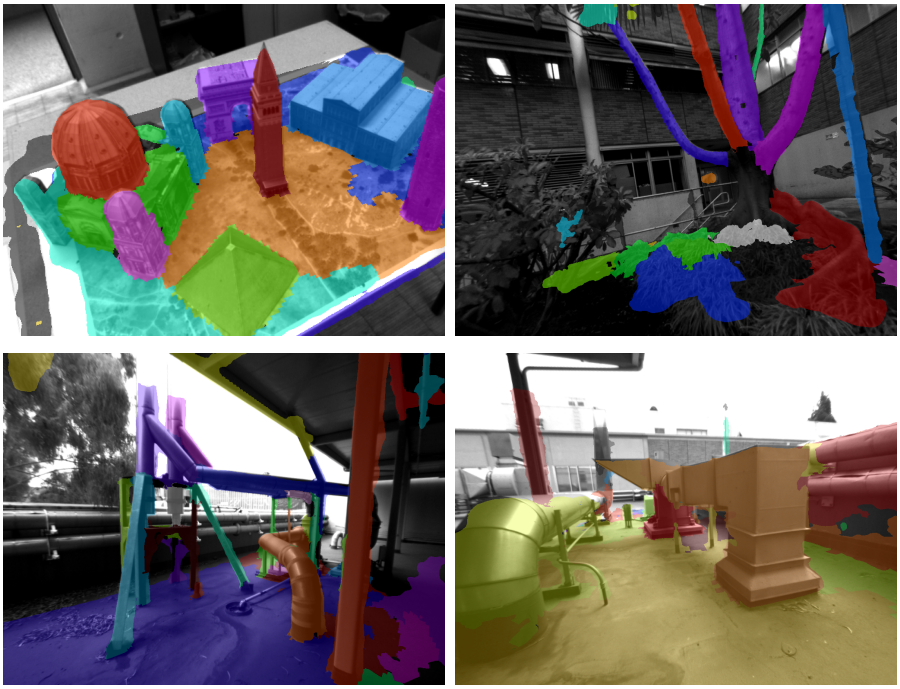


Figure 3.10: Sample re-projections of our baseline geometric segmentation on CoS (top left), Park (top right), Industrial1 (bottom left), Industrial2 (bottom right). Different colors indicate different segments.

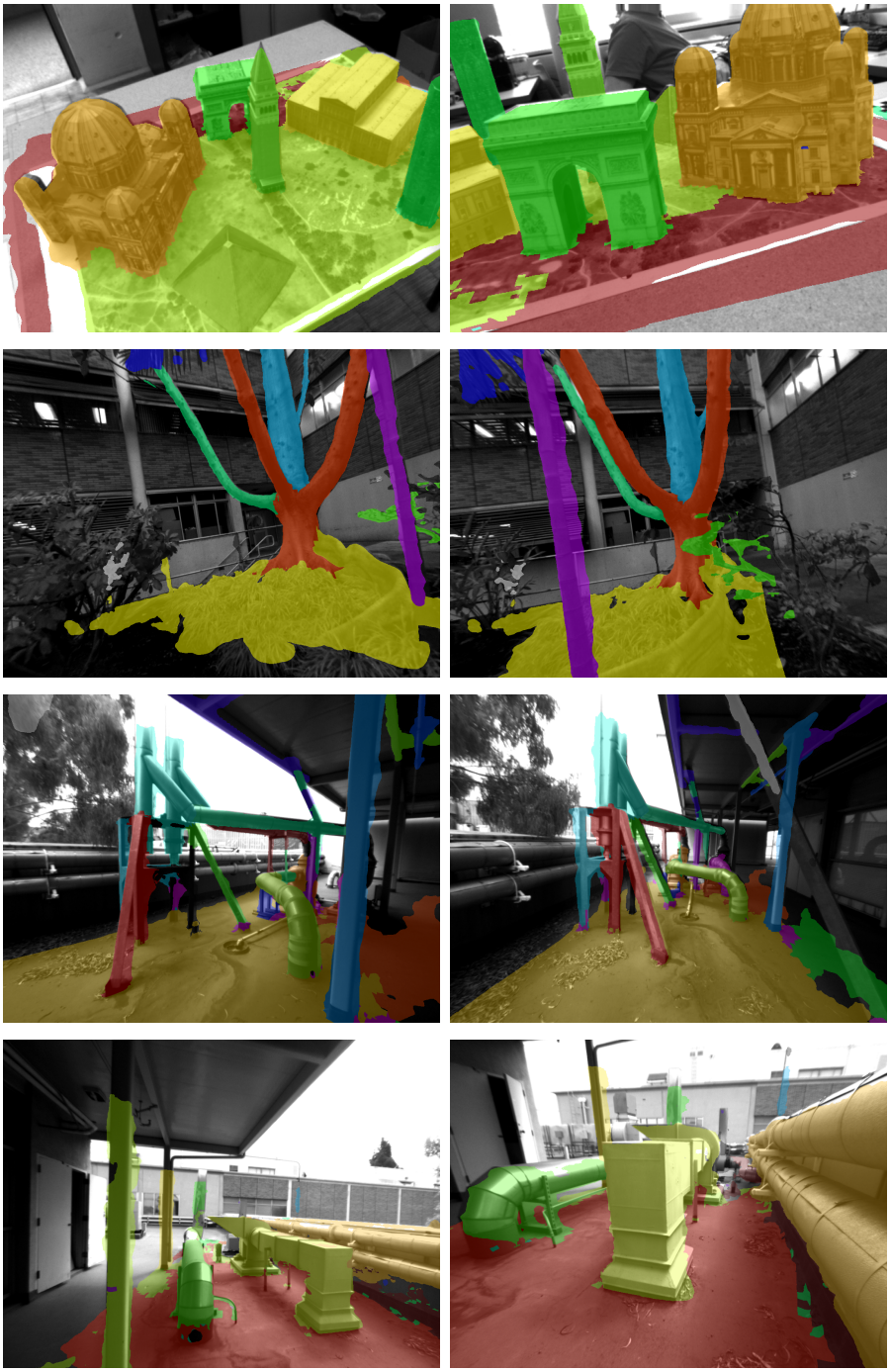


Figure 3.11: Sample re-projected segmentation results using our occlusion-constrained geodesics for CoS (top), Park, Industrial1, Industrial2 (bottom). Different colors indicate different segments.

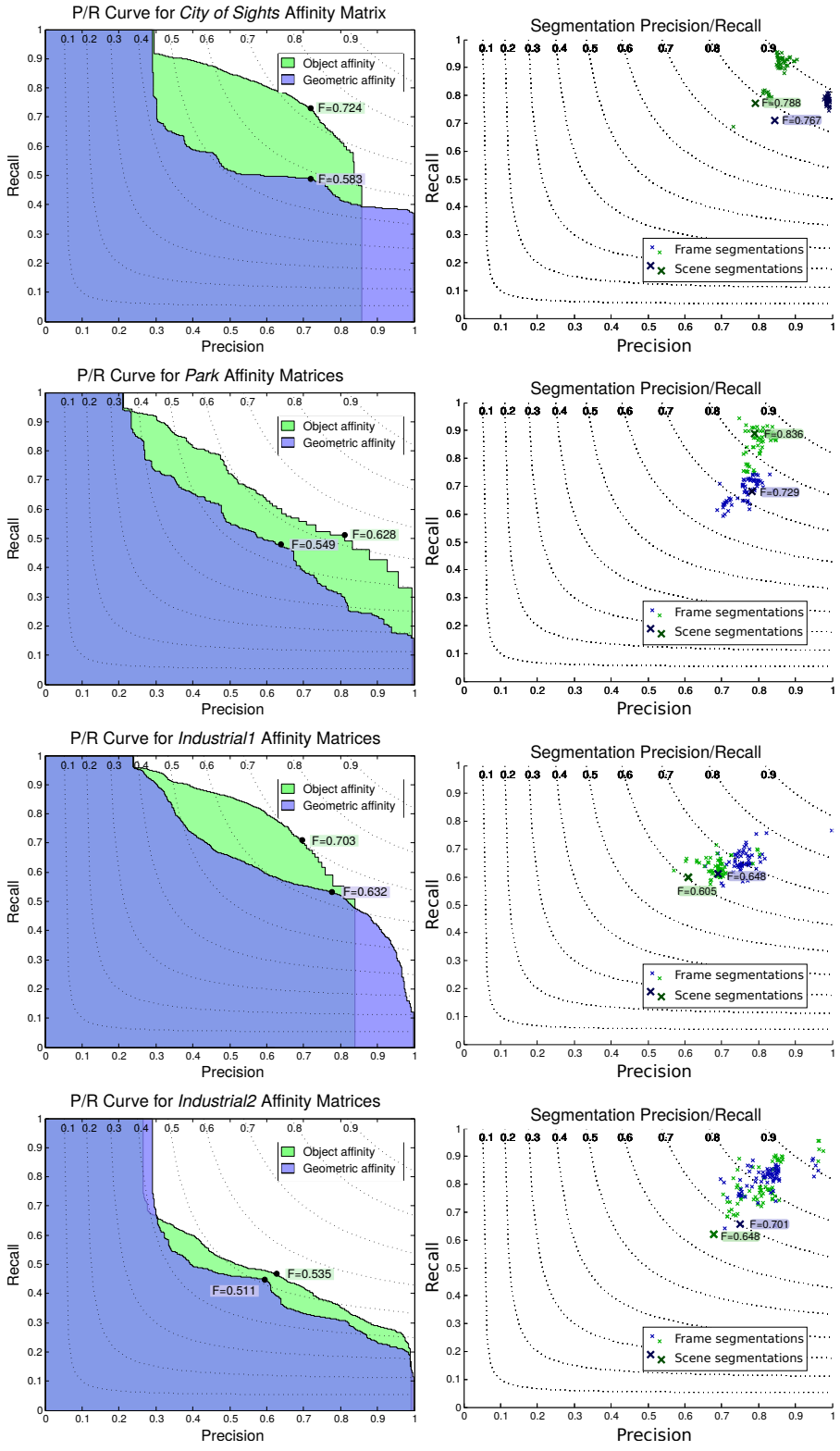


Figure 3.12: Quantitative evaluation of our segmentations vs. groundtruth. The curves at left show precision and recall scores induced by thresholding the affinity matrix at levels running from 0 to 1. Highlighted points indicate the best threshold for clustering, vis-à-vis the ground truth. At right, precision and recall are computed from final

Component	Time	Hardware
1. Dense Reconstruction	30Hz Realtime	GPU
2. Image Segmentation	3s / frame	1 × CPU
3. Construct Mesh	.2s / frame	1 × CPU
4. Compute Geodesics	0.6s / traversal	N × CPU
5. Segmentation	1.2s	1 × CPU

Figure 3.13: Approximate timings on the CoS sequence with a final mesh size of roughly 57000 points. GPU is an Nvidia GTX 780. Note that traversing the mesh to compute geodesics for the sampled subgraph is parallelized. Typically a subgraph of several hundred nodes is used.

3.4 Conclusions

We have presented a method to endow a scene, as densely reconstructed from monocular video, with a metric that incorporates geometric and topological information as seen by the viewer, as well as back-projected image statistics. While the latter are temporally inconsistent (they change with the video), the way they change is spatially consistent, an observation key to defining distances or affinities that allow us to partition the scene into coherent "objects". While one could employ trained object detectors at the outset to arrive at a semantic segmentation of the scene (and, by simple forward projection, of the video), we focus on low-level geometric and topological cues first, to segment the image into coherent regions, where one could then deploy object detectors if so desired. Semantic analysis of the scene involves object identities and relations, and knowledge of scene geometry and topology is key to infer the latter. This is our focus in this work.

CHAPTER 4

Information-Driven Autonomous Exploration

4.1 Introduction

We describe an information-gathering approach for exploring an unknown scene using a range sensor. Our explorer maintains a map of known and likely obstacles, under an Ising-like prior distribution, and follows a greedy best-next-view policy, whereby uncertainty at the next timestep is minimized by maximizing the uncertainty of the next range measurement. Uncertainty is efficiently approximated by a novel Poisson disk sampling technique. Our algorithm improves the performance of recent visibility-based planning approaches that come with guaranteed performance bounds on the expected path length to complete exploration, and extends them to allow exploration of an unbounded region.

4.2 Prior Work

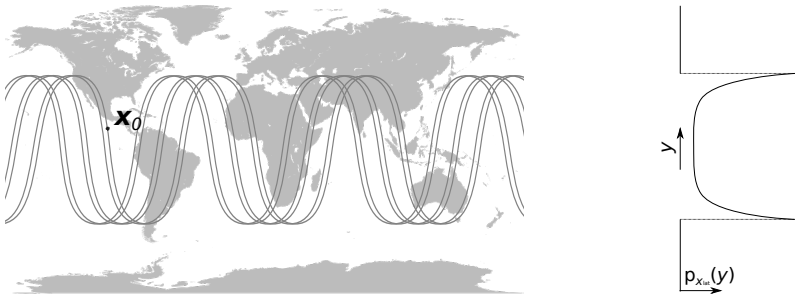
Information-driven visual exploration has a long history, both for the case of eye movement and for the more general case of full mobility. The former is relevant to visual attention and oculomotor control, where saccadic motions are hypothesized to be related to the uncertainty on the irradiance in different locations of the visual field. Information gain (uncertainty reduction) occurs due to the uneven distribution of sensing elements in the retina (foveal vision). Since we are interested in uniformly-sampled omnidirectional sensors, customary in robotics applications, no information gain can occur as a result of gaze control. Therefore,

we focus on the case of parallax motion (translation of the optical center), where uncertainty is due to *occlusions* as well as *scaling* phenomena that can be reduced by control of the vantage point. The use of information-theoretic criteria for visual exploration dates back to the literature on Active Vision ([89], [5], [94], [12]). For range sensors, where most of the uncertainty is due to occlusions, entropy can be related to visibility, and therefore several have adopted geometric criteria ([17], [95], [91], [70], [47]). Much of this work is concerned with a greedy approach to information maximization by seeking the "best next view" for a particular task, which could be recognition or manipulation ([70], [18], [67]). The problem is also addressed in the context of optimal control ([4], [78]), sequential decision [2], "Value of Information" [66], partially-observable path planning ([31], [56]), among others. In some cases, the problem exhibits submodular characteristics that make it amenable to be solved with efficient algorithms with provable guarantees [64]. For instance [23] perform underwater inspection, assuming a known map, exploiting submodular optimization. Unfortunately, our setting is not submodular, and therefore proper formalization in terms of optimal control or sequential decision processes would yield an intractable inference problem. As customary, therefore, we seek for surrogate criteria that yield algorithms with provable guarantees.

4.3 Information-Driven Exploration

In this chapter we consider the problem where the variable of interest is the geometry and topology of a two-dimensional scene (e.g. the map of a building) represented by an indicator function defined on a domain $\Omega \subseteq \mathbb{R}^2$, supported on a closed subset $\mathcal{O} \subseteq \Omega$ that describes unknown "obstacles" or "objects." The explorer follows a continuous path $x(t)$ in the space $\Omega \setminus \mathcal{O}$. Let $\mathcal{A} \subseteq \Omega \setminus \mathcal{O}$ be the path-component of $\Omega \setminus \mathcal{O}$ containing $x(0)$.

At discrete timesteps t_i, $i = 1, \ldots, L$, the explorer at $x_i = x(t_i)$ measures an

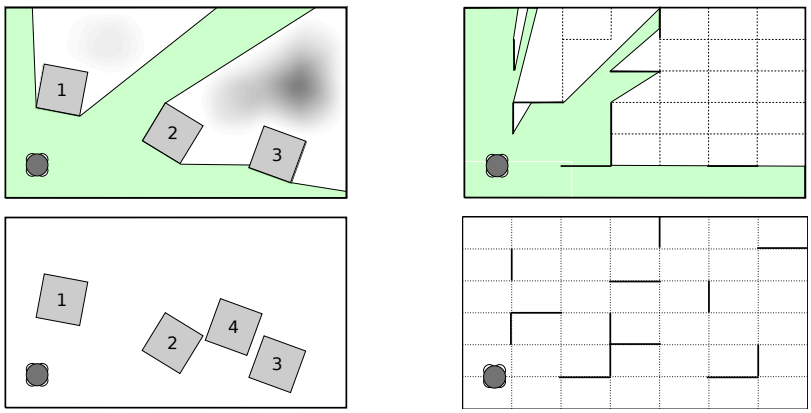


https://upload.wikimedia.org/wikipedia/commons/e/ec/World_map_blank_without_borders.svg

Figure 4.1: *Lost Astronaut Prior Distribution* The crash site $\mathbf{x}(t)$ is defined by a random time t, chosen uniformly in $[0, T]$. This induces prior marginal probabilities on the latitude and longitude of the crash site. The limiting distribution for as $T \to \infty$ is shown at right, while the limiting distribution on longitude is uniform.

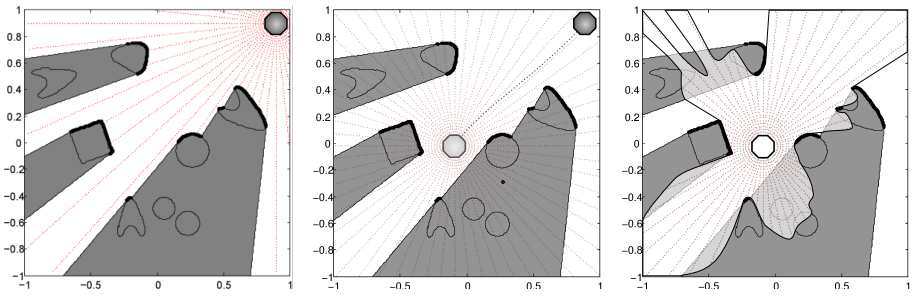
n-directional range map $\mathcal{Y}(x_i) \in (s\mathbb{N})^n$, which gives, to the nearest multiple of $s > 0$, the distance to the first obstacle along each sensor element's line of sight. Let $\mathcal{Y}^t = \{\mathcal{Y}(x_i) \,:\, t_i < t\}$ denote the history of measurements up to time t.

Observe that each measurement is a function $\mathcal{Y}(x) = h(x, \Omega, \mathcal{O})$ of the scene $(\Omega, \mathcal{O})$ and the vantage point x. If we interpret this as a realization of a stochastic process, each measurement $\mathcal{Y}_i$, on average, reduces the uncertainty in $\mathcal{O}$. Although each range sensor is affected by uncertainty due to scaling, spatial quantization, noise, etc. most of the uncertainty is due to occlusions, for objects are opaque and therefore at the outset we have no knowledge of the scene behind an occluder. An information-gathering controller then aims to design the control $\{u_j\}_{j=i}^{L}$ that maximizes the reduction in the entropy of $\mathcal{O}$. Here we will assume for simplicity that we can control the instantaneous velocity, so that $x_{i+1} = x_i + u_i$.



(a) Freely-placed obstacles (b) Gridded obstacles

Figure 4.2: *Exploring a Random Room* (a) In the first case, the explorer knows that there are four identical boxes scattered within a room of known dimensions, three of which he can see. There is a "shadow region" where the disposition of space is not known. Based on allowable configurations of the missing block (blocks can not intersect, unseen blocks can not lie in known free space), the explorer can compute, at each point in the shadow region, the probability that that point is covered by an obstacle. (b) In the second case, partitions (black line segments), constrained to grid lines (dotted lines) divide up a room. Seeing any point of a grid segment reveals the disposition of any other point on that grid segment.



(a) First View (b) New Vantage Point (c) Hypoth. Second View

Figure 4.3: *Range Measurement Planning* An explorer

4.4 Next-View Entropy

In order to compute the information gain of a measurement at time t, we have to define suitable distributions on the objects of interest. We maintain an estimate of the explorer's state at time t as a distribution $\mathbb{P}_t[\,\cdot\,] = \mathbb{P}[\,\cdot\,|\mathcal{Y}^t]$ on possible realizations of $\mathcal{A}$. Let $\mathcal{A}_t = \{x \in \Omega : \mathbb{P}_t[x \in \mathcal{A}] = 1\}$, $\mathcal{O}_t = \{x \in \Omega : \mathbb{P}_t[x \in \mathcal{A}] = 0\}$, and $\mathcal{U}_t = x \setminus (\mathcal{A}_t \cup \mathcal{O}_t)$ be the points which at time t are known to be visible, known to be invisible, or unknown, respectively. Finally, let $\mathcal{V}^t$ denote the set of points that are visible to the observer from point x_t.

4.4.1 Measurement Uncertainty

Let

$$G = \big\{g_{ij} := s\,i\big(\cos(2\pi j/n), \sin(2\pi j/n)\big)\big\}^{\,i,j\in\mathbb{N}}_{\,j<n}$$

be a uniform radial grid, centered at the origin, with radial spacing s and angular spacing $2\pi/n$ (see Figure 4.3). Although we define $\mathcal{Y}(x) = h(x, \Omega, \mathcal{O})$ as an $s\mathbb{N}$-valued random n-tuple, it is natural to define an "alter-ego" $\mathcal{Y}_G(x)$, a Boolean-valued random vector indexed by sample points $g_{ij} \in G$, with the constraint that

$$\mathcal{Y}_G(x)_{ij} = 0 \implies \mathcal{Y}_G(x)_{i+1,j} = 0.^{1} \tag{4.1}$$

Let $\mathcal{Y}(x)_{ij} = 1$ iff the point $g_{ij} + x$ is visible from x. Observe that

$$\mathcal{Y}_G(x)_{ij} = \mathbb{1}\{\mathcal{Y}(x)_j < s\,i\}.$$

This defines a bijective relation between $\mathcal{Y}_G(x)$ and $\mathcal{Y}(x)$, so $\mathbb{H}_t[\mathcal{Y}_G(x)] = \mathbb{H}_t[\mathcal{Y}(x)]$.

Our next-view energy $E_t(x)$ is the expected decrease in uncertainty at time t, due to a measurement $\mathcal{Y}(x)$:

$$\begin{aligned} E_t(x) :&= \mathbb{E}_t\big[\mathbb{H}_t[\mathcal{O}] - \mathbb{H}_t[\mathcal{O}|\mathcal{Y}(x)]\big] = \mathbb{E}_t\big[\mathbb{H}_t[\mathcal{Y}(x)]\big] \\ &= \mathbb{H}_t[\mathcal{Y}(x)] = \mathbb{H}_t[\mathcal{Y}_G(x)]. \end{aligned}$$

[1] This formalizes the notion that a visible object blocks the view of objects further along its line of sight.

We can decompose the latter into a sum of relative entropies, given a ordering $g_{o_1}, g_{o_2}, \ldots$ on G that respects radial ordering (i.e. if $o_k = (i,j)$, $o_\ell = (i',j')$, and $i' < i$, then $\ell < k$):

$$= \sum_k \mathbb{H}_t\big[\mathcal{Y}_G(x)_{o_k} \mid \{\mathcal{Y}_G(x)_{o_\ell}\}_{\ell<k}\big].$$

Now, if $\mathcal{Y}_G(x)_{i',j} = 0$ for $i' < i$, then by (4.1), we know $\mathcal{Y}(x)_{i,j}$ must equal 0 as well. In that case, the relative entropy at g_{ij} is zero, and so

$$= \sum_{o_k=(i,j)} \underbrace{\mathbb{P}_t\big[\mathcal{Y}_G(x)_{i'j} = 1 \text{ for all } i' < i\big]}_{\text{Probability that } x+g_{ij} \text{ is visible from } x}$$
$$\cdot \underbrace{\mathbb{H}_t\big[\mathcal{Y}_G(x)_{o_k} \mid \{\mathcal{Y}_G(x)_{o_\ell} \mid \ell < k\},\ \mathcal{Y}_G(x)_{i'j} = 1 \text{ for all } i' < i\big]}_{\text{Value of revealing scene at } x+g_{ij}}. \tag{4.2}$$

Here, we will reference the first term through its complement, which we will call the "extinction probability".

4.4.1.1 Extinction Probability

Define a normalized, instantaneous "mean free path" function $\mathrm{MFP} : \Omega \to [0, \infty)$, such that for any curve $C \subseteq \Omega$,

$$\mathbb{P}_t\big[C \in \mathcal{A}\big] = \exp\left(\int_C \frac{\log\big(\mathbb{P}_t(x+g_{ij} \in \mathcal{A})\big)}{\mathrm{MFP(y)}}|dy|\right).$$

Now we compute extinction probability, using a posterior estimate of the scene

$$\mathbb{P}_t\big[\mathcal{Y}_G(x)_{ij} = 1 \mid \mathcal{Y}_G(x)_{i-1,j} = 1\big] = \exp\left(\int_{\mathrm{line}(x+g_{i-1,j},\, x+g_{ij})} \frac{\log\big(\mathbb{P}_t(y \in \mathcal{A})\big)}{\mathrm{MFP}(y)}|dy|\right)$$
$$\approx \mathbb{P}_t(x+g_{ij} \in \mathcal{A})^{s/\,\mathrm{MFP}(x+g_{ij})},$$

Induction on this chain of conditional probabilities gives

$$\mathbb{P}_t\big[\mathcal{Y}_G(x)_{ij} = 1\big] \approx \prod_{i'<i} \mathbb{P}_t(x+g_{i'j} \in \mathcal{A})^{s/\,\mathrm{MFP}(x+g_{i'j})}$$

4.4.1.2 View Value

The second term in (4.2) is a conditional entropy whose computation is only tractable in certain limited situations. We will discuss these in the next section.

4.5 Obstacle Models

4.5.1 Uniform Obstacle Density

The simplest obstacle model assigns, to all points in $\mathcal{U}^t$, equal and nonzero probability p of lying in an obstacle. Additionally, it treats each sampled point as independent of all others. Thus the difficult entropy term in (4.2) reduces to a Bernoulli entropy:

$$\begin{aligned}\mathbb{H}_t\big[\mathcal{Y}_G(x)_{o_k=(i,j)} \mid \{\mathcal{Y}_G(x)_{o_\ell} \mid \ell < k\},\ \mathcal{Y}_G(x)_{i'j} = 1 \text{ for all } \ i' < i\big] &= \mathbb{H}_t\big[\mathcal{Y}_G(x)_k\big] \\ &= -p\log p - (1-p)\log p,\end{aligned}$$

where we take $0\log 0$ as 0. Additionally, we treat $\mathrm{MFP}(y)$ as constant. define

$$E_t(x) = \sum_{ij}\big(-p_{ij}\log p_{ij} - (1-p_{ij})\log(1-p_{ij})\big)\prod_{i'<i} p_{i'j}^{s/\,\mathrm{MFP}}, \tag{4.3}$$

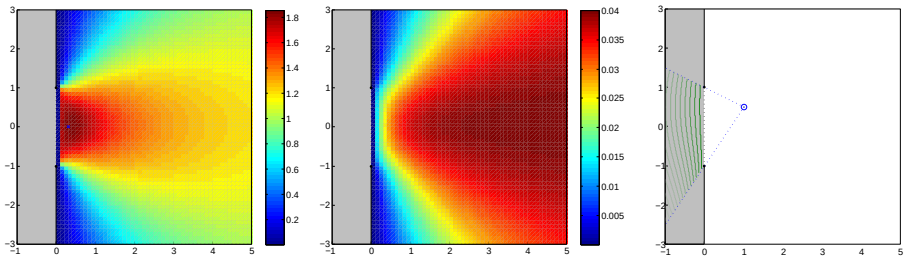
where

$$p_{i,j} = \begin{cases} 0 & \mathrm{line}(x+g_{i-1,j}, x+g_{ij}) \cap \mathcal{O}^t \neq \emptyset \\ 1 & \text{o.w, } x+g_{ij} \in \mathcal{A}^t \\ p & \text{o.w} \end{cases}.$$

The values p and MFP are model parameters.

4.5.1.1 Instant Extinction

A special case of uniform obstacle density, proposed by Valente et al. [84], this takes the limit as $p^{s/\,\mathrm{MFP}} \to 0$. To make sense of this, we treat it as a sum over



(a) Low p (b) High p (c) View Penetration

Figure 4.4: *Looking into a Closet with Uniform Obstacle Density* (a)-(b) Here we show the value of different vantage points for looking into an unexplored closet, for various values of p (MFP fixed). As p increases, a line of sight is less and less likely to penetrate deeply into the room, so the best vantage points, counter-intuitively, move further back where lines of sight are more parallel. (c) Green lines show the probability-0.5 penetration limits of the blue explorer, for different values of p. As $p \to 1$, the depth of penetration near a point $u \in \partial\mathcal{U}^t$ becomes proportional to the cosine of the incidence angle between $\vec{n}(u)$ and the line of sight.

visible shadow boundaries.

$$E_t(x) = \int_{\partial\mathcal{U}^t \cap \mathcal{V}^t} (r - x) \cdot \vec{n}(r) |dr|,$$

where $\vec{n}(r)$ is the outward-facing normal to the shadow boundary at point r. This gives greatest value to viewpoints which look directly (not obliquely) onto long, faraway (all lines of sight looking onto a faraway object are nearly parallel) shadow boundaries. The advantage of this method is that it can be integrated over shadow boundaries, although it assigns undue value to thin shadow regions which are unlikely to reveal much about the scene.

4.5.1.2 Independent Cells

A more general approach models an unknown scene as a random partition of t$\mathcal{D}$, with each cell of that partition assigned a random "color" (designation of "object"

or "free space"), independent of all other cells. Once a partition is known, all that remains is to learn the coloring. The information gain of sampling $S \subseteq \mathcal{D}$ of a randomly-colored, known partition is equal to the sum of the entropies of the colorings of all the cells sampled. Denote the probability that cell c is an obstacle by p_c. Then,

$$\begin{aligned}\mathbb{H}[\{\mathcal{Y}(x) \,:\, x \in S\}] &= \mathbb{H}[\{\mathcal{Y}(x) \,:\, x \in S\} \,|\, \mathcal{P}] \\ &= \sum_{c \,:\, c \cap S \neq \emptyset} -p_c \log p_c - (1-p_c) \log p_c.\end{aligned}$$

For an *unknown* partition $\mathcal{P}$,

$$\mathbb{H}[\mathcal{Y}(S)] = \mathbb{H}[\mathcal{Y}(S) \,|\, \mathcal{P}] + \mathbb{I}[\mathcal{P}; \mathcal{Y}(S)] \tag{4.4}$$

So we can express the view value as

$$\begin{aligned}&\mathbb{H}\big[\mathcal{Y}_G(x)_{o_k=(i,j)} \mid \{\mathcal{Y}_G(x)_{o_\ell} \mid \ell < k\},\ \mathcal{Y}_G(x)_{i'j} = 1 \text{ for all } \ i' < i\big] \\ &\quad = \mathbb{H}[\mathcal{Y}_G(x)_{o_k} \,|\, \mathcal{P}, \ldots] + \mathbb{I}[\mathcal{Y}_G(x)_{o_k}; \mathcal{P} \,|\, \ldots],\end{aligned} \tag{4.5}$$

where "..." abbreviates the condititioning expression on the LHS. The first term on the RHS of 4.5 is an average, over many known partitions $\mathcal{P}$, of the coloring information gained by sampling at $x + g_{ij}$. If no preceding nodes have sampled from the same cell, the gain is $-p_c \log p_c - (1-p_c) \log p_c$. Otherwise gain is zero. So computing marginal entropy amounts to computing the probability that a sample point will be the "first to find" the cell that it lies in.

If $\mathcal{P}$ is selected from a translationally- and rotationally-invariant probability distribution, this is equal to the expectation of the reciprocal of the number of grid points sharing a cell with $x + g_{ij}$. This expectation depends only on the position of g_{ij} in the grid, and is radially symmetric, so we can write

$$\mathbb{H}[\mathcal{Y}_G(x)_{o_k} \,|\, \mathcal{P}, \ldots] = w_i(-p_c \log p_c - (1-p_c) \log p_c), \tag{4.6}$$

where

$$w_i := \mathbb{E}\left[\frac{1}{\#(G \cap \mathcal{P}(x + g_{i,1}))}\right].$$

The second term on the RHS of 4.5 can be computed directly:

$$\mathbb{I}[\mathcal{Y}_G(x)_{o_k};\, \mathcal{P}|\, \ldots] = \sum_{\mathcal{P}} \sum_{\mathcal{Y}_G(x)_{o_k}} \mathbb{P}[\mathcal{P},\, \mathcal{Y}_G(x)_{o_k} \mid \ldots] \log \frac{\mathbb{P}[\mathcal{P},\, \mathcal{Y}_G(x)_{o_k} \mid \ldots]}{\mathbb{P}[\mathcal{P} \mid \ldots]\, \mathbb{P}[\mathcal{Y}_G(x)_{o_k} \mid \ldots]}$$

Let A be the event that $x + g_{o_k}$ is the first in its cell, let B_0 and B_1 be the event that $\mathcal{Y}_G(x)_{o_k}$ equals 0 or 1, respectively. Then

$$\mathbb{P}[A \mid \ldots] = w_i,$$

$$\mathbb{P}[B_0 \mid \ldots] = \mathbb{P}[A,\, B_0 \mid \ldots] = p_c w_i,$$

$$\mathbb{P}[A^c,\, B_1 \mid \ldots] = \mathbb{P}[A^c \mid \ldots] = 1 - w_i,$$

so

$$\mathbb{I}[\mathcal{Y}_G(x)_{o_k};\, \mathcal{P} \mid \ldots] = -p_c w_i \log w_i + (1 - p_c) w_i \log \frac{1 - p_c}{1 - p_c w_i} - (1 - w_i) \log(1 - p_c w_i).$$

Putting this together,

$$E_t(x) \approx \sum_{ij} \big(\mathbb{H}[\mathcal{Y}_G(x)_{o_k} \mid \mathcal{P},\, \ldots] + \mathbb{I}[\mathcal{Y}_G(x)_{o_k};\, \mathcal{P} \mid \ldots] \big) \cdot \prod_{i' < i} \mathbb{P}_t[x + g_{i'j} \in \mathcal{A}]^{s/\,\mathrm{MFP}(x + g_{i'j})} \tag{4.7}$$

4.5.2 Ising-Type Obstacles

A good obstacle model should incorporate assumptions about the shape, density and grouping behavior of obstacles. We have considered the obvious strategy of generating several hypothetical completions of partially revealed scenes, then averaging these completed scenes pointwise.

4.5.2.1 The Ising Model

We will define the Ising Model as a probability distribution $\mathcal{I}_{\mathcal{D}}$ on the set of binary, $\{-1, 1\}$-valued images $\big\{ I : \mathcal{D} \to \{-1, 1\} \big\}$, where $\mathcal{D}$ is a 2- (or 3-) dimensional, 4- (or 6-) connected lattice $\mathcal{D}$ (as shown in Fig. 4.5.2.2).

$$P(I) \propto \exp\big(\beta(I_{ij}I_{i+1,j} + I_{ij}I_{i,j+1} + I_{ij}I_{i-1,j} + I_{ij}I_{i,j-1})\big) \tag{4.8}$$

4.5.2.2 Sampling from the Ising Model

To generate obstacle hypotheses, we run a Gibbs-sampling implementation of the subcritical Ising model (as discussed in [24]), initialized with known obstacles, and halted after m iterations:

$$\mathbb{P}_t(I_{ij}^{m+1} = 1) = \begin{cases} \frac{\exp(-\beta(H_{ij}^m - 2))}{\exp(-\beta(H_{ij}^m - 2)) + \exp(-\beta(H_{ij}^m - 2))} & x_{ij} \in \mathcal{U}^t \\ 1 & x_{ij} \in \mathcal{A}^t \\ 0 & x_{ij} \in \mathcal{O}^t \end{cases} \tag{4.9}$$

with $H_{ij}^m := I_{i+1,j}^m + I_{ij}I_{i,j+1} + I_{ij}I_{i-1,j} + I_{ij}I_{i,j-1}$, and $\mathbb{P}_t(I_{ij}^0 = 1) = 0.5$ independently.

This defines a distribution $\mathcal{I}_{\mathcal{D}}^m$ on the set $\big\{I^m : \mathcal{D} \to \{0,1\}\big\}$. Samples and statistics of this family of distributions are shown in Fig. 4.5.2.2

A good choice of the scale parameter m produces obstacles which "look like" (have similar thickness and curvature to) obstacles that the explorer is likely to see. The parameter can be estimated by estimating mean free path in the real environment (e.g. by taking the average of the initial omni-directional range measurement) and choosing m to match. Figure 4.8 details how this is done.

Since these scenes are generated using a fixed coloring probability of 0.5, we normalize the MFP parameter so that $\mathbb{P}_t(y \in \mathcal{A})$ can vary:

$$\mathbb{P}_t\big[\mathcal{Y}_G(x)_{ij} = 0 \mid \mathcal{Y}_G(x)_{i-1,j} = 1\big] \approx \exp\left(\int_{\text{line}(x+g_{i-1,j},\, x+g_{ij})} \frac{\log\big(\mathbb{P}_t(y \in \mathcal{A})\big)}{\log(0.5)\ \text{MFP}_{0.5}}\, dy\right)$$

4.5.2.3 Poisson-Disk Approximation

In order to express the viewpoint quality in Ising-type in a simple form like (4.7), we make use of certain statistical properties shared between our Ising-model scenes and randomly-colored partitions of uniformly-sized Voronoi cells. A Poisson-disc cover $\mathcal{P}^r \subseteq [-1,1]^2$ is a uniformly random sampling of points in $[-1,1]^2$ such that no two points have distance less than r, and no new point can be added without violating this property. These sample points induce a Voronoi partition $\mathcal{L}^r = \mathcal{L}(\mathcal{P}^r)$ of Ω. Random colorings ξ of these Voronoi partitions produce scenes I^r which bear a superficial resemblance to those produced by the Ising process, as seen in Figure 4.9

We have found that the Poisson-Voronoi random checkerboard is a useful surrogate for entropy computations involving Ising priors. For a given set of points $\{x_k\} \subseteq \Omega$, the entropy $\mathbb{H}[I^m(\{x_k\})]$ of sampling a random Ising scene I^m, with scale parameter m, can be estimated from the entropy $\mathbb{H}[IPV^r(\{x_k\})]$, for a random Poisson-Voronoi checkerboard. We have found a strong empirical correspondence between these joint entropies, for sampling sets of ten or fewer (estimating the distribution of the boolean random variable $I^m(\{x\})$, for a *single* sampling set $\{x\} \subseteq \Omega$ requires on the order of $2^{|\{x\}|}$ random Ising scenes - as we will see, it is much less taxing to generate the Poisson-Voronoi checkerboards). Correlations between these entropies, for a range of values of m and r, are plotted in Figure 4.13. Good matches are highlighted in pink.

4.5.2.4 Computing the Weights w_g

The statistical correspondence between Ising scenes and Voronoi checkerboards allows for the efficient computation of weights w_i. The joint entropy of samples from the random checkerboards is easily expressed as a weighted sum of the marginal entropies of the samples.

Recall that we can define

$$w_i = \mathbb{E}\left[\frac{1}{\#(G \cap \mathcal{P}^r(g))}\right]. \tag{4.10}$$

That is, G is the average of the reciprocal of the number of grid points in G that share a Voronoi cell with g, as determined by a random Poisson disk sampling (see Figure 4.9)

These weights can be computed numerically, for a given Poisson disk radius r and sampling pattern G, simply by generating many random Poisson-Voronoi partitions of the region surrounding G and averaging the reciprocal of the number of g's neighbors in G. If the sampling pattern is radially symmetric, weights only depend on radial index. Weight profiles for several Poisson disk radii r are shown in Figure 4.5.

4.6 Numerical Optimizations

4.6.1 Bit-Parallel Monte Carlo

Recall that the update probability for our Monte Carlo Ising process is

$$P(I_{ij}^{t+1} = 1 | I^t) = \frac{\exp\big(\beta(H_{ij}^t - 2)\big)}{\exp\big(\beta(H_{ij}^t - 2)\big) + \exp\big(-\beta(H_{ij}^t - 2)\big)} \tag{4.11}$$

where

$$H_{ij}^t := I_{i+1,j}^t + I_{i,j+1}^t + I_{i-1,j}^t + I_{i,j-1}^t$$

This computation, which produces a single random bit, requires a new random floating-point number to be generated each time it is run. Many such bits need to be averaged to compute the marginal obstacle probabilities. If this operation could be approximated using only bit-wise operations, we could run these operations in parallel, one for each bit in a multi-bit datatype.

4.6.1.1 Overview

We define a random process $\bar{I}_{ij}^t$ to approximate I_{ij}^t. For simplicity, $\bar{I}$ will take values in $\{0,1\}$, where I took values in $\{-1,1\}$. First, we generate a random bit $B \sim \text{Bernoulli}(0.5)$. Recursively define

$$B^0 := B, \quad \text{and} \quad B^k := B_1^{k-1} \wedge B_2^{k-1}, \tag{4.12}$$

where $B_{1,2}^{k-1} \sim B^{k-1}$ are i.i.d. copies (their generation is discussed in the next section). Thus $B^k \sim \text{Bernoulli}(2^{-2^k})$. Define

$$\bar{I}_{ij}^{t+1} := \begin{cases} B(\bar{H}_{ij}^t) & \bar{H}_{ij}^t < 2 \\ \sim B(\bar{H}_{ij}^t) & \bar{H}_{ij}^t > 2 \\ B^0 & \bar{H}_{ij}^t = 2 \end{cases}, \tag{4.13}$$

where

$$B(H) := \bigwedge_{k=1}^{K} ((\sim C_k(H)) \vee B^k), \tag{4.14}$$

$$C_k(H) := k\text{-th bit of } \bar{\beta}|H-2| + 1. \tag{4.15}$$

Here $\bar{\beta}$ must be an integer. Observe that

$$\exp_2\big(-\sum_{k=1}^{K} 2^k\, C_k(H)\big) = 2^{-\bar{\beta}|H-2|-1},$$

and thus

$$E(\bar{I}_{ij}^{t+1}) = \begin{cases} 2^{-\bar{\beta}|H-2|-1} & H_{ij}^t < 2 \\ 1 - 2^{-\bar{\beta}|H-2|-1} & H_{ij}^t > 2 \\ 0.5 & H_{ij}^t = 2 \end{cases} \tag{4.16}$$

4.6.1.2 Computing the counts C_k

The counting bits C_k are computed as with a standard binary adder. This requires $O(k)$ bitwise AND and/or XOR operations for binary sum and carry, and about $2k$ boolean variables as registers.

4.6.1.3 Generating I.I.D. Bits

equivalent to the conjunction of 2^k independent coin-flips B_i^0:

$$B^k \sim B_1^0 \wedge B_2^0 \wedge \cdots B_{2^k}^0.$$

Instead of calling the random number generator 2^k times, we generate a single N-bit ($N \geq 2^k$) integer L (for "long"), and conjoin it with k bit-shifted copies of itself. Recursively define

$$L^0 := L, \qquad L^k := (L^{k-1} \wedge (L^{k-1} \ll 2^{k-1})) \ll 3 \cdot 2^{k-1} - 2. \tag{4.17}$$

Observe that

$$\begin{aligned}
L^k &= (L^{k-1} \wedge (L^{k-1} \ll 2^{k-1})) \ll 3 \cdot 2^{k-1} - 2 \\
&= (L^{k-1} \ll 2^{k-1}) \wedge (L^{k-1} \ll 2 \cdot 2^{k-1}) \ll 2^k - 2 \\
&= (L^{k-1} \ll 2^{k-2}) \wedge (L^{k-2} \ll 2 \cdot 2^{k-2}) \\
&\qquad \wedge (L^{k-1} \ll 3 \cdot 2^{k-2}) \wedge (L^{k-2} \ll 4 \cdot 2^{k-2}) \ll 2^k - 2 \\
&\;\;\vdots \\
&= \Big(\bigwedge\nolimits_{j=0}^{2^k-1} (L^0 \ll j)\Big) \ll 2^k - 1.
\end{aligned}$$

Thus, for the i-th bit of L^k:

$$(L^k)_i = \bigwedge\nolimits_{j=2^k-1}^{2^{k+1}-2} L_{i-j}^0$$

where bit indices are taken modulo N. This is a 2^k-fold conjunction of independent random bits, so

$$P((L^k)_i = 1) = 2^{-2^k},$$

Observe, also, that the bits

$$(L^{k_1})_i = \bigwedge\nolimits_{j=2^k-1}^{2^{k+1}-2} L_{i-j}^0 \quad \text{and} \quad (L^{k_2})_i = \bigwedge\nolimits_{j=2^k-1}^{2^{k+1}-2} L_{i-j}^0$$

are independent for $k_1 \neq k_2$ (the ranges $[2^{k_1} - 1, 2^{k_1+1} - 2]$ and $[2^{k_2} - 1, 2^{k_2+1} - 2]$ do not overlap). Note, however that when indices i_1 and i_2 differ by less than 2^k, the bits $(L^k)_{i_1}$ and $(L^k)_{i_1}$ are highly correlated, especially for large k.

In modern processors, the bit-shift operation runs in constant time, so the random lookup integers $\{L^k, \ldots, L^k\}$ are generated in $O(K)$ time.

4.7 Exploration

Having the range detector measurements (the observations) available $\mathcal{Y}(x_i) \in \mathcal{Y}$ at time t_i, we estimate $p_t(x)$ for a regular sampling of points in $\mathcal{A}_t$. Our next waypoint x_{t+1} is selected as

$$x_{t+1} = \arg\max_{x \in \mathcal{A}} E_t(x) \tag{4.18}$$

Exploration terminates when $E_t(x)$ falls below a certain threshold for all points on the sample grid.

4.7.1 Performance Bounds

These bounds depend on some assumed properties of the Ising-model marginal visibility probabilities $p_t(x) = \mathbb{P}_t[x \in \mathcal{V}]$.

The result in [48], that the zero-temperature Ising model acts (in the scaling limit) as a curve-shortening flow on the level curves of binary images, suggests that the iterative low-temperature simulations that generate our hypothetical scenes will produce predictably-rounded hypothetical obstacles. In turn, it suggests that their pointwise average will have predictably-rounded level curves. Here we will assume that, for any threshold $\epsilon > 0$, there is a finite constant $K = K(\epsilon)$, independent of $\mathcal{V}_t$ and $\mathcal{O}_t$, such that the boundary curvature of the level set $\{x : p_t(x) \leq 1 - \epsilon\}$ is bounded above by K.

Now, let $\mathcal{V}_t^r$ be the set of points in $\mathcal{V}_t$ that lie at distance r or greater from points in $\mathcal{O}_t$, and let $\mathcal{V}_t^r(x_0)$ be the path-component of $\mathcal{V}_t^r$ containing x_0. Next, we make a sort of equicontinuity assumption:

Conjecture 4.7.1. For p sufficiently close to 1, and radius $r > 0$, there is a second radius $0 < r' \leq r$, independent of $\mathcal{V}_t$ and $\mathcal{O}_t$, such that

$$\begin{aligned}\big(x \in \mathcal{V}_t^r \wedge p_t(x) > p \ \wedge \ d(x,z) < r'\big) \\ \text{implies that} \quad p_t(z) \geq \tfrac{1}{2}.\end{aligned}$$

This would guarantee all points sufficiently far from known obstacles, with sufficiently high marginal probability, lie within disks of high marginal probability. These conjectures, if true, imply the following:

Lemma 4.7.1. Given a probability threshold $p_{\text{thresh}} > 0$, and radius r, there is an energy threshold $E_{\text{thresh}} > 0$ such that any point $x \in \mathcal{V}_t^r$ with $p_t(x) > p_{\text{thresh}}$ will induce an energy $E_t(y) > E_{\text{thresh}}$ at a nearby point y.

Proof: Let r' be as described in the conjecture, where r and p_{thresh} take the place of r and p. Then let

$$\begin{aligned}E_{\text{thresh}} = 2^{-\lambda r'/2}\pi \min\left\{K(p_{\text{thresh}})^{-2}, \tfrac{1}{4}r'^2\right\} \\ \cdot\left(-\log p_{\text{thresh}}\right).\end{aligned}$$

Each of the three terms on the RHS is a lower bound of the corresponding term in the energy calculation, centered at x. The first bounds the area of the intersection between the level set $\{x : p_t(x) > p_{\text{thresh}}\}$ and the disk of radius r', centered at x. For points z in this region, $\frac{1}{2} \leq p_t(z) \leq p_{\text{thresh}}$, and so

$$2^{-\lambda r'/2} \leq \exp\Big(\lambda \int_0^{r'} \log p_t(z(s))\, ds\Big)$$

and

$$-\log p_{\text{thresh}} \leq -\log p_t(z) - (1 - p_t^{\lambda\beta s}(z))\log(1 - p_t(z)).$$

□

Now, define

$$f(p,\beta,\lambda) = -\log(p_t(z)) - (1 - p_t^{\lambda\beta s}(z))\log(1 - p_t(z))$$

$$M(\beta,\lambda) = \sup_{p:[0,\infty)\to[0,1]} \int_{t=0}^{\infty} f(p(t),\beta,\lambda) \exp\Big(\lambda \int_{u=0}^{t} \log p(u)\, du\Big)\, dt. \tag{4.19}$$

A restarting argument can be used to show that (4.19) is maximised by a constant function $p(t) = a$, so

$$M(\beta,\lambda) = \sup_a \; f(a,\beta,\lambda) \int_{t=0}^{\infty} \exp(\lambda t \log a)\, dt \tag{4.20}$$

$$= \sup_a \; \frac{f(a,\beta,\lambda)}{\lambda \log a}. \tag{4.21}$$

The last expression is continuous with respect to a, and bounded as $a \to 0$ and $a \to 1$, so $M(\beta,\lambda)$ does indeed exist.

Lemma 4.7.2. If an explorer has visited $y \in \mathcal{V}_t^r(x_0)$ before time t, then $E_t(y') < E_{\text{thresh}}$ for all points y' within distance $\delta = \delta(E_{\text{thresh}}, r, \beta, \lambda)$ of y, where $\delta = r\sin(\alpha/n_{\text{sh}})$, $\alpha = E_{\text{thresh}}/M(\beta,\lambda)$, and n_{sh} is the maximum number of disjoint shadow boundaries cast from a single point.

Proof: Traveling a distance δ in any direction will reveal a sliver of at most α along any shadow boundary. The maximum energy of such a vantage point is thus $n_{\text{sh}}\,\alpha\, M(\beta,\lambda)$.

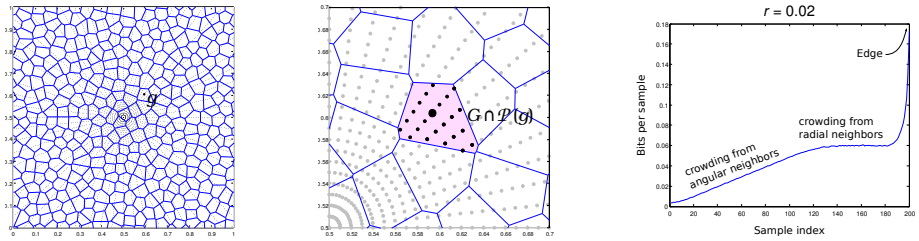
□

Combining these two lemmas we get

Proposition 4.7.1. If A is the area of $\mathcal{V}^r(x_0)$, then, the greedy algorithm will take at most $A/\pi\delta(E_{\text{thresh}}r,\beta,\lambda)^2$ steps to bring $p_t(x)$ below p_{thresh}, for all points $x \in V^r(x_0)$.

4.7.2 Exploration Results

Using the exploration testbed provided by the authors of [84], we were able to compare our algorithm with theirs, in terms of (1) number of planning steps (2) total distance traveled by the explorer (3) unexplored area vs. total distance traveled, and (4) unexplored area at termination. Results are shown in Figure 4.17.

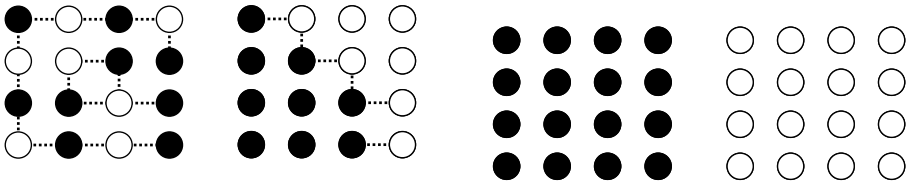


(a) Sampling Grid over random partition

(b) Counting neighbors

(c) Weight Profile

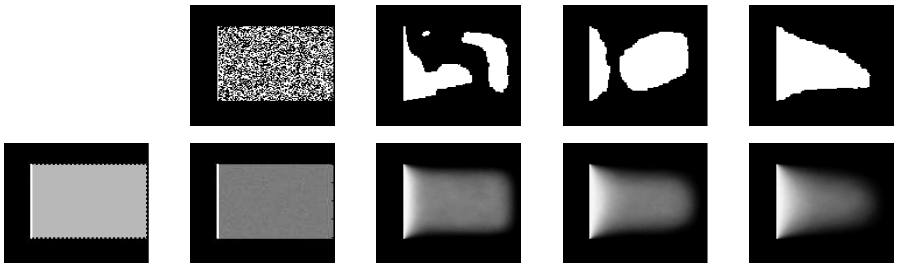
Figure 4.5: *Computing Sample Weights for a Radial Sampling Grid* **(a)** The weight w_{ij} is the expected energy gain from a sample at position $x + g_{ij}$ in $G(x)$. **(b)** The entropy of a sample from a random coloring of a random partition is taken as the expected value of the reciprocal of the number of samples with which it shares a cell (argument detailed in 4.5.2.4). **(c)** Since the sampling pattern G is radially symmetric, the weights w_{ij} can be described by a radial "profile". The weight profile can be divided into three phases: The first phase, when nodes are likely to share cells with angular neighbors, shows a increase in weight as radius increases and radial neighbors get further and further away. The second phase, when angular neighbors are too far away to share cells, has constant weight. The third phase, as we reach the edge of the sampling pattern, shows a rapid doubling of weight, as nodes with many inner and outer (smaller or greater radius) neighbors give way to nodes with few or no outer neighbors.



(a) High / Low Energy Configurations

(b) Minimum Energy Configurations

Figure 4.6: *Ising Model on a* 4×4 *Lattice* Likelihood under the Ising Model decreases exponentially with the number of edges between disagreeing nodes. Samples from the Ising Model are concentrated near the two minimum energy configurations: all-white or all-black.



(a) Visibility (b) 0 Iterations (c) 100 Iter. (d) 200 Iter. (e) 400 Iter.

Figure 4.7: *Ising-type Obstacle Posterior* (a) The random process described in 4.5.2.2 is run on a scene with a rectangular shadow region (gray) produced by a flat obstacle edge (white). Black denotes known free space. (b)-(e) Top row: Images produced by the process. Bottom row: Average of 1000 runs.

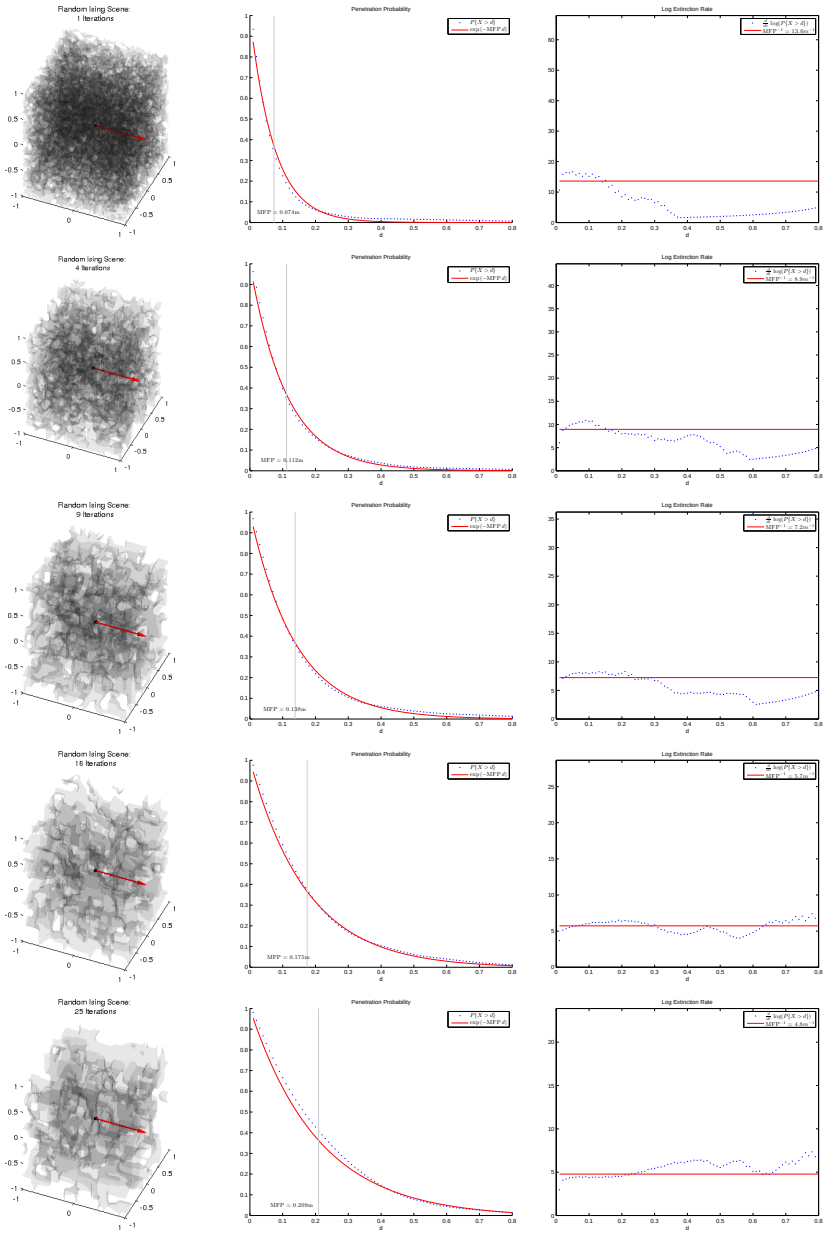
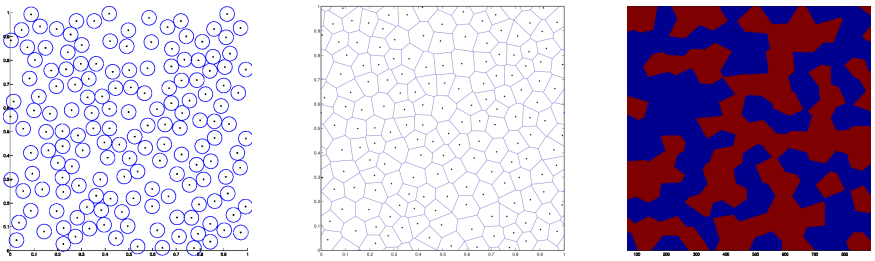
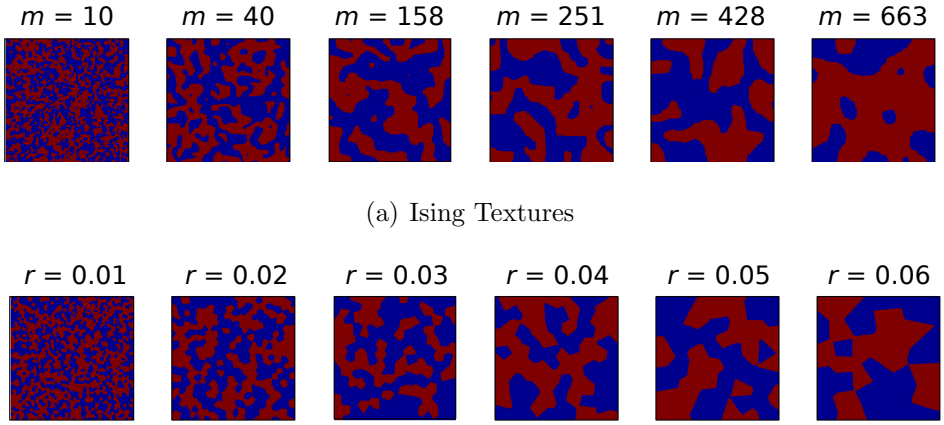


Figure 4.8: *Extinction Profiles* Here we show extinction profiles for Ising scenes with various granularities, indexed by number of update iterations. Instantaneous mean free path (derivative of log survival probability) and best-fit MFP are plotted together, showing that extinction in these scenes is well-modeled by an exponential decay.



(a) Poisson-Disk Sampling (b) Voronoi Partition (c) Checkerboard

Figure 4.9: *Generating Poisson-Voronoi Checkerboard* (a) To produce a scene of desired granulation, we first produce a Poisson-Disc sampling of the domain, which gives a maximal set of points satisfying the property that the distance between two points is no less than r, a granularity parameter. (2) We compute a Voronoi partition of the scene based on that sampling (3) We color the cells independently, denoting them as obstacle or free space, with probability 1/2.



(a) Ising Textures

(b) Corresponding Voronoi Checkerboards

Figure 4.10: *Poisson-Voronoi Checkerboard* There is a superficial resemblance Ising-type obstacles and Voronoi checkerboards with appropriate scale parameters.

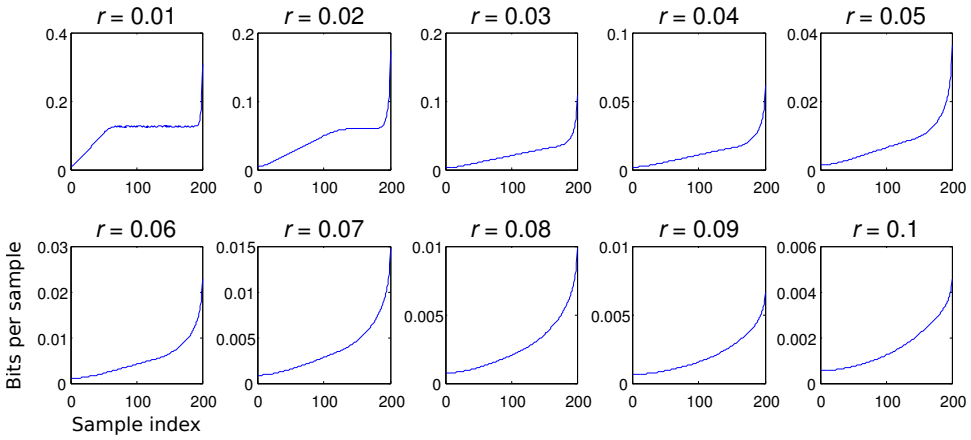
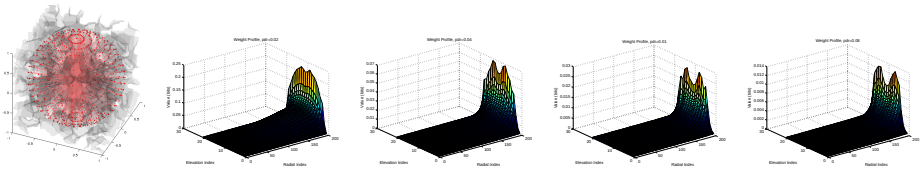


Figure 4.11: *2D Weight Profiles for Various r* Using the counting technique described in 4.5.1.2, we compute the information contribution, in bits, of samples at different (radial) positions on the radial sampling pattern (See Fig. 4.3), with regard to Poisson-Disc Voronoi scenes of various scales. In this case, we have 30 angular divisions and 200 radial divisions within a maximum radius of one unit.



(a) Pattern (b) r=0.02 (c) r=0.04 (d) r=0.08 (e) r=0.06

Figure 4.12: *3D Weight Profiles for Various r* The counting technique described in 4.5.1.2 is applied to 3D Poisson-Disc Voronoi Scenes, drawing from a spherical sampling pattern (a) with 30 azimuthal, 200 radial, and 11 elevation divisions. Again, radial symmetry around the vertical axis allows us to ignore azimuthal index.

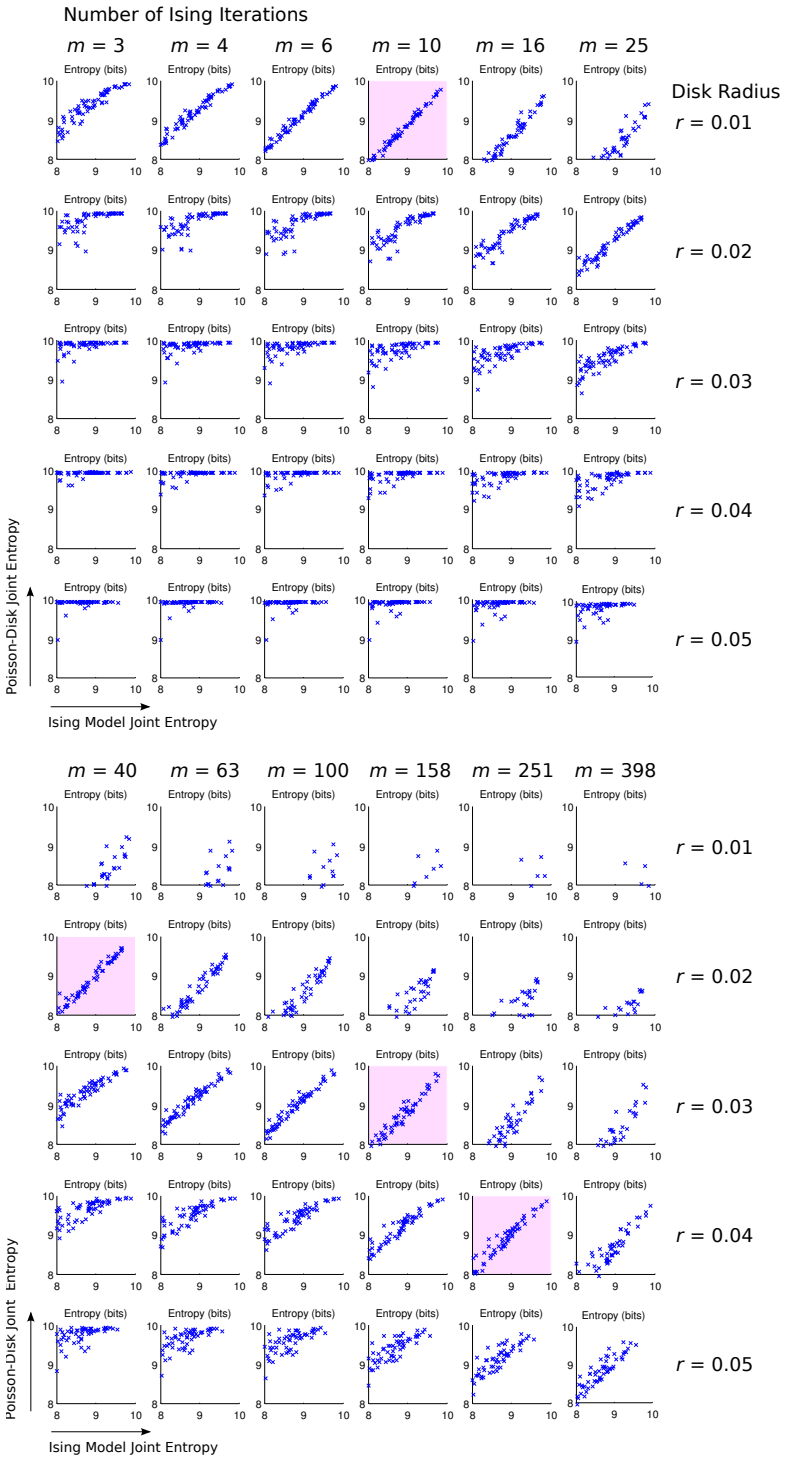


Figure 4.13: *Entropy Comparisons*

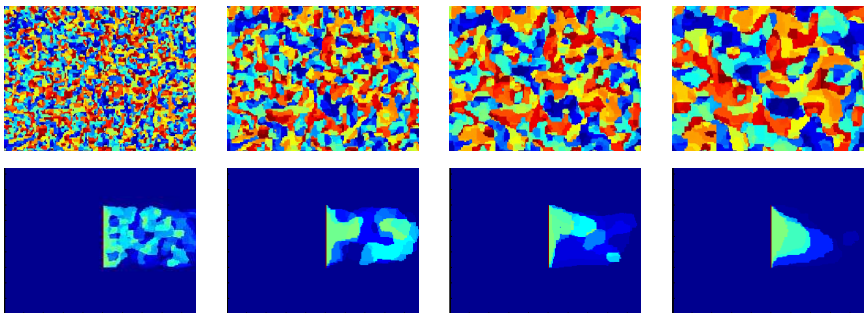


Figure 4.14: *Bit-Parallel Computation of Ising Marginal* "Bit twiddling" allows us to compute 64 simultaneous realizations of the Ising process on a single array of long integers. In the plots above, color is assigned according to nominal integer value, i.e. each realization contributes according to the significance of its bit position. This is the most direct way of visualizing the data, and the exponential decay of bit significance gives a sense of transparency and depth. Cf. Fig. 4.5.2.2

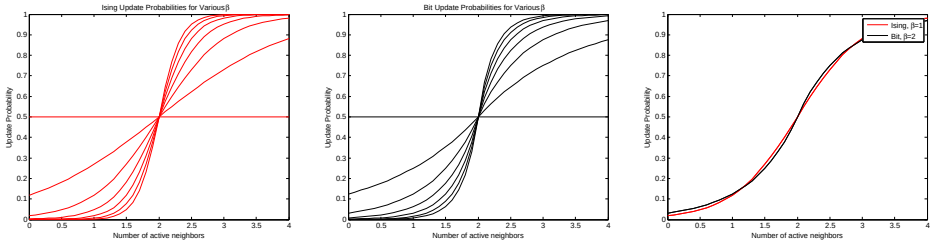
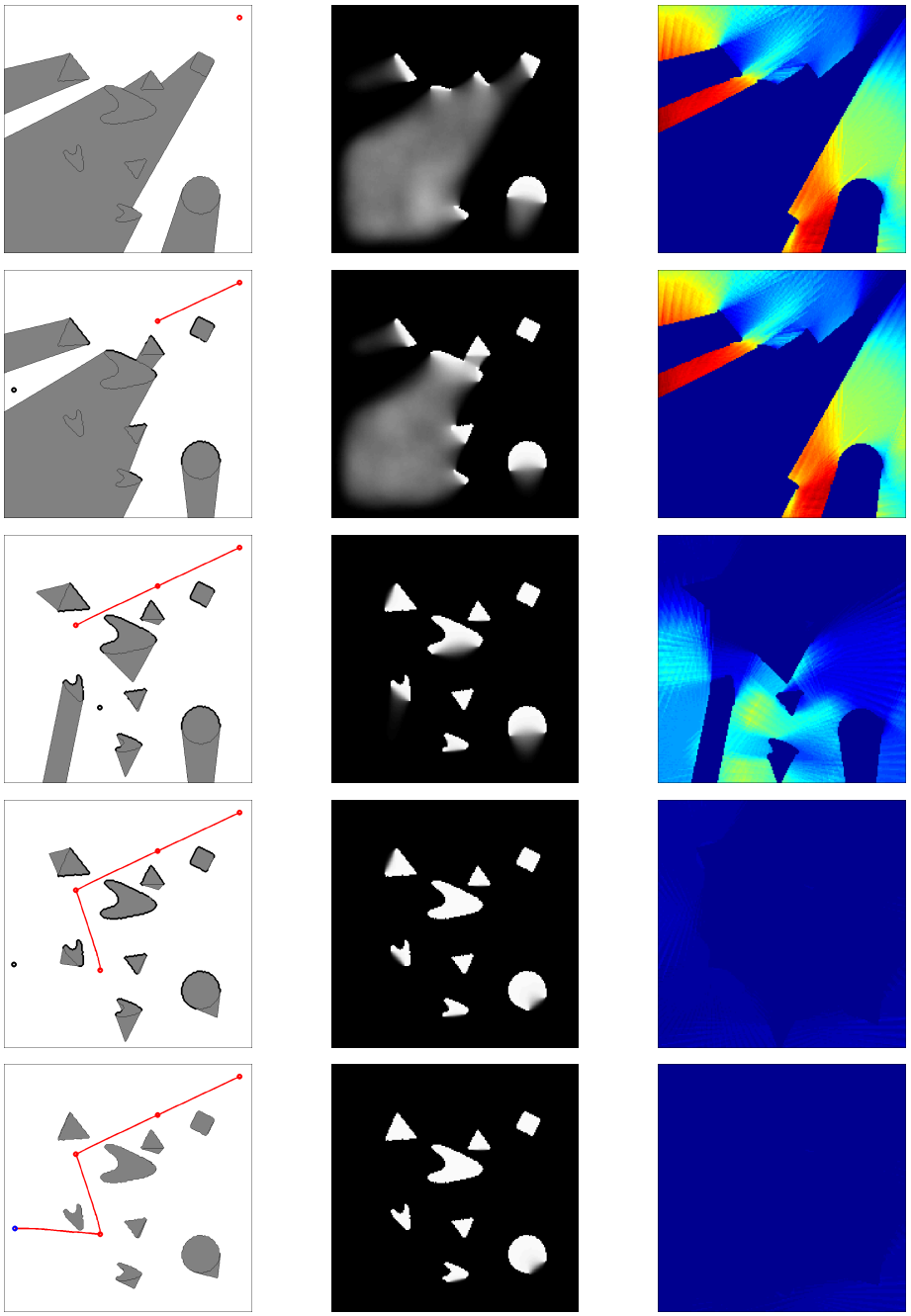
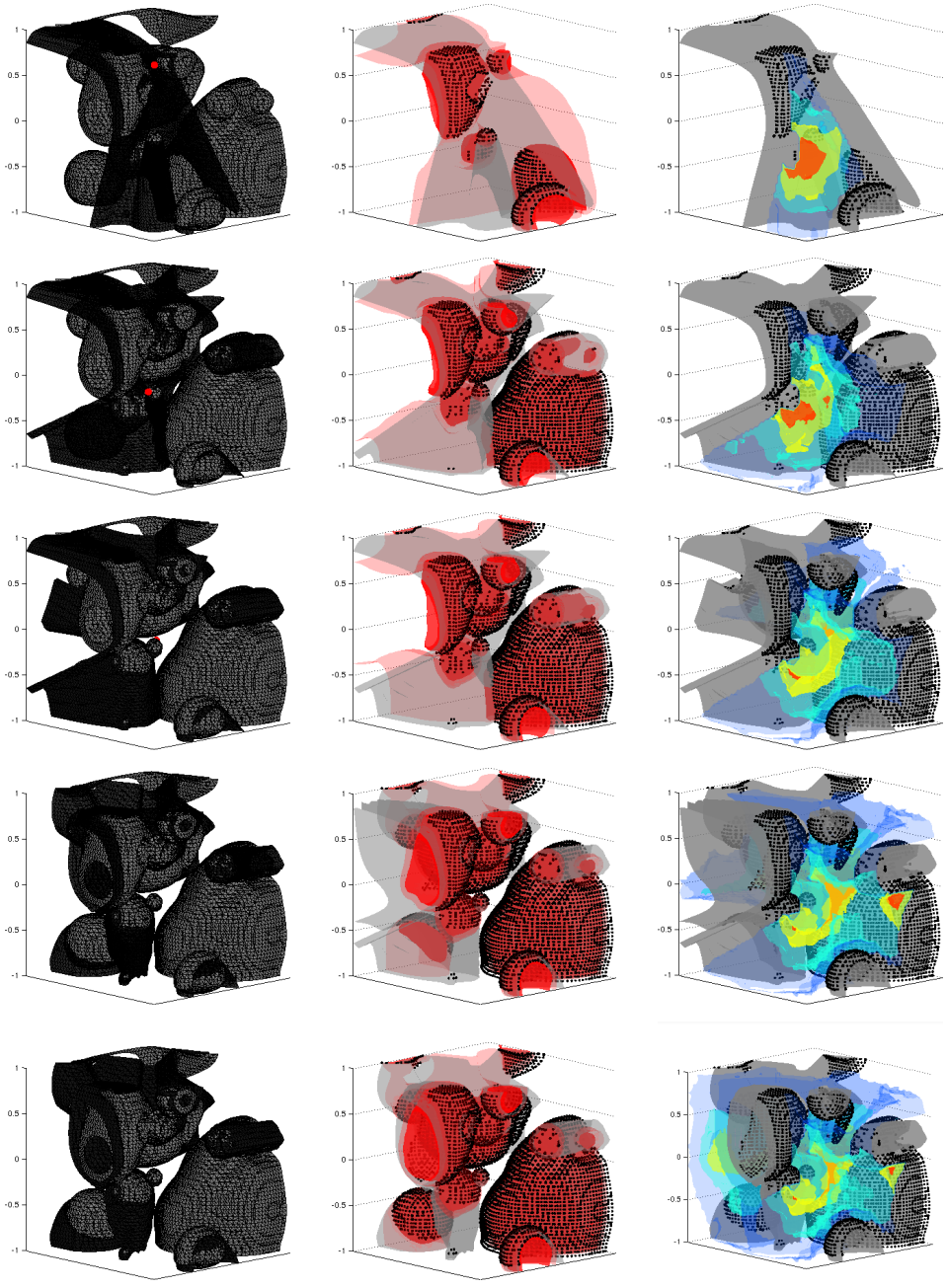


Figure 4.15: *Comparison of Update Probabilities:* Values of $\bar{\beta}$ are taken from 0 to 6. Update probabilities for the Ising Model and the parallel bit estimate are computed on the continuum $[0, 4]$, although only integers $\{0, \ldots 4\}$ are used in practice.

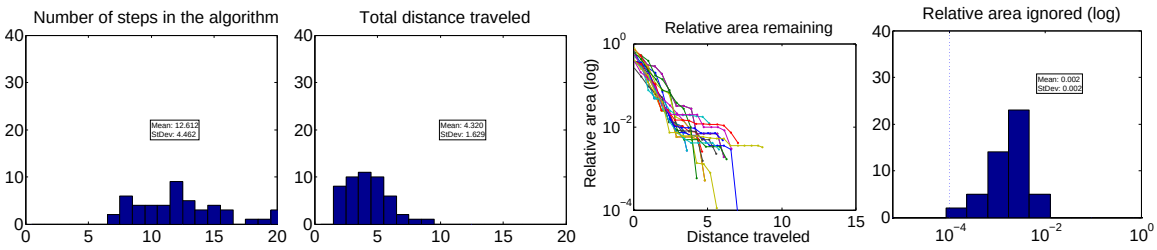


(a) Visibility (b) Marginal Likelihood (c) Viewpoint Quality

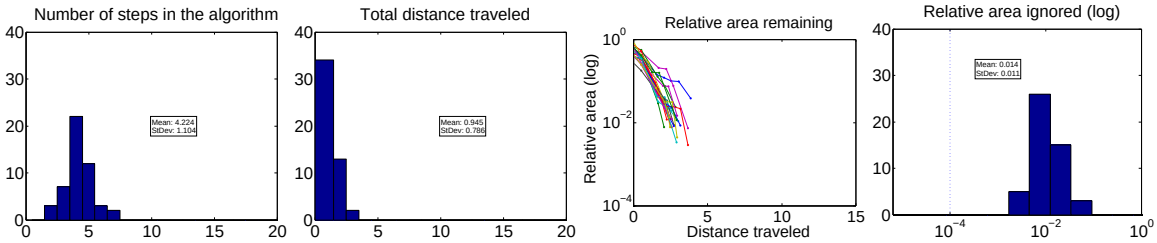
Figure 4.16: *Exploration of a 2D Scene* An explorer (red dot) attempts to efficiently map an unknown scene (a) by greedily choosing informative viewpoints (c). Observe that we are not merely uncovering area. Once the disposition of a site has been determined with high confidence (b), that site loses its informative value.



(a) Visibility (b) Marginal Likelihood (c) Viewpoint Quality



(d) Exploration Stats, Valente et al.



(e) Exploration Stats, Ising Marginals

Figure 4.17: Exploration comparison with Valente et al.

CHAPTER 5

Desiging Agents with Task-Specific Minimal Representation

5.1 Introduction

We are interested in obtaining agents that perform optimally on a task involving uncertainty while being as simple as possible, as measured by the size of their internal representation. While this is a problem relevant in many fields, our motivation comes from the fields of robotics and computer vision. Embodied agents have sensors providing very high-dimensional data; the worlds state is similarly high dimensional. Representing the agents belief about the world is complicated enough, and often inference is intractable. However, it has been observed that, for certain tasks, it is not necessary to represent and plan using the entire belief. Therefore, it is interesting to understand whether it is possible to solve a planning problem using a smaller representation than the entire belief. We will show that it is indeed possible and we will provide a constructive algorithm to obtain representation-minimal agents.

5.1.1 Previous Work

The topic of reduction of logical functions has been extensively studied. The formalism of finite state machines was developed in the 1950s and 60s, and by 1971, Hopcroft [32] published an $(n \log n)$-time algorithm for reducing completely-specified FSMs. Incompletely-Specified FSMs were studied in turn, and various

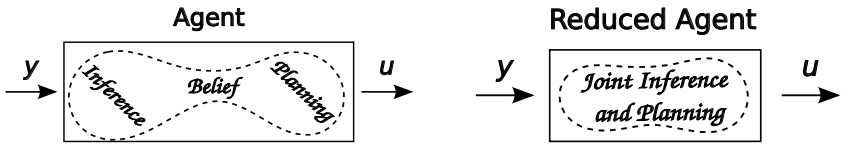


Figure 5.1: *Representation Reduction* Canonical belief spaces can become a serious informational bottleneck between inference and planning modules. An agent's understanding of its environment need not be any richer than necessary to support the task at hand.

heuristics were developed to quickly approximate minimum[1] reductions. [65, 27, 25]. Representation reduction in artificial intelligence has been dealt with in various guises: Dimensionality reduction, belief space compression etc. – most heuristics can be considered as implicit hard-coded representation reductions.

5.1.2 Contributions

This chapter re-casts problem of representation reduction in terms of well-studied computational constructs, and finds absolute minimum reductions in the case of discrete time, discrete input and output. It evaluates three algorithms for approximating minimum reductions, and clarifies their strengths and weaknesses. This work is the result of a collaboration of Andrea Censi, who introduced the formalism of representation reduction in the context of POMDP solvers and proposed the bit-at-a-time algorithm

[1]Here, we distinguish "minimal" from "minimum" reductions. A minimal reduction is one that cannot be further reduced by combining any of its states, whereas a minimum reduction is a minimal reduction with the fewest possible states

5.2 Formalization

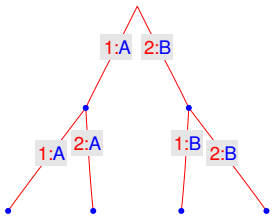
Example 5.2.1 (Equivalent Decision Tables)**.** Suppose $\mathcal{Y} = \{1, 2\}$, $\mathcal{U} = \{A, B\}$, $\mathcal{C} = \bigcup_{i=1}^{2} \mathcal{Y}^i$ and

$$\mathcal{T}(c) = \begin{cases} A & c \in \{(1), (1,2)\} \\ B & c \in \{(2), (2,1)\} \\ \mathcal{T}'(c) & \text{otherwise.} \end{cases},$$

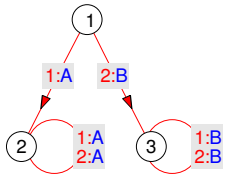
is an optimal policy, for arbitrary $\mathcal{T}' : \mathcal{C} \to \mathcal{U}$. However, depending on the choice of $\mathcal{T}'$ (highlighted in the tables below), the completed policy can have differently-sized minimal representations:

$\mathcal{T} : \mathcal{C} \to \mathcal{U}$	
(1)	$\mapsto A$
(2)	$\mapsto B$
$(1,1)$	$\mapsto A$
$(1,2)$	$\mapsto A$
$(2,1)$	$\mapsto B$
$(2,2)$	$\mapsto B$

(a) Decision Table



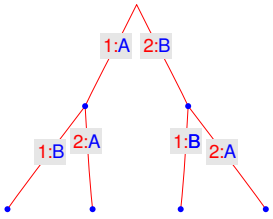
(b) Decision Tree



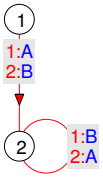
(c) Minimal Policy

$\mathcal{T} : \mathcal{C} \to \mathcal{U}$	
(1)	$\mapsto A$
(2)	$\mapsto B$
$(1,1)$	$\mapsto B$
$(1,2)$	$\mapsto A$
$(2,1)$	$\mapsto B$
$(2,2)$	$\mapsto A$

(d) Decision Table



(e) Decision Tree



(f) Minimal Policy

Instead, we propose an "incompletely-determined" formalization:

Definition 5.2.1 (Policies)**.** Given a set $\mathcal{Y}$ of observations, recursively construct

$$\mathcal{C}_0 = \{\emptyset\} \quad \text{and} \quad \mathcal{C}_{i+1} = \{(c, y_{i+1}) : c \in \mathcal{C}_i,\ y_{i+1} \in \mathcal{Y}_c\}, \tag{5.1}$$

where $\mathcal{Y}_c \subseteq \mathcal{Y}$ are the observations that may be seen in context c. Let $\mathcal{C} = \bigcup_{i=0}^{\infty} \mathcal{C}_i$. A **policy** P is then a tuple $\langle \mathcal{C}, \mathcal{U}, \mathcal{T}, \mathcal{Y} \rangle$, where $\mathcal{U}$ is some decision set and $\mathcal{T} : \mathcal{C} \setminus \{\emptyset\} \to \mathcal{U}$.

Definition 5.2.2 (Completely-Determined Policies)**.** If $\mathcal{Y} = \bigcup_{\mathcal{C}} \mathcal{Y}_c$ and $\mathcal{C} = \mathcal{Y}^{\leq n}$ for some $n \in \mathbb{N}$, then $P = \langle \mathcal{C}, \mathcal{U}, \mathcal{T}, \mathcal{Y} \rangle$ is **completely determined**. A **completion** of P is a policy $P' = \langle \mathcal{C}', \mathcal{U}', \mathcal{T}', \mathcal{Y} \rangle$ such that

$$\mathcal{Y} \subseteq \mathcal{Y}', \quad \mathcal{C}' = \bigcup_{i=0}^{\infty} (\mathcal{Y}')^i, \quad \mathcal{U} \subseteq \mathcal{U}', \quad \text{and} \quad \mathcal{T}'|_{\mathcal{C}} = \mathcal{T}. \tag{5.2}$$

Let $\mathrm{Comp}(P)$ be the set of completions of the policy P.

Definition 5.2.3 (FSM Representations)**.** An **FSM representation** (or just **representation**) is a tuple $\langle \mathcal{C}, \mathcal{R}, \mathcal{U}, \mathcal{S}, \mathcal{T}, \mathcal{Y} \rangle$ (abbreviated to $\langle \mathcal{R}, \mathcal{S} \rangle$ when $P = \langle \mathcal{C}, \mathcal{U}, \mathcal{T}, \mathcal{Y} \rangle$ is given), with "states" $\mathcal{S} \subseteq \mathbb{N}$ and state assignments $\mathcal{R} : \mathcal{C} \to \mathcal{S}$, such that

$$\mathcal{R}(c) = \mathcal{R}(c') \quad \text{and} \quad y \in \mathcal{Y}_c \cap \mathcal{Y}_{c'} \implies \mathcal{T}(c, y) = \mathcal{T}(c', y). \tag{5.3}$$

Let $\mathrm{Rep}(P)$ be the set of representations of the policy P.

Definition 5.2.4 (Minimal Representations)**.** The **size** of an FSM representation is the cardinality of its state set. A representation $\langle \mathcal{R}, \mathcal{S} \rangle$ of P is **minimal** if $|\mathcal{S}| = \min\{|\mathcal{S}'| : \langle \mathcal{R}', \mathcal{S}' \rangle \in \mathrm{Rep}(P)\}$. A representation $\langle \mathcal{R}', \mathcal{S}' \rangle$ is a **reduction** of the representation $\langle \mathcal{R}, \mathcal{S} \rangle$ if there is a surjection $\phi : \mathcal{S} \to \mathcal{S}'$ such that $\mathcal{R}' = \phi(\mathcal{R})$.

Example 5.2.2. If $\mathcal{C} = \{c_1, c_2, \ldots\}$, then we have a canonical representation $\langle \mathcal{R}, \mathcal{S} \rangle$, where

$$\mathcal{S} = \{1, \ldots, |\mathcal{C}|\} \quad \text{and} \quad \mathcal{R} : c_k \mapsto k. \tag{5.4}$$

It can be shown that the size of a minimal representation of a policy P is equal to the minimum size of the minimal representations of its completions, i.e.

$$\min\{|\mathcal{S}'| \,:\, \langle \mathcal{R}', \mathcal{S}' \rangle \in \mathrm{Rep}(P)\} = \min\{|\mathcal{S}'| \,:\, \langle \mathcal{R}', \mathcal{S}' \rangle \in \mathrm{Rep}(P'),\ P' \in \mathrm{Comp}(P)\} \tag{5.5}$$

Incompletely-determined policies allow more freedom in representation reduction, as shown in the next example.

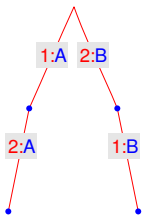
Example 5.2.3 (Incompletely-Determined Policies)**.** Let $\mathcal{C} = \{\emptyset, (1), (2), (1,2), (2,1)\}$, $\mathcal{U} = \{A, B\}$, and

$$\mathcal{T}(c, y) = \begin{cases} A & c \in \{(1), (1,2)\} \\ B & c \in \{(2), (2,1)\} \end{cases}.$$

Observe that the minimal policy is the same as that of the completely-determined policy in Example 5.2(f).

$\mathcal{T} : \mathcal{C} \setminus \{\emptyset\} \to \mathcal{U}$	
(1)	$\mapsto A$
(2)	$\mapsto B$
$(1,2)$	$\mapsto A$
$(2,1)$	$\mapsto B$

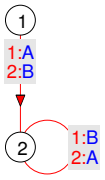
(a) Decision Table



(b) Decision Tree

$\mathcal{R} : \mathcal{C} \to \mathcal{S}$	
$\emptyset$	$\mapsto 1$
(1)	$\mapsto 2$
(2)	$\mapsto 2$
$(1,2)$	$\mapsto 2$
$(2,1)$	$\mapsto 2$

(c) Representation



(d) Minimal Policy

5.2.1 FSM Reduction

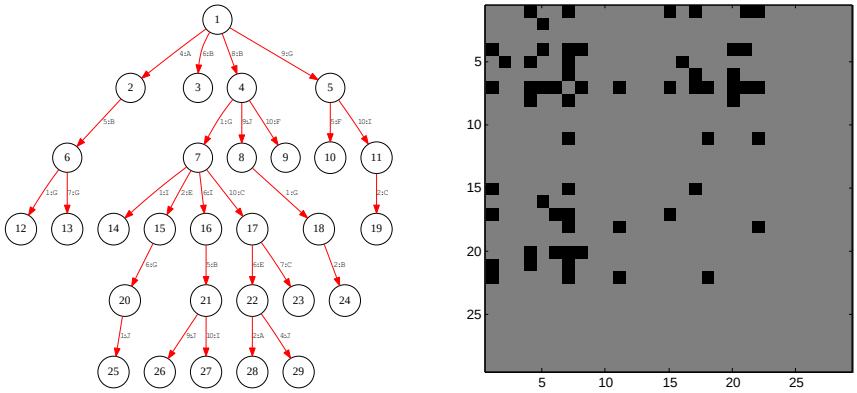
Given a decision policy (or an ISFSM) how do we find an obedient (or equivalent) ISFSM with the smallest possible state set?

for completely-specified FSM, this can be done in $n \log n$ time by Hopcroft's algorithm (Alg. 5.3.1).

To find a minimum representation of a given policy, we first compute a graph of reducibility relations, then compute a minimal clique-covering.

Cliques on the equivalence graph identify sets of states that can be collapsed into a single state. The minimal clique-covering, that is the smallest collection of disjoint cliques that covers the equivalence graph, correponds to a minimal reduction of the FSM.

5.3 Representation Reduction Strategies



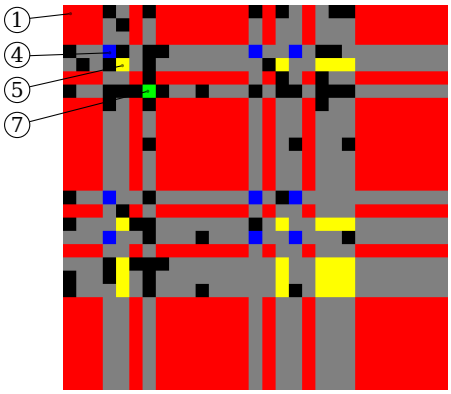
(a) Canonical Representation (b) Compatibility Matrix

For practical computation of reducibility, we'll start with the weaker condition of compatibility.

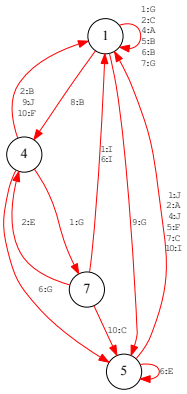
5.3.1 Reducibility Relations

Definition 5.3.1 (Reducibility)**.** For a given policy $P = \langle \mathcal{C}, \mathcal{U}, \mathcal{T}, \mathcal{Y} \rangle$, two contexts $c_1, c_2 \in \mathcal{C}$ are **reducible** (write $c_1 \sim c_2$) if there exists a representation $\langle \mathcal{R}, \mathcal{S} \rangle$ of P such that $\mathcal{R}(c_1) = \mathcal{R}(c_2)$. Likewise, for a given representation $R = \langle \mathcal{R}, \mathcal{S} \rangle$, two states $s_1, s_2 \in \mathcal{S}$ are **reducible** if there exists a reduction $(\phi, \langle \mathcal{R}', \mathcal{S}' \rangle)$ of R such that $\phi(s_1) = \phi(s_2)$.

Observe that for any representation $\langle \mathcal{C}, \mathcal{R}, \mathcal{U}, \mathcal{S}, \mathcal{T}, \mathcal{Y} \rangle$, the contexts $c_1, c_2 \in \mathcal{C}$



(c) Greedy Clique Covering of 5.4(b)



(d) Reduced Representation

Figure 5.4: *Greedy Reduction Algorithm* A "running clique" is kept, to which new states are added until all remaining states are incompatible with at least one state in the clique. Then, a new clique is begun. Here, the cliques are $\{1, 2, 3, 6, 8, 9, 10, 11, 12, 13, 14, 16, 19, 22, 26, 27, 28, 29, 30, 31, 32\}$ (red), $\{4, 15, 18\}$ (blue), $\{5, 17, 21, 22, 23\}$ (yellow), and $\{7\}$ (green). (d) shows the resulting policy graph once

are reducible if and only if the states $\mathcal{R}(c_1)$ and $\mathcal{R}(c_2)$ are reducible. Observe also that for incompletely-determined policies, reducibility is a symmetric but not-necessarily-transitive relation

Example 5.3.1 (Non-Transitive Reducibility)**.** Suppose $\mathcal{Y} = \{1, 2, 3\}$, $\mathcal{C} = \{\emptyset, (1), (2), (1, 3), (2, 3)\}$, $\mathcal{U} = \{A, B\}$, and

$$\mathcal{T}(c) = \begin{cases} A & c \in \{(1), (1, 3)\} \\ B & c \in \{(2), (2, 3)\} \end{cases}. \tag{5.6}$$

Observe that, under this policy, $\emptyset \sim (1)$ and $\emptyset \sim (2)$, but $(1) \not\sim (2)$, since $\mathcal{T}(1, 3) \neq \mathcal{T}(2, 3)$.

However, it can be shown that, under a completely-determined policy, reducibility induces an equivalence relation. In either case, we compute reducibility using the following criterion:

Lemma 5.3.1. Two contexts $c_1, c_2 \in \mathcal{C}$ are reducible iff

$$\mathcal{T}(c_1, s) = \mathcal{T}(c_2, s) \quad \text{for all} \quad s \in \mathcal{Y}^* \quad \text{such that} \quad (c_1, s), (c_2, s) \in \mathcal{C} \tag{5.7}$$

This informs the following algorithm

Algorithm 1: (Hopcroft) Compute Reducibility Relations

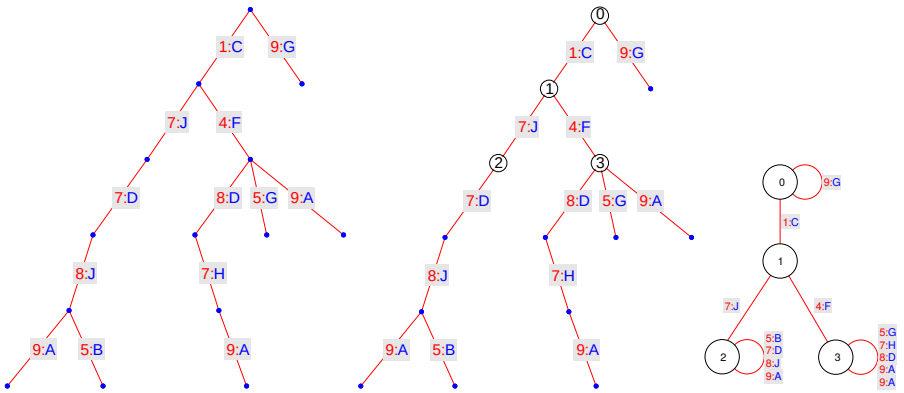
Input: A representation $\langle \mathcal{C}, \mathcal{R}, \mathcal{U}, \mathcal{S}, \mathcal{T}, \mathcal{Y} \rangle$

Output: A reducibility matrix $A : \mathcal{S} \times \mathcal{S} \to \{\textbf{true}, \textbf{false}\}$.

$A(s_1, s_2) \Leftarrow \textbf{true}$ for all $s_1, s_2 \in \mathcal{S}$.
repeat
  $isChanged \Leftarrow \textbf{false}$
  for $s_1 < s_2 \in \mathcal{S}$ **do**
    if $A(s_1, s_2) = \textbf{true}$ **then**
      for $c_1 \in \mathcal{R}^{-1}(s_1)$, $c_2 \in \mathcal{R}^{-1}(s_2)$ **do**
        for $y \in \mathcal{Y}_{c_1} \cap \mathcal{Y}_{c_2}$ **do**
          if $\mathcal{T}(c_1, y) \neq \mathcal{T}(c_2, y)$ or $\sim A(\mathcal{R}(c_1, y), \mathcal{R}(c_2, y))$ **then**
            $A(s_1, s_2) \Leftarrow \textbf{false}$.
            $isChanged \Leftarrow \textbf{true}$.
          end if
        end for
      end for
    end if
  end for
until $\sim isChanged$

5.3.2 Bit-at-a-Time

The Bit-at-a-Time reduction, proposed by Andrea Censi, generates a set of states one at a time, separating ambiguous contexts recursively. An ambiguous context



(a) Decision Tree (b) Partitioned Tree (c) Reduced FSM

Figure 5.5: *Censi's Bit-at-a-time Algorithm*

is an (input, state) pair for which more than one output is defined. Ambiguities which arise in shorter sequences are separated first. This results in an FSM whose graph structure is a subtree of the original decision tree.

5.3.3 Greedy Covering

Although reducibility is not an equivalence relation, any reduction $\phi : \mathcal{S} \to \mathcal{S}'$ induces an equivalence relation, partitioning $\mathcal{S}$ into cliques of mutually-reducible states, i.e.

$$\mathcal{S} = \bigsqcup_{s' \in \mathcal{S}'} \phi^{-1}(s'), \quad \text{where} \quad \phi(s_1) = \phi(s_2) \implies A(s_1, s_2) \tag{5.8}$$

Thus, a minimum reduction induces a minimum clique partition of the reducible states of a representation.

5.3.4 Assembling Cliques

Notation 1 (Arrow notation)**.** For a policy $P = \langle \mathcal{C}, \mathcal{U}, \mathcal{T}, \mathcal{Y} \rangle$, write $c \to c'$ if $c = (c_1, \ldots, c_i) \in \mathcal{C}_i \subseteq \mathcal{C}$ and $c' = (c_1, \ldots, c_i, y) \in \mathcal{C}_{i+1} \subseteq \mathcal{C}$, for some i. For a representation $\langle \mathcal{R}, \mathcal{S} \rangle$ of P, write $s_1 \to s_2$ if there are $c_1 \in f^{-1}(s_1)$ and $c_2 \in f^{-1}(s_2)$

such that $c_1 \to c_2$.

We propose the following, greedy, approximate algorithm In order find the

---

Algorithm 2: Greedy Clique Covering

Input: A representation $\langle \mathcal{C}, \mathcal{R}, \mathcal{U}, \mathcal{S}, \mathcal{T}, \mathcal{Y} \rangle$ with $s_1 < s_2$ only if $s_2 \not\to s_1$.

Input: A reducibility matrix $A : \mathcal{S} \times \mathcal{S} \to \{\textbf{true}, \textbf{false}\}$ as computed by Algorithm 1.

Output: A partition function $\phi : \mathcal{S} \to \mathcal{S}'$ with $\phi(s_1) = \phi(s_2)$ only if $A(s_1, s_2)$.

$\mathcal{S}' \Leftarrow \mathcal{S}$

$\phi \Leftarrow id_{\mathcal{S}}$

$unused \Leftarrow \mathcal{S}$

while $|unused| > 0$ **do**

  $s_1 \Leftarrow \min(unused)$

  $unused \Leftarrow unused \setminus \{s_1\}$

  for $s_2 \in unused$ **do**

    if $A(s_1, s_2)$ **then**

      $\phi(s_2) \Leftarrow s_1$

      $unused \Leftarrow unused \setminus \{s_2\}$

    end if

  end for

end while

---

absolute minimum representation of a given policy, it suffices to run the greedy algorithm on all possible orderings of its states: Given a minimum clique covering $S = \{s_{c11}, s_{c12}, \ldots, s_{c1N_1}\} \cup \{s_{c21}, s_{c22}, \ldots, s_{c2N_2}\} \cup \cdots \cup \{s_{cK1}, s_{cK2}, \ldots, s_{cKN_K}\}$, feed states to the greedy algorithm in the order in which they are written. Failure to add states to a running clique will occur only once per clique in the minimal covering (exactly K times) [2], so the greedy algorithm will produce a minimum

[2] If more than K times, then some running clique

covering. Of course, exhaustively checking every ordering will end up taking exponential time. In fact, it has been shown that the minimum clique cover problem is NP-hard [37].

5.3.5 Maximal Anticlique

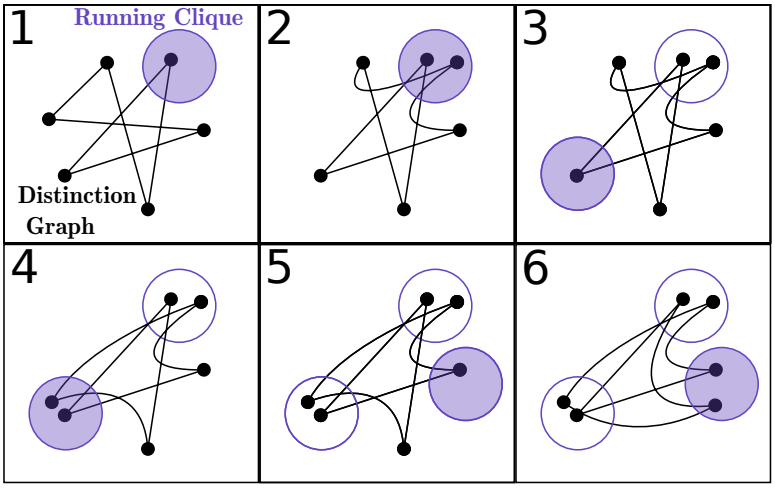
Alberto and Simão [1] propose a heuristic to increase the likelihood that a greedy algorithm produces a minimum clique covering – First, a maximum anticlique is found in the compatibility graph (This is also an NP-hard problem [37], but the size of the maximum anticlique generally grows more slowly than the graph itself). Now, each state in the maximum anticlique must belong to a different clique in the minimum clique covering. Also, every remaining state must be compatible with at least one of the states in the anticlique (or else violate the maximality of the anticlique). The greedy algorithm then proceeds, taking first the states in the maximum anticlique, then the remaining states It can easily be shown that an ordering of this type will produce a minimum reduction. The number of such orderings still grows exponentially with the number of states, but in practice it grows significantly more slowly.

5.3.6 Comparisons

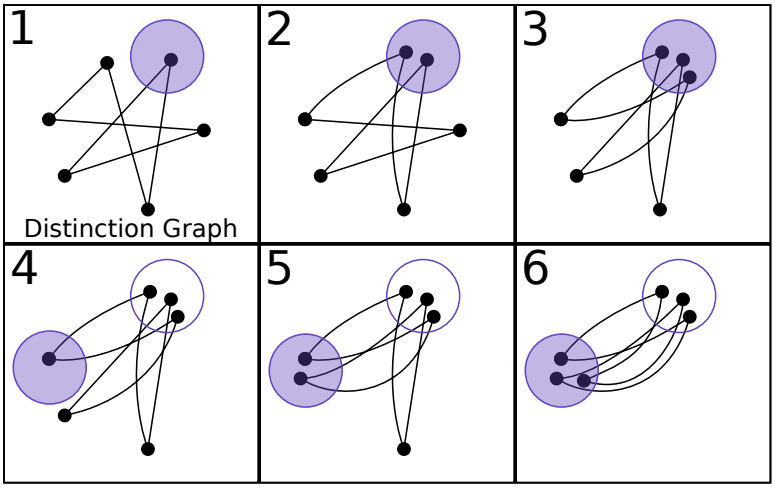
Two types of random FSMs were generated to test the correctness and numerical efficiency of the algorithms described above.

5.3.6.1 Poisson Random Tree

A Poisson random tree is an decision policy generated by recursively adding $X \sim$ Poisson(λ) children to each new node of an existing policy. We conditioned this result on the outcome that the process not terminate before the tree reaches depth H (contains a sequence of length H or greater). These decision policies are well-

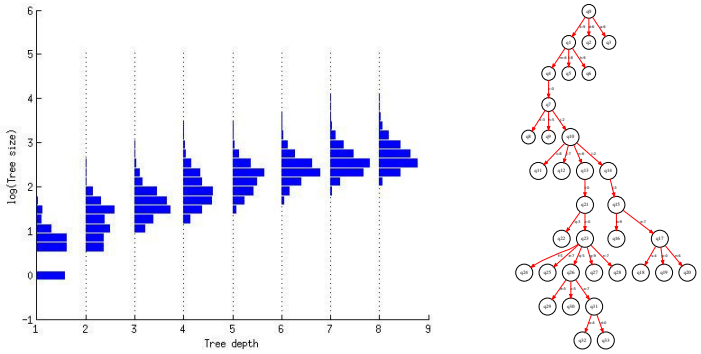


(a) Bad Ordering



(b) Good Ordering

Figure 5.6: *Greedy Algorithm on Pathological Trees* Here we illustrate the dependence of greedy algorithms on the order of their inputs. In the first set of figures (a), we proceed randomly, greedily adding random states to a running clique until none can be added (no two states sharing an edge in the distinction graph can be part of the same clique). This results in one more clique than is necessary - proceeding counterclockwise, we cover the set with only two cliques. Alberto and Simão's maximum anticlique algorithm fares no better, as the maximal anticlique has size 2.



(a) Poisson Tree Growth Trend (b) Tree, Depth 10

Figure 5.7: *Poisson Random Tree*

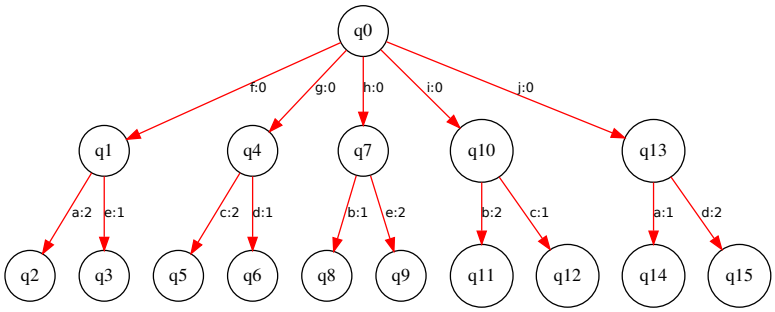
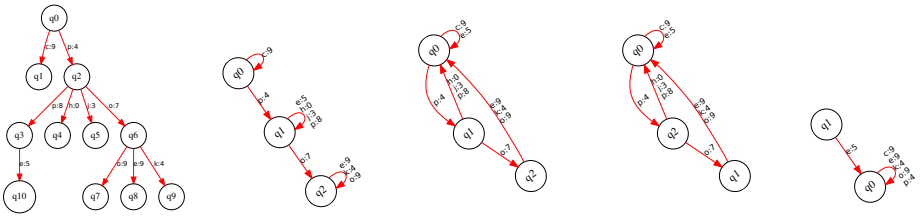


Figure 5.8: *Pathological Tree*

studied in decision theory, as they model birth/death processes where individuals continuously produce offspring at a rate of λ per lifetime.

5.3.6.2 Pathological Tree

The "pathological" tree is a policy of depth 2 with $6n + 1$ contexts. was designed to frustrate the algorithm of Alberto and Simão. Each of its states at depth 1 is incompatible with exactly two others. The resulting distinction graph consists of disjoint rings. The order in which states are added to cliques is critical. Half of all orderings result in a three-state FSM, whereas the minimum number of states is two.



(a) Original (b) Bit-at-a-Time (c) Greedy (d) Alberto-Simão (e) Exact

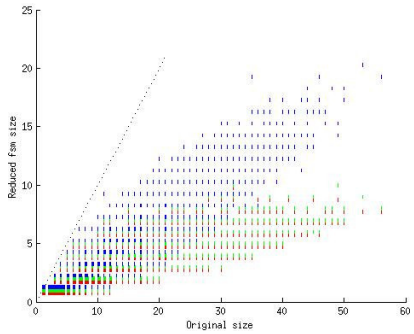
Figure 5.9: *Typical Reductions* Here we contrast the results of the various reduction algorithms introduced above. The Bit-at-a-Time method (b) produces a minimal sub-tree of the canonical policy. The last three methods are equivalent up to a reordering of states, so reductions (c) and (d) are practically identical.

5.3.7 Discussion

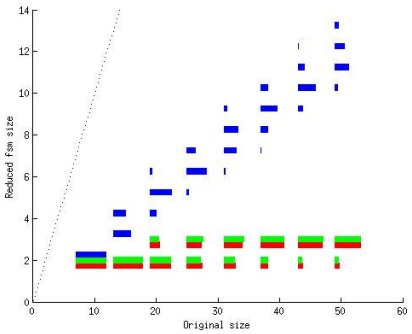
The Bit-at-a-time algorithm consistently underperformed all of the greedy clique-assembly algorithms, both in time requirements (finding the earliest ambiguous contexts took $O(N^3)$ time), and in reduction performance. However, the subtree property of its output may make it better suited for applications that require running "in-place".

Greedy algorithms seem to be the best choice, although they left something to be desired. Although better heuristics helped to reduce the likelihood of non-optimal reduction, it was not hard to find policies that could trip them up.

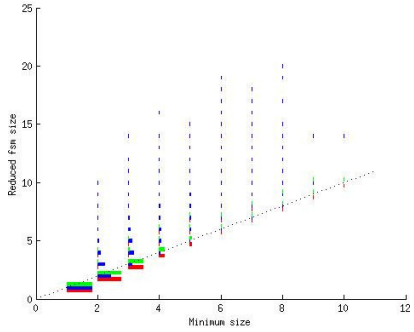
Although it is not certain how much of this work can be generalized to more interesting robotics applications (e.g. continuous time, continuous input/output, probabilistic systems), our results suggest that some general-purpose ambiguity-splitting scheme could eventually be applied to all types of POMDP solvers.



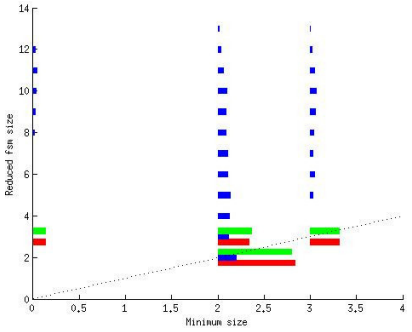
(a) Poisson Trees Reduced vs. Orig. Size



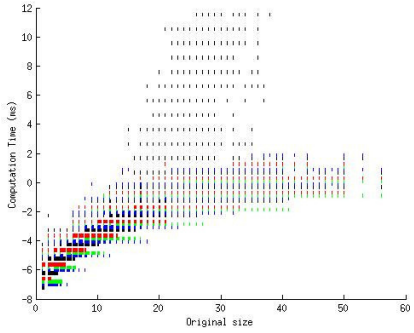
(b) Patho. Trees Reduced vs. Orig. Size



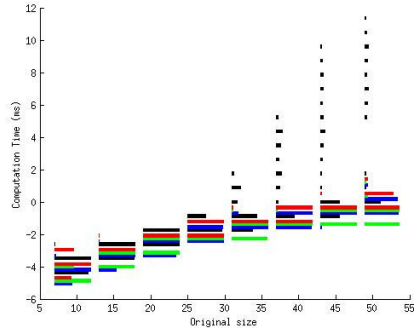
(c) Poisson Trees Reduced vs. Min. Size



(d) Patho. Trees Reduced vs. Min. Size



(e) Algo. Runtimes on Poison Trees



(f) Algo. Runtimes on Patho. Trees

Figure 5.10: Red bars pertain to Alberto and Simão's method, blue bars pertain to the bit-at-a time method, green bars pertain to the greedy clique completion method, and black bars pertain to an exhaustive search for a minimal reduction. The black dotted line in the first two rows of figures plots the line $y = x$. Note the greater proportion (50% or greater) of nonoptimal reductions in th pathological ("Patho.") examples, and the clear separation in performance between the bit-at-a-time and greedy methods.

Bibliography

[1] A. Alberto and A. Simao. Minimization of incompletely specified finite state machines based on distinction graphs. In *Test Workshop, 2009. LATW '09. 10th Latin American*, pages 1–6, March 2009.

[2] P. Algoet. The strong law of large numbers for sequential decisions under uncertainty. pages 609–633, 1994. *IEEE Transactions on Information Theory* 40, pp.1994.

[3] Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, page 0278364912461538, 2012.

[4] L.P.K̃aelbling A.R.C̃assandra and M.L. Littman. Acting optimally in partially observable stochastic domains. 1994.

[5] T. Arbel and F.P. Ferrie. Entropy-based gaze planning. 2001. Image and Vision Computing, vol. 19, no. 11, pp. 779786, 2001.

[6] Nikolay Atanasov, Bharath Sankaran, Jerome Le Ny, George J Pappas, and Kostas Daniilidis. Nonmyopic view planning for active object detection. *arXiv preprint arXiv:1309.5401*, 2013.

[7] Oscar Kin-Chung Au, Youyi Zheng, Menglin Chen, Pengfei Xu, , and Chiew-Lan Tai. Mesh segmentation with concavity-aware fields. In *IEEE Trans. Vis. Comp. Graphics*, page To appear, 2011.

[8] A. Ayvaci and S. Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, October 2012.

[9] G. Basile and G. Marro. On the observability of linear, time-invariant systems with unknown inputs. *Journal of Optimization theory and applications*, 3(6):410–415, 1969.

[10] S. Bezzaoucha, B. Marx, D. Maquin, J. Ragot, et al. On the unknown input observer design: a decoupling class approach with application to sensor fault diagnosis. In *1st International Conference on Automation and Mechatronics, CIAM'2011*, 2011.

[11] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Extracting 3d scene-consistent object proposals and depth from stereo images. In *Computer Vision–ECCV 2012*, pages 467–481. Springer, 2012.

[12] B. Burns and O. Brock. Information-theoretic construction of probabilistic roadmaps. pages 650–655, 2003. *Proceedings of the International Conference on Intelligent Robot and Systems*, pp.Las Vegas, 2003.

[13] Trevor Campbell, Miao Liu, Brian Kulis, Jonathan P. How, and Lawrence Carin. Dynamic clustering via asymptotics of the dependent dirichlet process mixture. In *Advances in Neural Information Processing Systems 26*, 2013.

[14] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *Computer Vision–ECCV 2012*, pages 430–443. Springer, 2012.

[15] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3D mesh segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.

[16] Robert Collins. A space-sweep approach to true multi-image matching. In *IEEE Computer Vision and Pattern Recognition*, pages 358–363, June 1996.

[17] C.I. Connolly. The determination of next best views. pages 432–435, 1985. *Proceedings of the IEEE International Conference on Robotics and Automation* , pp.1985.

[18] J. Denzler and C. Brown. Information theoretic sensor data selection for active object recognition and state estimation. 2002. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 145157, 2002.

[19] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

[20] H. Dimassi, A. Loría, and S. Belghith. A robust adaptive observer for nonlinear systems with unknown inputs and disturbances. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 2602–2607. IEEE, 2010.

[21] Wei Feng, Jiaya Jia, and Zhi-Qiang Liu. Self-validated labeling of markov random fields for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1871–1887, 2010.

[22] Vittorio Ferrari, Tinne Tuytelaars, and Luc Van Gool. Object detection by contour segment networks. In *Computer Vision–ECCV 2006*, pages 14–28. Springer Berlin Heidelberg, 2006.

[23] F. Hover U.M̃itra G.A.H̃ollinger, B.Ẽnglot and G.S. Sukhatme. Uncertainty-driven view planning for underwater surface inspection. 2012. *Accepted, International Conference on Robotics and Automation*, May 2012.

[24] C.J. Geyer. Markov chain monte carlo maximum likelihood. pages 156–163, 1991. *Proc. 23rd Symp. Interface. Computing Science and Statistics.*

[25] Sezer Gören and F Joel Ferguson. On state reduction of incompletely specified finite state machines. *Computers & Electrical Engineering*, 33(1):58–69, 2007.

[26] G. Graber, T. Pock, and H. Bischof. Online 3d reconstruction using convex optimization. In *1st Workshop on Live Dense Reconstruction From Moving Cameras, ICCV 2011*, 2011.

[27] Antonio Grasselli and Fabrizio Luccio. A method for minimizing the number of internal states in incompletely specified sequential networks. *Electronic Computers, IEEE Transactions on*, (3):350–359, 1965.

[28] C. Haene, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *IEEE Computer Vision and Pattern Recognition*, 2013.

[29] F Hamano and G Basile. Unknown-input present-state observability of discrete-time linear systems. *Journal of Optimization Theory and Applications*, 40(2):293–307, 1983.

[30] H. Hammouri and Z. Tmar. Unknown input observer for state affine systems: A necessary and sufficient condition. *Automatica*, 46(2):271–278, 2010.

[31] M. Hauskrecht. Value-function approximations for partially observable markov decision processes. pages 33–94, 2000. *Journal of Artificial Intelligence Research*, vol. 13, pp.2000.

[32] John Hopcroft. An n log n algorithm for minimizing states in a finite automaton. Technical report, DTIC Document, 1971.

[33] Hanqing Jiang, Haomin Liu, Ping Tan, Guofeng Zhang, and Hujun Bao. 3d reconstruction of dynamic scenes with multiple handheld cameras. In *Computer Vision–ECCV 2012*, pages 601–615. Springer, 2012.

[34] E. Jones and S. Soatto. Visual-inertial navigation, localization and mapping: A scalable real-time large-scale approach. *Intl. J. of Robotics Res.*, april 2011.

[35] E. S. Jones, A. Vedaldi, and S. Soatto. Inertial structure from motion and autocalibration. In *Workshop on Dynamical Vision*, October 2007.

[36] Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. Learning 3d mesh segmentation and labeling. *ACM Transactions on Graphics (TOG)*, 29(4):102, 2010.

[37] RichardM. Karp. Reducibility among combinatorial problems. In RaymondE. Miller, JamesW. Thatcher, and JeanD. Bohlinger, editors, *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Springer US, 1972.

[38] J. Kelly and G. Sukhatme. Fast Relative Pose Calibration for Visual and Inertial Sensors. In *Experimental Robotics*, pages 515–524, 2009.

[39] Olaf Khler and Ian Reid. Efficient 3d scene labeling using fields of trees. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013.

[40] Georg Klein and David Murray. Parallel Tracking and Mapping for Small AR Workspaces. *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–10, November 2007.

[41] Maria Klodt and Daniel Cremers. A convex framework for image segmentation with moment constraints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2236–2243. IEEE, 2011.

[42] Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012.

[43] M Pawan Kumar and Daphne Koller. Efficiently selecting regions for scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3217–3224. IEEE, 2010.

[44] M Pawan Kumar, Philip HS Torr, and Andrew Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76(3):301–319, 2008.

[45] Sanjiv Kumar and Martial Hebert. A hierarchical field framework for unified context-based classification. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1284–1291. IEEE, 2005.

[46] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8694 of *Lecture Notes in Computer Science*, pages 703–718. Springer International Publishing, 2014.

[47] Y.-H. R.T̃sai L.Ṽalente and S. Soatto. Information gathering control via exploratory path planning. 2012. *Proceedings of the Conference on Information Sciences and Systems*, Mar 2012.

[48] HUBERT Lacoin, FRANÇOIS Simenhaus, and FABIO LUCIO Toninelli. Zero-temperature 2d ising model and anisotropic curve-shortening flow. *arXiv preprint arXiv:1112.3160*, 2011.

[49] Florent Lafarge, Renaud Keriven, and Mathieu Bredif. Insertion of 3d-primitives in mesh-based representations: Towards compact models preserving the details. *IEEE Transactions on Image Processing*, 19(7):1683–1694, 2010.

[50] Dmitry Laptev, Alexander Vezhnevets, Sarvesh Dwivedi, and Joachim M Buhmann. Anisotropic sstem image segmentation using dense correspon-

dence across sections. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 323–330. Springer, 2012.

[51] Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman. A pylon model for semantic segmentation. 2011.

[52] José Lezama, Karteek Alahari, Josef Sivic, and Ivan Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3369–3376. IEEE, 2011.

[53] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[54] D. Liberzon, P. R. Kumar, A. Dominguez-Garcia, and S. Mitra. Invertibility and observability of switched systems with inputs and outputs. 2012.

[55] Dahua Lin, Eric Grimson, and John Fisher. Construction of dependent dirichlet processes based on poisson processes. In *Neural Information Processing Systems*, 2010.

[56] Y.M̃ansour M.K̃earns and A.Y. Ng. Approximate planning in large pomdps via reusable trajectories. 1999.

[57] Aleix M. Martnez, Pradit Mittrapiyanuruk, and Avinash C. Kak. On combining graph-partitioning with non-parametric clustering for image segmentation. *Computer Vision and Image Understanding*, 95(1):72 – 85, 2004.

[58] A. Mourikis and S. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 3565–3572. IEEE, 2007.

[59] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[60] Richard A Newcombe and Andrew J Davison. Live dense reconstruction with a single moving camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1498–1505. IEEE, 2010.

[61] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011.

[62] Claudia Nieuwenhuis and Daniel Cremers. Spatially varying color distributions for interactive multilabel segmentation. 2012.

[63] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187 – 1200, Jun 2014. Preprint.

[64] Near optimal Observation Selection using Submodular Functions. A. krause and c. guestrin. 2007. *American Association for Artificial Intelligence (AAAI)*, 2007.

[65] C. P. Pfleeger. State reduction in incompletely specified finite-state machines. *IEEE Trans. Comput.*, 22(12):1099–1102, December 1973.

[66] N.F̃riedman R.D̃earden and D. Andre. Model based bayesian exploration. pages 150–159, 1999. *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp.1999.

[67] Active recognition through next view planning: A survey. S.d. roy, s. chaudhury and s. banarjee. 2004. *J. Pattern Recognition*, vol. 37, no. 3, pp. 429446, 2004.

[68] Ian Reid and Keith Connor. Multiview segmentation and tracking of dynamic occluding layers. *Image and Vision Computing*, 28(6):1022–1030, 2010.

[69] Xiaofeng Ren, Liefeng Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766, June 2012.

[70] Pradeep Khosla Robert Grabowski and Howie Choset. Autonomous exploration via regions of interest. 2003.

[71] Stergios I Roumeliotis, Andrew E Johnson, and James F Montgomery. Augmenting inertial navigation with image-based motion estimation. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 4, pages 4326–4333. IEEE, 2002.

[72] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1605–1614. IEEE, 2006.

[73] R. F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H. J. Kelly, and Andrew J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conference on Computer Vision and pattern Recognition*, 2013.

[74] Thomas Schoenemann and Daniel Cremers. A combinatorial solution for model-based image segmentation and real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1153–1164, 2010.

[75] Thomas Schoenemann and Daniel Cremers. A coding-cost framework

for super-resolution motion layer decomposition. *Image Processing, IEEE Transactions on*, 21(3):1097–1110, 2012.

[76] Sunando Sengupta, Eric Greveson, Ali Shahrokni, and Philip H.S. Torr. Semantic modelling of urban scenes. In *International Conference on Robotics and Automation*, 2013.

[77] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.

[78] R. Sim and G. Dudek. Online control policy optimization for minimizing map uncertainty during exploration. pages 1758–1763, 2004. *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, pp.2004.

[79] S. Soatto. 3-d structure from visual motion: modeling, representation and observability. *Automatica*, 33:1287–1312, 1997.

[80] Jorg Stuckler, Nenad Biresev, and Sven Behnke. Semantic mapping using object-class segmentation of rgb-d images. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3005–3010. IEEE, 2012.

[81] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[82] Yee Whye Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, New York, 2010.

[83] Johannes Ulén, Petter Strandmark, and Fredrik Kahl. An efficient optimization framework for multi-region segmentation based on lagrangian duality.

[84] L. Valente, R. Tsai, and S. Soatto. Information-seeking control under visibility-based uncertainty. *J. Math. Imaging and Vision*, 2013.

[85] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann. Weakly supervised structured output learning for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2012.

[86] J. Wang and E. Adelson. Representing moving images with layers. In *IEEE Trans. on Image Processing*, volume 3(5), pages 625–638, 1994.

[87] Andreas Wedel, Clemens Rabe, Tobi Vaudrey, Thomas Brox, Uwe Franke, and Daniel Cremers. *Efficient dense scene flow from sparse or dense stereo data*. Springer, 2008.

[88] Andreas Wendel, Michael Maurer, Gottfried Graber, Thomas Pock, and Horst Bischof. Dense reconstruction on-the-fly. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1450–1457. IEEE, 2012.

[89] P. Whaite and F.P. Ferrie. Autonomous exploration: driven by uncertainty. pages 193–205, 1997. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, no.3, pp.Mar 1997.

[90] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1625–1632. IEEE, 2013.

[91] B. Yamauchi. A frontier-based approach for autonomous exploration. pages 146–151, 1997. *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp.10-11 Jul 1997.

[92] Julian Yarkony, Alexander Ihler, and Charless C Fowlkes. Fast planar correlation clustering for image segmentation. In *Computer Vision–ECCV 2012*, pages 568–581. Springer, 2012.

[93] Stella Yu and Jianbo Shi. Multiclass spectral clustering. In *Proceedings of the Ninth International Conference on Computer Vision*, 2003.

[94] Y. Yu and K. Gupta. An information theoretic approach to viewpoint planning for motion planning of eye-in-hand systems. 2000. *Proceedings of the International Symposium on Industrial Robotics*, 2000.

[95] K.W̃ang Z.Ỹu and R.-G. Yang. Next best view of range sensor. pages 185–188, 1996. *Proceedings of the IEEE International Conference on Industrial Electronics, Control, and Instrumentation* , vol.1, pp.vol.1, 5-10 Aug 1996.

[96] Christopher Zach. Fast and high quality fusion of depth maps. *Proc. 3DPVT*, 2008.

[97] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In *ICCV*, pages 1–8. IEEE, 2007.

[98] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems*, 2001.

[99] Guofeng Zhang, Jiaya Jia, and Hujun Bao. Simultaneous multi-body stereo and segmentation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 826–833, 2011.

[100] Guofeng Zhang, Jiaya Jia, Wei Hua, and Hujun Bao. Robust bilayer segmentation and motion/depth estimation with a handheld camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2011.

[101] Guofeng Zhang, Jiaya Jia, Wei Hua, and Hujun Bao. Robust bilayer segmentation and motion/depth estimation with a handheld camera. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):603–617, 2011.

[102] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Consistent depth maps recovery from a video sequence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):974–988, 2009.

[103] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3127–3134. IEEE, 2013.