Information-Driven Adaptive Structured-Light Scanners

Guy Rosman, Daniela Rus, John W. Fisher III
CSAIL, Massachusetts Institute of Technology
{rosman,rus,fisher}@csail.mit.edu

Abstract

Sensor planning and active sensing, long studied in robotics, adapt sensor parameters to maximize a utility function while constraining resource expenditures. Here we consider information gain as the utility function. While these concepts are often used to reason about 3D sensors, these are usually treated as a predefined, black-box, component. In this paper we show how the same principles can be used as part of the 3D sensor.

We describe the relevant generative model for structured-light 3D scanning and show how adaptive pattern selection can maximize information gain in an open-loop-feedback manner. We then demonstrate how different choices of relevant variable sets (corresponding to the subproblems of locatization and mapping) lead to different criteria for pattern selection and can be computed in an online fashion. We show results for both subproblems with several pattern dictionary choices and demonstrate their usefulness for pose estimation and depth acquisition.

1. Introduction

Range sensors have revolutionized computer vision in recent years, with commodity RGB-D scanners allowing us to easily tackle challenging problems such as articulated pose estimation [27], Simultaneaous Localization and Mapping (SLAM) [16, 31, 6], and object recognition [15, 21]. The use of 3D sensors often relies on a simplified model of the resulting depth images that is loosely coupled to the photometric principles behind the design of the scanner. Given this intermediate representation, we deploy computer vision algorithms to understand the world and take actions based on the acquired scene information.

Significant efforts have been devoted to optimal planning of sensor deployment under resource constraints, *e.g.*, on energy, time, or computation. *Sensor planning* has been employed in many aspects of vision and robotics, including positioning of 3D sensors and cameras, as well as other *active sensing* problems, see for example [25, 3, 2, 37, 32]. The goal is to focus sensing on the aspects of the environment or scene most relevant to the specific inference task.

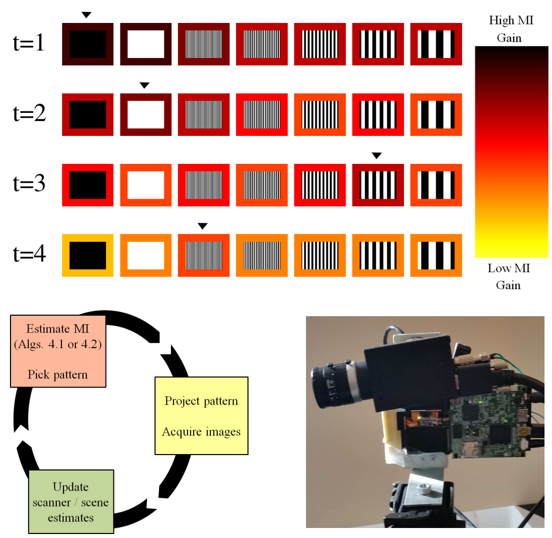

Figure 1. Illustration of patterns selection. Each row illustrates another turn of pattern selection. For each pattern, the information gain is estimated, shown by different border color around each pattern, and the different stem heights in the plot on the left. Black arrowheads and red circles in the plot mark the selected pattern at each turn. Note the different patterns selected, and diminishing information gain over time. Bottom row: Left: the proposed open-loop w/ feedback 3D scanning with pattern selection flowchart. Right: the project/camera system used for 3D scanning.

However, the same principles are generally not used to examine the operation of the 3D sensor itself. At a finer scale, each acquisition by a photosensitive sensor is a measurement, and the parameters of the sensors, including any active illumination, are an *action* parameter (in the decision-theoretic sense [29]) to be optimized and planned.

In this paper we reformulate adaptive selection of patterns in structured-light scanners as the following resource constrained sensor-selection process. We treat the choice of the projected pattern at each time as a planning choice, and the number of projected patterns as a resource. Our goal is to minimize the number of projected patterns while maximizing the task-specific *information gain*. We compute information gain from the (predicted) observation of the scene given previous observations and a new proposed projected pattern. This allows us to pick the next projected pattern in an online fashion, corresponding to the greedy selection regime in sensor selection.

The contributions of this paper are: (i) We devise a probabilistic generative graphical model for the 3D scanning process, depicted in Figure 2. We estimate mutual information between the observed images and variables in the model in Algs. 1,2. (ii) For the task of range estimation, we demonstrate greedy open-loop pattern selection for the projector in Subsec. 4.1. (iii) For the task of pose estimation, we show which parts of the scene are informative, for several cases of interest, in Subsec. 4.2.

We note that sensor planning is an instance of experimental design, studied in a variety of domains, including economics [9], medical decision making [7], robotics [17, 11], and sensor networks [4, 33, 13, 38, 5, 14]. While many optimality criteria have been proposed, one commonly used criterion is information gain. It is well-known that selection problems have intractable combinatorial complexity. However, it has been shown that tractable greedy selection heuristics, combined with open-loop feedback control [1] guarantee near-optimal performance [13, 34], due to the submodular property of conditional mutual information (MI). This assumes one can evaluate the information measure for the set of sensing choices (patterns in our current context). We derive a physics-based model for structured-light sensing that simultaneously lends itself to tractable information evaluation while producing superior empirical results in a real system. We also characterize the informational utility of a given pattern (or class of patterns) in the face of varying *relevant* versus *nuisance* parameter choices [18]. In the process, we demonstrate that the value of a given structured-light pattern changes depending on the specific inference task. We exploit commonly available graphics hardware to efficiently estimate the information gain of a selected pattern and reason about the effect of the dependency structure in the probabilistic model.

The choice of parameterization for the latent variables in the model is crucial for efficient information gain estimation. This can be seen in the common tasks of range sensing and pose estimation. We consider these two important applications and demonstrate how a careful choice of the scene and scanner representation lends itself to estimation of conditional mutual information.

In the field of structured-light reconstruction, several studies have suggested adaptive scanners (see for example [8, 19, 20, 37]), and energy-efficient designs [24]. However, unlike previous attempts that observed specific image features and addressed a specific pattern decoding technique, we show how given a generative model for the sensing process we can obtain an adaptive scanner for various tasks, forming a decision-theoretic *purposive* [22] 3D scanner.

We formulate 3D acquisition as a probabilistic inference process within a detailed model for the scene and sensor in Section 2. We discuss methods of representing uncertainty in a manner appropriate for a specific task. In Section 3 we show how MI estimation can be combined with standard approaches for reconstruction in several cases of interest, and demonstrate the integration of MI estimation into a structured-light scanner. Section 4 demonstrates the proposed system in several experiments that exemplify the usefulness of the proposed approach. Section 5 concludes the paper and describes possible new directions.

2. Modelling Active 3D Computer Vision

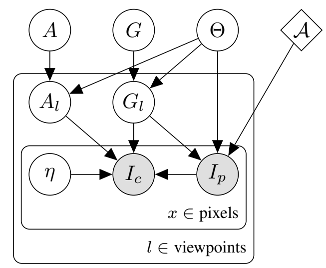

Figure 2. Proposed model for classification with active illumination.

We now describe the generative model used for pattern selection and inferring depth. We adopt a model that describes structured-light and time-of-flight imaging devices and standard cameras or camera-and-projector systems. Estimation of information gain is central to our method and thus impacts the choice of parameterization. We emphasize that approximations we use for estimating information gain and choosing patterns generally do not carry over when we compute the reconstruction. To our knowledge, this is the first analysis of active information-based planning in this setting. The model parameters are roughly partitioned into agent pose, geometry of the scene, and photometry of the scene. We summarize the notation below (see the supplement for further details):

- A and G denote the photometric and geometric properties of the scene and are modeled as Gaussian per scene element as described in Section 3.
- Θ denotes the scanner/agent pose. It is distributed as a Gaussian in the Lie-algebra se(3). If range estimation is solely of interest, Θ is assumed to be fixed.
- A_l, G_l denote the view-dependent representations of the scene. They are not deterministic functions of A, G, Θ due to unmodeled aspects (*e.g.* occlusions). The geometry and pose determine camera and projector coordinates at each pixel.

- I_c and I_p denote the camera and projector intensity values corrupted by additive per-pixel noise $\eta(x)$. $x \in \mathbb{R}^2$ denotes pixels in the camera image plane.
- $\mathcal{A}$ denote the pattern selection.

The generative graphical model of Figure 2 depicts the relationships of the variables. Observations are denoted by shaded circles, latent variables by white circles, and parameters by diamonds. As shown in Figure 2, the model factorizes as

$$\begin{aligned} &p\left(A, G, \Theta, A_l, G_l, \eta, I_c, I_p; \mathcal{A}\right) \\ &\quad = p\left(\Theta\right) p\left(A\right) p\left(G\right) \\ &\quad\quad \prod_{l} p\left(A_l | A, \Theta\right) p\left(G_l | G, \Theta\right) \\ &\quad\quad \prod_{l,x} p\left(I_c | A_l, G_l, I_p, \eta\right) p\left(I_p | G_l, \Theta; \mathcal{A}\right) p\left(\eta\right), \end{aligned} \tag{1}$$

where the first line includes prior terms for the scene. The second incorporates projection onto a specific viewpoint of the projector images and world model, and the last line involves sensor image rendering, and noise realization.

We note that depending on the inference task various latent variables alternate their roles as either relevant or nuisance. We choose patterns in order to maximize *focused* information gains [18], i.e., information regarding the *relevant* set, rather than information of the non-relevant, or nuisance, variables. We follow the notation of [18] where $\mathcal{R} \subseteq \mathcal{U}$ denotes the relevant set and $\mathcal{U}$ denotes the set of all nodes. Nuisance parameters have certainly been considered in existing 3D reconstruction methods. Examples include the standard *binarize-decode-reconstruct* approach for time-multiplexed structured-light scanners or the choice of view-robust descriptors for 3D reconstruction from multiple views [28]. The utility of the generative model is that nuisances are dealt with in a mathematically-consistent fashion.

2.1. Inference and Sensor Planning in 3D Vision

We consider several inference tasks of interest in 3D computer vision and the pattern selection issues which arise. For example, inference of G_l given I_c, I_p, Θ amounts to 3D reconstruction, where G_l is assumed to approximate G and A_l is treated as a nuisance. Previous methods adopt a probabilistic model for improving structured-light reconstruction [30, 26], but assume a predetermined set of patterns. Alternatively, *Simultaneous Localization and Mapping* (SLAM) methods incorporate inference steps for the geometry and pose parameters alternating between pose (Θ) updates conditioned on the geometry (G_l) and vice-versa. Updates to the 3D map may be posed as inference of G given G_l, Θ. In all cases, limiting assumptions regarding occlusions, the relation of appearance parameters and 3D geometry, and the relation between different range scans of the same scene are typically invoked.

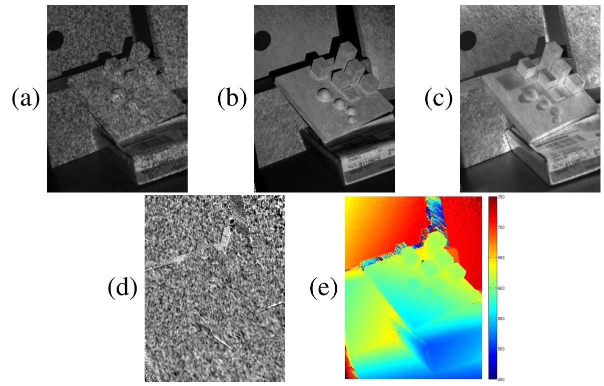

Figure 3. Left-to-right: a) I_c, b) coefficients a and c) b from Equation 2 for the MAP-estimated range d) I_p in the camera image plane, e) the range image in mm. Note how parameter b captures scene illumination, whereas parameter a captures the reflectance coefficient of the surface with respect to the projector.

For structured-light acquisition, one can associate pixels in I_c and I_p given the range r at each pixel x (which is a choice for G_l) and the pose Θ. The set of pixels in I_p are obtained via $\Pi_{r,\Theta}(x) \in \mathbb{R}^2$ by back-projecting x into the 3D world and projecting it into the projector image plane. The relation between the intensity values of these pixels can be given as

$$I_c(x) = a(x) I_p\left(\Pi_{r,\Theta}(x)\right) + b(x) + \eta(x), \tag{2}$$

where a, b depend on the ambient light, normals, and albedo of the incident surface. For sufficiently large photon count, η is assumed Gaussian accounting for sensor noise and unmodeled phenomena such as occlusions and non-Lambertian lighting components. Utilizing time-multiplexed structured-light, plane-sweeping [26] enables efficient inference of G_l from I_c, I_p, and incorporation of priors on the scene structure G. For our purposes, one can assume a fixed pose, and limit the inference to estimation of G_l. Figure 3 provides an example of I_c, I_p, a, b, r for a reconstructed scene with random smoothed patterns (as described in Subsection 4.1). The resulting 3D reconstruction is superior to the classic binarize-decode-triangulate pipeline with respect to robustness to artifacts such as specularities and low SNR conditions.

Our goal is to efficiently compute the relevant mutual information quantities $\mathcal{I}^{\mathcal{A}}\left(\mathbf{x}_{\mathcal{R}}; I_C\right)$ for different definitions of $\mathcal{R}$, and choices from the set $\mathcal{A}$, alternately considering Θ, G, and A as the relevant variable set $\mathbf{x}_{\mathcal{R}}$. Nonlinear correspondence operators (back-projection and projection) linking I_c, I_p complicate dependency analysis within the model and preclude analytic forms. We exploit common graphics hardware for a straightforward and efficient sampling approach that follows the generative model.

2.2. Photometric Entropy in Active Illumination 3D Scanning

When describing 3D scanner, the interplay of photometric models and the reconstruction can lead to improved results [35, 23] and warrants examination. In Equation 2, coefficients a and b capture illumination variability. A slightly more detailed description of the photometric model

$$I_c = \rho \frac{1}{r_p(x)^2} \langle n(x), l \rangle \, I_p(\pi_r(x)) + \rho I_{amb}, \tag{3}$$

aids in our understanding of the contributions of the different factors. Here, ρ is the albedo coefficient, $n(x)$ is the surface normal at a given image location x, l is the projector direction, and I_{amb} is the ambient lighting. r_p is the distance from the projector, and $I_p(\pi_r(x))$ is the projector intensity, assumed pixel-wise independent. Observing the pixel intensity entropy associated with different simplifications of this model provides us with intuition on the relative importance of various factors and gives us some bounds on how much information can be gained from modification of the patterns. Specifically, the difference in image entropy between an arbitrary i.i.d. pattern, and a deterministic pattern that deforms according to the geometry gives us a bound on the maximum information gain. In the supplement, we construct a synthetic experiment that evaluates the sensitivity of entropy and information measures to each factor.

3. Estimating Uncertainty in 3D Scanners

We present two important cases of estimating mutual information gain for pattern selection in structured-light scanners. In each, we consider inference over different subsets of variables, and the mutual information between them and the observed images. Differing assumptions on the fixed/inferred variables and dependency structure in the image formation model lead to different algorithms for MI estimation given as Algorithms 1 and 2.

An important observation is that given the pose, range measurements and camera image pixel values can be approximated as an independent estimation problem per-pixel (here we model the effect of surface self-occlusions as noise). This provides an efficient and parallelizable estimation procedure for the case of range estimation. This assumption has been exploited in plane-sweeping stereo, and we now utilize it for MI estimation. We note that even where the inter-pixel dependency is not negligible, we can compute an upper bound for the information gain. For example, for the case of pose and range estimation we obtain

$$\begin{aligned} \mathcal{I}(I_c;\Theta,r) =& H(I_c) - H(I_c|\Theta,r) \leq \\ & \sum_x H(I_c^x) - \sum_x H(I_c^{(x)}|\Theta,r) \triangleq \hat{\mathcal{I}}(I_c;\Theta), \end{aligned} \tag{4}$$

where $\hat{\mathcal{I}}$ is the *pixel-wise mutual information* between the sensor and the inferred parameter.

3.1. Range Image MI Estimation

We start with the simple, yet instructive, case of estimating mutual information between the scene geometry and the observed images given a known set of illumination patterns. Here, inference is over G_l as represented by the range at each camera pixel $r \equiv r(x)$. We assume a Gaussian prior for a and b.

We compute the pixel-wise mutual information individually and sum the results. In this subsection, we assume a deterministic choice of pose; the patterns are deterministic throughout the paper, and hence omitted from the notation for $\mathcal{I}$. The mutual information between I_c and G_l given θ,I_p is given by

$$\begin{aligned} \hat{\mathcal{I}}\left(I_c;G_l|\theta\right) &= \sum_x \mathcal{I}\left(I_c(x);r(x)|\theta\right) \\ &= \sum_x E_{I_c,r|\theta}\left[\left(\left(\log \frac{p(I_c|r,\theta)}{p\left(I_c|\theta\right)}\right)\right)\right]. \end{aligned} \tag{5}$$

While computing $p(I_c|r,\theta)$ is straightforward, we are still forced to estimate $p(I_c|\theta)$, which can be done by marginalizing over r according to our posterior estimates,

$$p(I_c|\theta) = E_r[p\left(I_c|r,\theta\right)]. \tag{6}$$

For each sample of θ, r, we can then compute the log of the likelihoods ratio, and integrate it. We note the existence of alternatives such as using GMMs or Laplace approximations, for efficient implementation.

We perform one sampling loop in order to estimate $p(I_c|\theta)$. We then use another set of samples in order to estimate $\mathcal{I}\left(I_c;G_l|\theta\right)$. Algorithm 1 describes computation of the MI gain for frame T.

Since $a, b, \eta^{(0..T)}$ are all are assumed to be Gaussian conditioned on r, $p\left(a, b, I_c^{(t)}|I_p^{(0...t)}, I_c^{(0...t-1)}\right)$ is Gaussian. We can compute the pdf of a, b and $I_c^{(T)}$ given $I_p^{(0...T)}$ and $I_c^{(0...T-1)}$, by conditioning on each image t at a time, computing $p\left(a, b, I_c^t|I_c^{0..t-1}\right)$ for each $t = 0..T$ iteratively. This allows fast computation on parallel hardware such as graphics processing units (GPUs), without explicit matrix inversion or other costly operations at each kernel.

3.2. Pose MI Estimation with Structured-Light

A second important case we explore is typical of pose estimation problems, where we try to infer a low-dimensionality latent variable set with global influence, in addition to range uncertainty. In 3D pose estimation, we usually estimate Θ given a model of the world G. In visual SLAM, G, A, A_l are commonly used to infer Θ, G_l, either as online inference [31], or in batch-mode [12], where usually a specific function of the input (feature locations from different frames, or correspondence estimates) is taken. In

Algorithm 1 MI estimation / pattern selection for range image

for pattern p, in each pixel x **do**
  for samples $i = 1, 2, \dots, N_{hist}$ **do**
    Sample a range value for x according to $p(r)$.
    Raytrace I_p, sample I_c. Compute the statistics of a, b, I_c conditioned on previous image measurements.
    Compute probability $p(I_c|r)$.
    Update the estimated per-pixel histogram, $p(I_c)$.
  end for
  for samples $i = 1, 2, \dots, N_{MI}$ **do**
    Draw a new range value for x according to a proposal distribution $p(r)$.
    Raytrace I_p, sample I_c. Compute the statistics of a, b, I_c conditioned on previous image measurements.
    Compute probability $p(I_c|r)$, estimate $\log\left(\frac{p(I_c|r)}{p(I_c)}\right)$.
    Update the estimated mutual information.
  end for
end for
Pick pattern p with maximum MI sum over the image

Algorithm 2 MI estimation / pattern selection for pose estimation

1: **for** pattern p, in each pixel x **do**
2:   **for** samples $i = 1, 2, \dots, N_{hist}$ **do**
3:     Draw pose sample θ_i, compute T_{θ_i}
4:     **for** each sampled range value $r(x)$ **do**
5:       Back-project x^3, compute $\bar{x}^3 = T_{\theta_i, r}(x)$.
6:       Project $\bar{x}^3$ and sample $I_p^{1 \dots t}$, sample $I_c^{1 \dots (t-1)}$.
7:       Compute the statistics of $a, b, I_c^{(t)}$ conditioned on previous image measurements and r sample.
8:       Update the estimated per-pixel histogram, $P(I_c)$
9:     **end for**
10:   **end for**
11:   **for** samples $i = 1, 2, \dots, N_{MI}$ **do**
12:     Draw pose sample θ_i and associated transformation T_{θ_i}
13:     **for** each sampled range value $r(x)$ **do**
14:       Back-project x^3, compute $\bar{x}^3 = T_{\theta_i, r}(x)$.
15:       Project $\bar{x}^3$ and sample $I_p^{1..t}$, sample $I_c^{1..(t-1)}$.
16:       Compute $a, b, I_c^{(t)}$ estimates conditioned on previous image measurements, and r sample.
17:       Estimate $log\left(\frac{P(I_c|a,b,I_p,T_{\theta_i})}{P(I_c)}\right)$.
18:       Update the mutual information gain estimate.
19:     **end for**
20:   **end for**
21: **end for**
22: Pick pattern p with maximum MI sum over the image.

depth-sensor based SLAM, the range sensors obtain a measurement G_l under some active illumination. Θ is then approximated from G, G_l.

We now describe computation of the MI between the pose and the images. As before, we parameterize G_l by $r(x)$, and given (Θ, r) we re-establish a correspondence between I_p and I_c. This is done by computing a back-projected point x_j^3 (denoting it is a 3D point), transforming it according to Θ to get $\bar{x}_j^3$, and projecting $\bar{x}_j^3$ onto the camera and projector image. A similar situation would arise where inferring a class variable, where instead of merely inferring Θ we also infer a categorical variable C that determines the class of the observed object. Here too, we can still use the following observations: (i) given the pose parameters, the problem can still be approximated as a per-pixel process – this assumption underlies most visual servoing approaches. (ii) the pose parameter space is low-dimensional and can be sampled from, as is often done in particle filters for pose estimation. We can therefore write

$$\mathcal{I}\left(I_c^{(x)}; \Theta | G_l\right) = E_{I_c, \Theta, r}\left(\log \frac{P(I_c^{(x)}|\Theta)}{P\left(I_c^{(x)}\right)}\right), \quad (7)$$

where as before, $P\left(I_c|\theta\right)$ is computed by marginalization over r. This procedure is detailed as Algorithm 2. When computing $p(I_c^{(x)}|\Theta)$, $p(\Theta)$ can be conditioned on previous observations, and sampled from the current uncertainty estimate for the pose and range.

We note that when sampling the pose, different variants of the range images can be used, allowing us to marginalize w.r.t. range uncertainty as well.

When sampling a conditioned image model per pixel, collisions in the projected pixels can occur. While these can be arbitrated using atomic operations on the GPU, the semantics of write hazards on GPUs are such that invalid pixel states can be avoided. Furthermore, to allow efficient computation on the GPU, we must consider memory access patterns. In our implementation we compute proposal image statistics given θ, and then aggregate the contribution into the accumulators for the mutual information per pixel.

Extension to classification we could incorporate categorical variables, including object classes as part of Θ. This requires merely changing lines 4,14, in Algorithm 2 to sample a distribution over $\bar{x}_j^3(\theta, C, r)$ instead of $\bar{x}_j^3(\theta, r)$. This allows us to choose patterns for object classification tasks, which is beyond the scope for this paper.

While sampling the full space of appearance and range per-pixel is computationally expensive, running the algo-

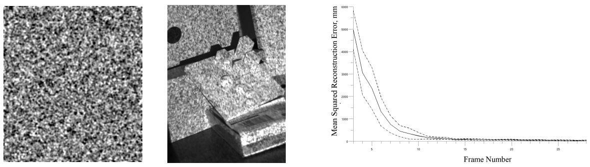

Figure 4. Left-to-right: a projected Gaussian-smoothed pattern, a captured image, average reconstruction error as a function of the number of patterns used. Dashed lines mark the standard deviation over pattern sequences.

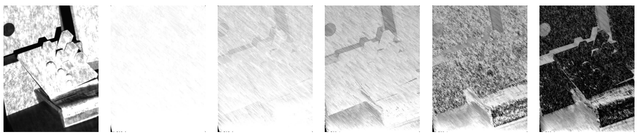

Figure 5. Left-to-right: An indicator image of reflected patterns amplitudes, followed by the mutual information between the image and the range, for random Gaussian-smoothed patterns. The initial patterns are dominated by well-illuminated areas, followed by poorly-illuminated areas (a secondary trend relates to the surface illumination angle).

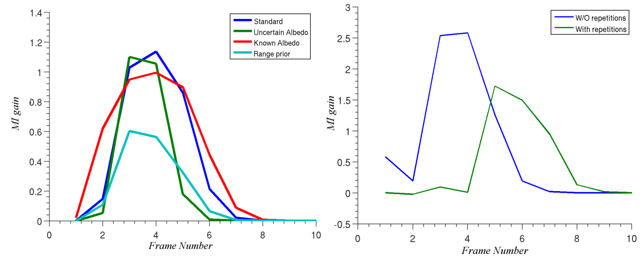

Figure 6. Left: Mutual information gain under different assumptions on the scene: Blue line - the standard case of large range and albedo uncertainty of $\sigma_r = 300mm$, $\sigma_a = 3, \sigma_b = 300$. Red line - $\sigma_a = 30, \sigma_b = 3000$ (high uncertainty of the appearance). Green line - $\sigma_a = 0.3, \sigma_b = 20$ (strong prior on the appearance). Cyan line - $\sigma_r = 7mm$ (low initial uncertainty of the range). Given a good prior on the nuisance parameters of the albedo, range is estimated more quickly in terms of frames. Given a strong range prior, the region does not require as many patterns for estimation, and overall MI gain is smaller. Right: Blue - information gain for a set of different patterns. Green - where only half of the patterns are shown, but they are repeated twice. The information gain is much lower in the second case.

rithm without any optimizations on a GPU takes approximately one second on an Nvidia Quadro K2000.

4. Numerical Results

We conducted several experiments aimed at giving an intuition for the approach proposed in this paper, and demonstrating its utility, with several choices of projector patterns and scenes. In terms of the relevant sets of variables, we have focused on range sensing and pose estimation.

4.1. Pattern Choice for Range Sensing

We first demonstrate the setup used. For pattern libraries we used a set of random patterns generated by smoothing i.i.d. Gaussian noise with Gaussian filters of various scales, and striped patterns of the sort used for gray-code structured-light. They are shown in Figures 5 and 9, respectively. We used as test objects both fabricated models with various scales of features, see Figure 5, and coated/raw wooden art models. The PointGrey Grasshopper II camera and TI LightCrafter projector used are shown in Figure 1. Pixel noise standard deviation was about $2.5/255$ for most experiments. We validate the use of the smoothed Gaussian patterns for reconstruction in Figure 4, demonstrating the decrease in the average range L2 error measured as we use more patterns for reconstruction. We use the reconstruction from a set of 120 patterns as a ground-truth estimate, making the assumption that the reconstruction is an unbiased estimator, so that reconstruction using all patterns is considered a ground-truth.

In Figure 5 we show the MI gain collected over the scene, averaged over 50 random pattern sequences. The amount of information gained from the patterns decreases as we add more patterns, as expected with MI, and surfaces that are well-illuminated and frontal-facing having faster uncertainty reduction. We look at the average MI gain per pattern over various random sequences of patterns, in Figure 6. We highlight several interesting cases. The first case (which often occurs in practice) assumes high uncertainty of the range or the appearance coefficients. The second and third cases involve less and more certainty in the appearance coefficients respectively. The fourth case involves having a good initial guess (std. of 7mm) for the range. As expected, the certainty of the appearance coefficients increases the MI between the images and the range. Having a good range prior decreases the amount of information gained per frame and the overall MI.

We then proceed to perform selection according to MI gain based on the proposed model. Although we perform greedy (pattern at a time) selection, there are bounds guaranteeing the performance of a greedy vs. optimal selection of the whole pattern sequence – see [34] for such bounds and the relevant terminology. In our test we initialize each attempt from a pair of randomly chosen patterns. At each turn we try ten randomly chosen patterns and compute their image-range MI. We pick the the most informative pattern, and contrast this with a random pattern selection. The MI gains for two scenes are measured in Table 1, collected over ten instantiations.

In one scenario, we modulate the patterns by spatial bands in the projector's image plane: 14 bands in the x and in the y directions with 15 random textures instantiations for each band, see example in Figure 7(a). From these we greedily select patterns in ten sequences, and unify them into 69 unique patterns. The patterns are mostly those that illuminate the region of interest, as expected by their high MI gain. The region of interest is defined as the silhoutte of an object (the hand) in the image. A similar test was done with patterns modulated by an exponentially, radially

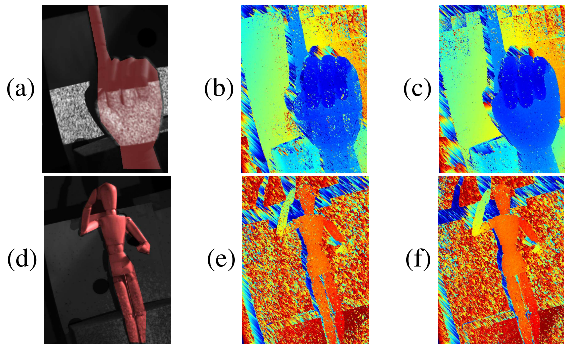

Figure 7. Left-to-right: camera image with a projected pattern on the marked object (red overlay marks the mask used for MI integration). The area covered by the mask received significantly more pattern coverage and the reconstruction with these bands is considerably better than random selection. Top: reconstruction with a random set of 69 bands (range RMS=24.1mm) vs. reconstruction with the set of 69 bands selected by a greedy selection (range RMS=18.9mm). Bottom: reconstruction with a random set of 65 blobs (range RMS=59.1mm) - random vs. greedy.

decreasing envelope, illuminating local regions of the projector field of view at each time (see Figure 7(d)). 20 random patterns are taken, modulated by 15 random locations. Of these, 65 are selected after removing repetitions. Here the region of interest was the mannequin. We use these pattern sets to reconstruct the range image, and compare to randomly choosing the same number of patterns. Qualitatively, the selected patterns often illuminated parts of the objects which were poorly reconstructed, as expected. As we show in Figure 7, we get significantly more accurate reconstruction compared to random selection—18.9mm RMS, compared to 24.1mm RMS for the hand example, and 51.3mm compared to 59.1mm in the mannequin example. This demonstrates the usefulness of our selection criteria when judged by reconstruction accuracy.

Finally, in order to demonstrate that greedy selection improves reconstruction, on average, per pattern selection, we perform ten greedy selection steps, selecting a single pattern out of ten randomly drawn ones, and demonstrate the resulting reconstruction. We take striped gray-code patterns modulated by radially-decreasing piece-wise smooth masks, centered at various locations, for a total of 240 patterns. The results of adding patterns at random vs. greedy selection show that even when we do not yet have reasonable reconstruction, greedy selection according to MI improved L2 reconstruction error. Despite the fact the L2 reconstruction error does not directly coincide with MI, we show that computing MI gain according to our model results early on in the reconstruction sequence in improved reconstruction results, as shown in Figure 8. For example, the depth reconstruction error obtained by 10 random patterns is obtained with less than six patterns in the greedy case, representing a 40% speedup.

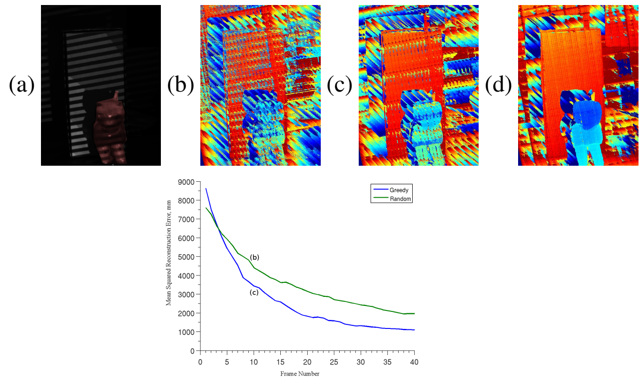

Figure 8. Top, Left-to-right: camera image with a projected pattern on the marked object (MI integration mask shown in red), range image of the scene as reconstructed by random selection of 10 patterns, greedy selection of 10 patterns and the full set of 240 patterns, reconstruction squared error as a function of the number of patterns addded, averaged over 20 trials. Bottom: error between partial frame sets reconstruction and the full 240 frames reconstruction, where frames are added at random (green) or using our approach (blue). Greedy selection based on our model improves reconstruction results with significantly fewer frames (50%), as demonstrated by Subfigures (b) and (c).

4.2. Pattern Choice for Pose Estimation

In Figures 9–12 we show computed per-pixel MI between a new camera image and the pose, assuming a highly certain range image, as estimated by Algorithm 2. We start in Figure 9 with a synthetic case where the results are easy to interpret, with a scene made of a single large corner. The pattern set for this experiement is the standard gray-code striped patterns, shown in the first row. We assume only translational uncertainty; we leave reasoning about the full SE(3) pose space to future work as it is less instructive. We use stripes going from coarse to fine, stopping at a pattern of four pixels stripe width in the projector image plane. At this phase, the appearance coefficients A, G are well estimated. In this example the camera and the projector are facing the z direction, and in front of them there is a large smoothed corner. We compare a case of uncertainty in the xy plane, to that of uncertainty in the z plane in terms of the pixel-wise MI gain. The large sloped corner and the edges are the main source of uncertainty reduction in xy since the rest of the scene is planar. In the z uncertain case, the full image is informative to the same extent. The intermediate case is a mix between the two, as expected.

For pattern selection, in Figure 10 we demonstrate pattern choice according to the proposed criteria for choosing patterns in a structured-light scanner. This shows that for an unknown pose information can be obtained from edges and corners; given a reasonable model of the scene, we can use mutual information to suggest which pattern to use to project only informative parts of the scene. The patterns chosen consist of a striped pattern projected only along a partial band of the projector screen. Figure 11 demostrates a

	Hand Mean MI, Greedy	STDev, Greedy	Mean MI, Random	STDev Random	Mannequin Mean MI, Greedy	STDev, Greedy	Mean MI, Random	STDev Random
Step 1	0.4168	0.2820	0.1267	0.0957	0.1688	0.0561	0.0756	0.0504
Step 2	0.7904	0.2803	0.3263	0.2457	0.2404	0.0694	0.0653	0.0484
Step 3	0.8129	0.1820	0.2686	0.1694	0.3030	0.0916	0.1199	0.0695
Step 4	0.6232	0.1125	0.2125	0.1591	0.2911	0.0806	0.0997	0.0939
Step 5	0.1562	0.0995	0.0903	0.1317	0.1334	0.0450	0.0744	0.0656
Step 6	0.0229	0.0264	0.0376	0.0433	0.0400	0.0232	0.0482	0.0486

Table 1. MI gain starting from two random patterns, when using greedy selection, compared to random pattern selection. Resulting MI gains are shown for the hand and mannequin examples from Figure 7. Our MI-greedy approach obtains a larger information gain, and does so faster (in frame counts) than a random ordering of frames.

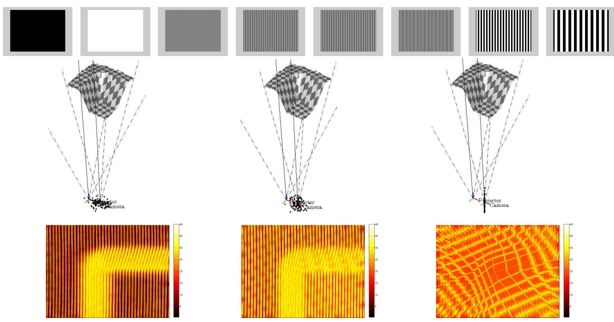

Figure 9. Per-pixel information gain for the case of initial uncertain scanner position. Left-to-right, top row: a set of patterns used for 3D sensing for pose estimation. Middle row: a rendering of the scene with sensor pose samples (black dots) in 3 scenarios, and the fields of view of the projector and camera. Bottom row: pixelwise mutual information estimates: with high uncertainty in the x-y plane of the scanner, uncertainty in x-y-z, and z-only uncertainty in scanner position. Yellow and red marking high and low information gain, respectively. Surfaces at sharp angles to the projector and camera provide greater uncertainty reduction in the x-y directions, whereas for uncertainty in the z axis, all surfaces are informative.

different set of patterns, of stripes modulated by a Gaussian mask, allowing to focus a pattern in a small region, which is important in practical applications. As can be seen, the top-ranking patterns are those that illuminate edges in the scene, which should give us high uncertainty reduction. MI for pose estimation can also be seen with real scenes. In Figure 12 we show pixelwise pose estimate for Gaussian smoothed patterns. The most informative pixels are edges and sloped areas, where the perceived projector intensity changes rapidly as a function of the pose.

5. Conclusions

In this paper we present a novel information-driven approach to planning into 3D sensors at the sensor level. We demonstrate how different uncertainty estimates and sensor models lead to different criteria for pattern selection. Future work includes the completion of a prototype scanner based on the proposed approaches. This decision-theoretic approach where action choice is identified with pattern selection in structured-light easily extends to other reconstruction techniques such as depth-from-focus (see for example [36]) and compressive sensing time-of-flight [10]. We intend to explore these in future work.

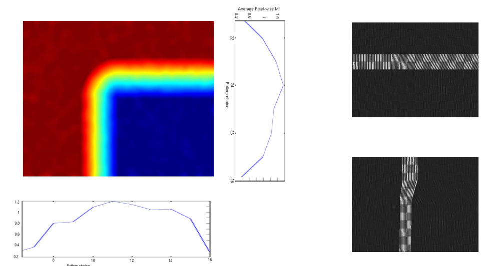

Figure 10. Left: the depth image and the MI scores of vertical and horizonal stripe masks of the patterns with respect to pose estimation in the xy plane. Right: the top-scoring horizontal and vertical patterns, as seen when projected onto the scene. As can be seen, the patterns that were selected are the ones illuminating the edges and corner.

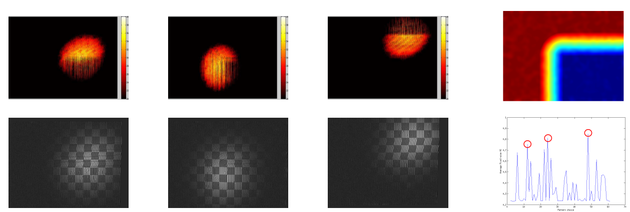

Figure 11. Left-to-right, top row: the top 3 selected masks from a set of 60 masks, and the range image. Bottom row: a MAP estimated images for the 3 masks, used when estimating the MI for each pattern, followed by the average MI scores for the patterns. Red circles mark the patterns shown.

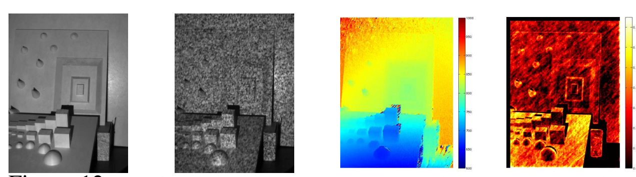

Figure 12. Left-to-right, top row: an image of the scene, one of the projected patterns as capture, the range image, the pixelwise mutual information with respect to the pose, which initial uncertainty in the camera's xy plane. The main informative areas are the cones, and regions that face the x, y directions.

Acknowledgements

The authors thank Christopher Dean for general and helpful discussions. Support for this research has been provided by ONR MURI N00014-09-1-1051, N00014- 11-1-0688, and ARO MURI W911NF-11-1-0391. We are grateful for this support.

References

[1] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000. 2

[2] F. S. Cohen and D. Cooper. A decision theoretic approach for 3-d vision. In *CVPR*, pages 964–972, Jun 1988. 1

[3] J. Denzler and C. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE-TPAMI*, 24(2):145–157, Feb 2002. 1

[4] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. pages 588–599. VLDB Endowment, 2004. 2

[5] E. Ertin, J. W. Fisher III, and L. C. Potter. Maximum mutual information principle for dynamic sensor query problems. In *IPSN*, pages 558–561. Springer, Feb 2003. 2

[6] M. F. Fallon, H. Johannsson, and J. J. Leonard. Efficient scene simulation for robust Monte Carlo localization using an RGB-D camera. In *ICRA*, May 2012. 1

[7] Y. Fumie. Value of information analysis in environmental health risk management decisions: Past, present, and future. *Risk analysis : an international journal.*, 2004,. 2

[8] L. V. Gool and T. P. Koninckx. Real-time range acquisition by adaptive structured light. *IEEE-TPAMI*, 28(3):432–445, March 2006. 2

[9] R. A. Howard. Information value theory. *IEEE Trans. Systems Science and Cybernetics*, 2(1):22–26, Aug 1966. 2

[10] J. C. Howell, G. A. Howland, A. Kirmani, A. Colaco, and V. K. Goyal. Compressive depth map acquisition using a single photon-counting detector: Parametric signal processing meets sparsity. In *CVPR*, pages 96–102, 2012. 8

[11] B. J. Julian, M. Angermann, M. Schwager, and D. Rus. Distributed robotic sensor networks: An information-theoretic approach. *I. J. Robotic Res.*, 31(10):1134–1154, 2012. 2

[12] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, Nara, Japan, Nov. 2007. 4

[13] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, June 2008. 2

[14] C. Kreucher, K. Kastella, and A. Hero. Sensor management using an active sensing approach. *Signal Processing*, 85(3):607–624, 2005. 2

[15] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *ICRA*, 2011. 1

[16] J. J. Leonard and H. F. Durrant-Whyte. Simultaneous map building and localization for an autonomous mobile robot. In *IROS*, pages 1442–1447, 1991. 1

[17] D. S. Levine. Information-rich path planning under general constraints using rapidly-exploring random trees. Master's thesis, MIT, Dept. of Aero.-Astro., June 2010. 2

[18] D. S. Levine and J. P. How. Sensor selection in high-dimensional Gaussian trees with nuisances. In *NIPS*, pages 2211–2219, 2013. 2, 3

[19] Q. Li, M. Biswas, M. Pickering, and M. Frater. Dense depth estimation using adaptive structured light and cooperative algorithm. In *CVPR Workshops*, pages 21–28, June 2011. 2

[20] X. Maurice, P. Graebling, and C. Doignon. Real-time structured light coding for adaptive patterns. *Journal of Real-Time Image Processing*, 8(2):169–178, 2013. 2

[21] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 1

[22] S. K. Nayar, V. Branzoi, and T. E. Boult. Programmable Imaging: Towards a Flexible Camera. *Int. J. of Computer Vision*, Oct 2006. 2

[23] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein. RGBD-fusion: Real-time high precision depth recovery. In *CVPR*, pages 5407–5416, 2015. 4

[24] M. O'Toole, S. Achar, S. G. Narasimhan, and K. N. Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Trans. on Graphics*, 34(4):35:1–35:13, 2015. 2

[25] L. Paletta, M. Prantl, and A. Pinz. Learning temporal context in active object recognition using bayesian analysis. In *ICPR*, volume 1, pages 695–699, 2000. 1

[26] G. Rosman, A. Dubrovina, and R. Kimmel. Sparse modeling of shape from structured light. In *3DIMPVT*, pages 456–463, Washington, DC, USA, 2012. IEEE Computer Society. 3

[27] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-Time human pose recognition in parts from single depth images. June 2011. 1

[28] S. Soatto. Steps towards a theory of visual information: Active perception, signal-to-symbol conversion and the interplay between sensing and control. *CoRR*, abs/1110.2053, 2011. 3

[29] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998. 1

[30] J. P. Tardif and S. Roy. A MRF formulation for coded structured light. In *3DIM*, pages 22–29, Washington, DC, USA, 2005. IEEE Computer Society. 3

[31] S. Thrun. Robotic mapping: A survey. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millenium*. Morgan Kaufmann, 2002. 1, 4

[32] L. Valente, Y.-H. R. Tsai, and S. Soatto. Information-seeking control under visibility-based uncertainty. *Journal of Mathematical Imaging and Vision*, 48(2):339–358, 2014. 1

[33] J. L. Williams, J. W. Fisher III, and A. S. Willsky. Approximate dynamic programming for communication-constrained sensor network management. *IEEE Transactions on Signal Processing*, 55(8):3995–4003, August 2007. 2

[34] J. L. Williams, J. W. Fisher III, and A. S. Willsky. Performance guarantees for information theoretic active inference. *JMLR*, 2:620–627, 2007. 2, 6

[35] L. Yu, S. K. Yeung, Y. Tai, and S. Lin. Shading-based shape refinement of RGB-D images. In *CVPR*, pages 1415–1422, 2013. 4

[36] X. Yuan, P. Llull, X. Liao, J. Yang, G. Sapiro, D. J. Brady, and L. Carin. Low-cost compressive sensing for color video and depth. In *CVPR*, 2014. accepted. 8

[37] Y. Zhang, Z. Xiong, P. Cong, and F. Wu. Robust depth sensing with adaptive structured light illumination. *J. Vis. Comm. and Image Representation*, 25(4):649–658, 2014. 1, 2

[38] F. Zhao, J. Shin, and J. Reich. Information-driven dynamic sensor collaboration. *Signal Processing Magazine, IEEE*, 19(2):61 –72, mar 2002. 2