Information-Driven Adaptive Structured-Light Scanners

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Rosman, Guy et al. “Information-Driven Adaptive Structured-Light Scanners.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30 2016, Las Vegas, Nevada, USA, Institute of Electrical and Electronics Engineers (IEEE), December 2016: 874-883 © 2016 Institute of Electrical and Electronics Engineers (IEEE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/CVPR.2016.101">http://dx.doi.org/10.1109/CVPR.2016.101</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td>Version</td>
<td>Author's final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Mon Jan 07 10:39:42 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/111676">http://hdl.handle.net/1721.1/111676</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
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 localization 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], Simultaneous 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.

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 in-
formation 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

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.

• $A$ 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

$$p(A, G, \Theta, A_l, G_l, \eta, I_c, I_p; A) = p(\Theta) p(A) p(G) \prod_l p(A_l|A, \Theta) p(G_l|G, \Theta) \prod_{i,x} p(I_c|A_l, G_l, I_p, \eta) p(I_p|G_l, \Theta; A) p(\eta)$$

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 $R \subseteq U$ denotes the relevant set and $U$ 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 $I_c, I_p, \Theta$ 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 ($\Theta$) updates conditioned on the geometry ($G_l$) and vice-versa. Updates to the 3D map may be posed as inference of $G$ given $G_l, \Theta$. 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.

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 $\Theta$. 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

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

where $a, b$ depend on the ambient light, normals, and albedo of the incident surface. For sufficiently large photon count, $\eta$ 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 $I^A(x_R; I_C)$ for different definitions of $R$, and choices from the set $A$, alternately considering $\Theta, G$, and $A$ as the relevant variable set $x_R$. 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.

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.
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

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

aids in our understanding of the contributions of the different factors. Here, $\rho$ 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

$$I(I_c; \Theta, r) = H(I_c) - H(I_c|\Theta, r) \leq \sum_x H(I_c^x) - \sum_x H(I_c^x|\Theta, r) \triangleq \hat{I}(I_c; \Theta), \quad (4)$$

where $\hat{I}$ 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 $I$. The mutual information between $I_c$ and $G_l$ given $\theta, I_p$ is given by

$$\hat{I}(I_c; G_l|\theta) = \sum_x I(I_c(x); r(x)|\theta)$$

$$= \sum_x E_{I_c, r|\theta} [\left(\log \frac{p(I_c|r, \theta)}{p(I_c|\theta)} \right)]. \quad (5)$$

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,

$$p(I_c|\theta) = E_r[p(I_c|r, \theta)]. \quad (6)$$

For each sample of $\theta, r$, 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 $I(I_c; G_l|\theta)$. 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 $r$, $p(a, b, I^{(t)}_c | I^{(0..t-1)}_c, I^{(0..T-1)}_c)$ is Gaussian. We can compute the pdf of $a, b$ and $I^{(T)}_c$ given $I^{(0..T-1)}_c$ and $I^{(0..T-1)}_c$, by conditioning on each image $t$ at a time, computing $p(a, b, I^{(t)}_c | I^{(0..t-1)}_c)$ 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 $\Theta$ given a model of the world $G$. In visual SLAM, $G, A, A_t$ are commonly used to infer $\Theta, G_t$, 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
depth-sensor based SLAM, the range sensors obtain a measurement \( G_t \) under some active illumination. \( \Theta \) is then approximated from \( G_t, G_t \).

We now describe computation of the MI between the pose and the images. As before, we parameterize \( G_t \) by \( r(x) \), and given \((\Theta, r)\) we re-establish a correspondence between \( I_p \) and \( I_c \). This is done by computing a back-projected point \( x^3_j \) (denoting it is a 3D point), transforming it according to \( \Theta \) to get \( \bar{x}^3_j \), and projecting \( \bar{x}^3_j \) onto the camera and projector image. A similar situation would arise when inferring a class variable, where instead of merely inferring \( \Theta \) 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

\[
\mathcal{I} \left( I^3_c \mid \Theta \right)_{G_t} = E_{I_c, \Theta, r} \left( \log \frac{P(I^3_c \mid \Theta)}{P(I^3_c)} \right),
\]

(7)

where as before, \( P(I_c \mid \Theta) \) is computed by marginalization over \( r \). This procedure is detailed as Algorithm 2. When computing \( p(I^3_c \mid \Theta), p(\Theta) \) 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 \( \theta \), 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 \( \Theta \). This requires merely changing lines 4,14, in Algorithm 2 to sample a distribution over \( x^3_j(\theta, C, r) \) instead of \( x^3_j(\theta, r) \). 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-
and we show the MI gain collected over the scene, considered a ground-truth.

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. Numerical Results

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 smooth 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 have 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 silhouette of an object (the hand) in the image. A similar test was done with patterns modulated by an exponentially, radially
demonstrates a
we demonstrate pat-
we show computed per-pixel MI be-
with a synthetic case where the results are easy
in Figure
which were poorly reconstructed, as expected. As we show
the selected patterns often illuminated parts of the objects
case, representing a 40% speedup.
randomly choosing the same number of patterns. Qualitatively ,
tern sets to reconstruct the range image, and compare to ran-
the depth reconstruction error obtained by 10 random pat-
top: reconstruction with a random set of 69 bands (range
RMS=24.1mm) vs. reconstruction with the full set of 240 patterns, reconstruction squared error
as a function of the number of patterns added, 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 recon-
struction 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 be-
tween 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 experiment 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 pat-
tern 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 demonstrates a

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 ran-
dom 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.

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 added, 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 recon-
struction results with significantly fewer frames (50%), as demonstrated by
Subfigures (b) and (c).
<table>
<thead>
<tr>
<th>Step</th>
<th>Hand Mean MI, Greedy</th>
<th>STDev, Greedy</th>
<th>Mannequin Mean MI, Greedy</th>
<th>STDev, Greedy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4168</td>
<td>0.2820</td>
<td>0.1267</td>
<td>0.0957</td>
</tr>
<tr>
<td>2</td>
<td>0.7904</td>
<td>0.2803</td>
<td>0.3263</td>
<td>0.2457</td>
</tr>
<tr>
<td>3</td>
<td>0.8129</td>
<td>0.1820</td>
<td>0.2686</td>
<td>0.1694</td>
</tr>
<tr>
<td>4</td>
<td>0.6232</td>
<td>0.1125</td>
<td>0.2125</td>
<td>0.1591</td>
</tr>
<tr>
<td>5</td>
<td>0.1562</td>
<td>0.0995</td>
<td>0.0903</td>
<td>0.1317</td>
</tr>
<tr>
<td>6</td>
<td>0.0229</td>
<td>0.0264</td>
<td>0.0376</td>
<td>0.0433</td>
</tr>
</tbody>
</table>

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.

Figure 10. Left: the depth image and the MI scores of vertical and horizontal stripe masks of the patterns with respect to pose estimation in the z axis 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 corners.

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 x-y plane. The main informative areas are the cones, and regions that face the x, y directions.

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.

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


