Sparse Modeling of Shape from Structured Light

Guy Rosman  Anastasia Dubrovina  Ron Kimmel
Dept. of Computer Science
Technion - IIT
Haifa 32000 Israel
rosman|nastyad|ron@cs.technion.ac.il

Abstract

Structured light depth reconstruction is among the most commonly used methods for 3D data acquisition. Yet, in most structured light methods, modeling of the acquired scene is crude, and is executed separately from the decoding phase. Here, we bridge this gap by viewing the reconstruction process via a probabilistic model combining illumination and shape. Specifically, an alternating minimization algorithm for structured light reconstruction is presented, incorporating a sparsity-based prior for the local surface model. Integrating this 3D surface prior into a probabilistic view of the reconstruction phase results in a robust estimation of the scene depth.

We formulate and minimize reconstruction error and demonstrate performance of the algorithm on data from a structured light scanner. The results demonstrate the robustness of our algorithm to scanning artifacts under low SNR conditions and object motion.

1. Introduction

Structured light and active illumination range scanners have become an important tool for scene understanding [12, 19, 16], robotics [21, 10, 26], object modeling [8, 2], indoor scene mapping [22], and human computer interaction [32], among other tasks. The scanner usually consists of a calibrated camera-projector pair; where coded light patterns emitted by the projector are acquired by the camera and allow robust triangulation and depth reconstruction. For a review of existing structured light techniques see, for example, [29].

Many of the techniques used to reconstruct 3D depth via structured light incorporate ad-hoc assumptions on the scene structure and the 3D imaging process. These include, for instance, smoothness of the acquired surface [40, 17], or temporal objects behavior [11, 40, 17]. Yet, modeling these assumptions in a more complete way is crucial when the captured illumination patterns are of low SNR, due to long scanning range and short camera exposure times. Furthermore, such assumptions can help when dealing with motion artifacts, where some of the captured images are subject to abrupt intensity changes due to motion of depth discontinuities or albedo boundaries. Failing to model the imaging process in a realistic manner may lead to outliers in the reconstructed depth image, as is often observed in structured light scanners.

Here, we improve upon results obtained by structured light based scanners [25, 29], especially in face of challenging illumination conditions, by providing strong priors for the imaging model and surface shape. Moreover, while strong shape priors are utilized for range image correction, i.e., surface denoising and completion, the approach we suggest incorporates shape and illumination priors into the reconstruction itself, giving us a principled approach of combining powerful surface priors and probabilistic understanding of the acquisition process. Here we introduce a patch-based image similarity prior, similar to those successfully utilized for images, depth images, and surface processing [1, 7, 35, 31, 39].

2. Regularized Structured Light Model

In shape from structured light, one is attempting to reconstruct the geometric structure of the scene, by illuminating the scene with a set of projected patterns \( I_P = \{I_P^{(n)}\}_{n=1}^N \), where \( N \) is the number of patterns, and taking a set of images \( I_C = \{I_C^{(n)}\}_{n=1}^N \) of the scene using a camera. We denote the optical centers of the camera and projector by points \( C \) and \( P \) respectively. The overall setup is shown in Figure 1. In our formulation, we denote the estimated range image as \( z(x) \), where \( x \in \mathbb{R}^2 \) denotes the (two-dimensional) camera image coordinates.

We assume a Lambertian surface model for objects in the scene, and a projector emitting directional light in a temporal sequence of patterns. The main source of imaging noise is assumed to be the sensor. The lighting conditions we deal with are such that the photon count per image sensor pixel is
jector illumination component, we can model every pixel’s motion, thus requiring short exposure intervals.

This is the typical scenario in real structured light systems with decoding the coded light patterns poses a challenge. This Gaussian, yet the signal is weak enough so that correctly high enough so that the image noise model is approximately constant factors we obtain

\[
  z = \arg\max_{z,a,b} P \left( z, a, b \middle| I_P, I_C \right) \\
  = \arg\max_{z,a,b} \frac{P \left( z, a, b, I_P, I_C \right)}{P \left( I_P, I_C \right)} \\
  = \arg\max_{z,a,b} \frac{P \left( I_P, I_C, a, b \middle| z \right) P \left( z \right)}{P \left( I_P, I_C \right)} \\
  = \argmax_{z,a,b} P \left( I_P, I_C, a, b \middle| z \right) P \left( z \right) \\
  = \arg\min_{z,a,b} \left( -\log P \left( I_P, I_C, a, b \middle| z \right) - \log P \left( z \right) \right).
\]

We incorporate the maximum-likelihood choice of \(a, b\) into \(P(I_P, I_C | z)\), minimizing the negative log-probability over \(a\) and \(b\),

\[
  \min_{z,a,b} \left[ \min_{a,b} \left[ \sum_i \frac{\left(a(x)I_P^{(i)}(\Pi_z(x)) + b(x) + n^{(i)}(x)\right)^2}{\sigma^2} \right] \right].
\]

The optimal values of \(a\) and \(b\) for this least-squares fitting problem are given in analytical form by solving the normal equations using \(I_C, I_P\) at points \(x, \Pi_z(x)\) respectively,

\[
  \begin{pmatrix}
    a \\
    b
  \end{pmatrix} = \left( \frac{\mu_{\Pi_P}}{\mu_{I_P}} \right)^{-1} \left( \frac{\mu_{I_C}}{\mu_C} \right),
\]

\[
  \mu_{I_P} = \sum_i I_P^{(i)}(\Pi_z(x)), \quad \mu_C = \sum_i I_C^{(i)}(x),
\]

\[
  \mu_{I_P} = \sum_i \left(I_P^{(i)}(\Pi_z(x)) \right)^2.
\]

In order to obtain an efficient algorithm for computing and optimizing photoconsistency in the structured light case, we note that we can incorporate the computation of the maximum-likelihood expressions for \(a, b\) into a plane-sweep operation [5] when seeking the optimum value of \(z\).

Inserting the optimal \(a, b\) as a function of \(z\) and noting the conditional independence (given \(z\)) of neighboring pixel values \(I_C(x), I_P(\Pi_z(x))\) provides us with a functional to minimize with respect to \(z(x)\), similar to [34],

\[
  \arg\min_z \int_x \left[ \min_{a,b} \left( -\log P \left( I_P, I_C, a, b \middle| z(x) \right) \right) dx + \psi(z) \right].
\]

The expression \(\rho_{SL}(z; I_C, I_P, x)\) denotes a penalty for the photoconsistency assumption. This term is often optimized per pixel by several steps, including binarization of the code letters, decoding of the code, and depth reconstruction. These separate steps, however (for any specific code) are sub-optimal, even if efficient to compute.

The term \(\psi(z)\) denotes our choice for approximating the negative log-probability prior for the surface shape,
- log P(z). There are several possible choices of surface shape priors. These can incorporate either smoothness assumptions and more elaborate geometric priors, assumptions on local shape of patches on surfaces, or reasoning on natural depth image statistics [38]. We now describe two such possible regularization priors for depth images.

**Total-Variation regularization** The minimum area [4] and total-variation [28] (TV) priors, and related smoothness measures have been suggested in several forms for regularization of range images [23] and surface reconstruction [15, 33]. TV regularization for structured light can be expressed as

\[
\arg\min_z \int_x \rho_{SL}(z; I_C, I_P, x) + \|\nabla z\| dx,
\]

where \( \|\nabla z\| \) is the total variation of the range image, for some coefficient \( \tilde{c} \). This form of regularization is strongly related to MRF-based structured light [34].

**Patch-based Priors for Structured Light** Another possibility for modeling range images involves assuming a local model for each patch of the surface. Regularizing the surface then expresses itself via the parameters of this model. This includes modelling via polynomials or similar functions, leading to the moving-least-squares [18] approach, or expressing the patch via a functional basis with sparse coefficients, leading to sparsity-based regularization. Priors for depth images based on patch-estimators are described, for example, in [31, 13, 20, 35].

In our case, we assume that the depth image can be locally viewed as a sparse combination of basis functions. We note by \( \tilde{\psi}(\cdot) \) our prior for surface patches. This leads to a patch-based regularizer of the reconstruction,

\[
\arg\min_z \int_x \rho_{SL}(z; I_C, I_P, x) dx + \tilde{c}_i \sum_j \tilde{\psi}(P_j z),
\]

where \( P_j z \) denotes extraction of a small neighborhood \( i \) from the surface \( z \). For example, for an \( L_1 \)-sparse representation prior, Equation 5 becomes

\[
\arg\min_{\tilde{e}_i, \alpha_j} \int_x \rho_{SL}(z; I_C, I_P, x) + \tilde{c}_i \left( \sum_j \| P_j z - D \alpha_j \|^2 + \lambda \| \alpha_j \|_1 \right),
\]

where \( D \) denotes a dictionary for depth image patches, \( P_j \) denotes a matrix extracting block \( j \) from the image in column-stacked notation, and \( \alpha_j \) denotes the representation of patch \( P_j z \) in that dictionary.

3. Alternating Minimization Algorithm for Regularized Structured Light

We assume the coded light pattern can be reconstructed by minimizing per-pixel the decoding error function \( \rho_{SL}(x, I_C, I_P; z) \). While this reconstruction is usually obtained by binarization and decoding of the time-multiplexed code, we view it as a photoconsistency term between the structured light patterns and the resulting camera image intensities [24], when estimating the illumination conditions. Note that this function depends only on the depth value and camera intensities per pixel. In order to regularize the solution we suggest to use an alternating minimization, adding an auxiliary variable to model each patch. We decouple the problems of regularization and structured light decoding, minimizing the functional in Equation 5, which is of a half-quadratic form [9]. Minimization with respect to \( \alpha_j \) given \( z \) results in a per-patch denoising algorithm of \( P_j z \), similar to the approach taken in [14]. We now describe the different steps of the algorithm, which is given as Algorithm 1.

**Solving for \( z \)** The update of \( z \) depends on the structured light patterns, and may not even be continuous. Since the similarity term relating \( z \) and \( D \alpha_j \) is quadratic, we can rewrite the term for each pixel \( x \) in \( z \) as the sum of a photoconsistency measure and a sum of squared distances from versions of \( z(x) \) in all of the patches containing this pixel, with an aggregate weight \( w(x) \),

\[
z^{n+1}(x) = \arg\min_z \rho_{SL}(z) + \tilde{c}_i w(x) \| z - \tilde{z}(x) \|^2.
\]

A solution can be obtained by sweeping the set of possible \( z \) values, similar to stereo [5]. Doing this plane-sweep is highly suitable for parallel implementation on graphics processing units (GPUs) [37]. Note that plane-sweeps are discrete by nature, as are the coded patterns in many cases. In order to obtain convergence, and allow sub-pixel precision, we minimize a linearly-interpolated photoconsistency, along with the quadratic distance in the second term of Equation 9. The depth estimated at each pixel is set according to the minimum of the interpolated cost function.

**Solving for \( \alpha_j \)** Given a patch estimate \( P_j z \), an update of the patch resorts to a standard sparse representation problem. Specifically, if we take our sparse prior to be of an \( L_1 \) type, we can update \( \alpha_j \) using iterative shrinkage [3],

\[
\alpha_j^{n+1} = S_M(\alpha_j^n - 2t D^T (D \alpha_j^n - P_j z)),
\]

where \( t \) is a gradient descent step, chosen to be small enough, and \( S_M(\cdot) \) denotes the soft shrinkage operator,

\[
S_M(y) = \begin{cases} 
0, & |y| \leq \lambda \\
y - \lambda, & y > \lambda \\
y + \lambda, & y < -\lambda
\end{cases}
\]

While faster iterative methods exist for \( L_1 \) minimization (see [36] for a few examples), because of the alternating minimization nature of our scheme, more complex steps may not lead to faster convergence. We therefore chose to use the original iterative shrinkage scheme.
Algorithm 1 Alternating Minimization Sparse Structured Light

1: for \( k = 1, 2, \ldots, \) until convergence do
2: \quad Update \( \alpha_j^k(x) \) for all \( j \), according to Equation (10).
3: \quad Update \( z^k(x) \), according to Equation (9).
4: end for

3.1. Learning a Sparse Depth Prior

In order to learn a surface model from range images, several properties of the data must be taken into account. Since reconstruction errors are of an outlier nature, algorithms such as KSVD [7] that assume an additive white Gaussian noise model require pre-processing and outlier removal. Furthermore, since many of the patches in range scans are of smooth surfaces, and since the KSVD algorithm is initialization-dependent, care must be taken to provide a diversified initial dictionary. We focus the algorithm on the scarcer edge patches by clustering the data first using the mean-shift algorithm [6]. The resulting dictionary obtained from a set of 50 range scans is shown in Figure 2. We note that the examples used for testing are not part of this dataset, avoiding overfitting for a specific subject. While the training data is from a specific class of human faces, the learned primitives are quite general, as can be seen in Figure 2. We leave the effect of different dictionary and training data choices for future research.

![Figure 2. An example of the dictionary of 300 words obtained from a set of 50 range scans.](image)

4. Results

In order to test the proposed scheme, we use a standard structured light setup similar to [30], with 10 striped patterns, along with an all-ones and all-zeros pattern. The camera images are sampled at 320 \( \times \) 240, and projector patterns are shot using a 1024 \( \times \) 768 DLP projector. In order to simulate low-SNR conditions, we have added Gaussian noise to the camera images before reconstruction. Results are shown in Figure 3 for the case of structured-light images with intensity Gaussian noise of standard deviations 5 and 10.

In order to quantitatively validate our method, we take as ground truth an almost-noiseless range image of the head statue, and measure range errors compare to it. We compare both \( L_1 \) and robustified \( L_2 \), truncated at 10 millimeters, and compared to median post-processing, taken with the smallest filter size that removed range outliers from the face, in order to avoid oversmoothing. The results of this comparison are given in Table 1. For 320 \( \times \) 240 images, the dictionary trained was of patch size 8 \( \times \) 8.

We compare our results to several approaches. A common way of removing reconstruction artifacts is by median filtering, as was done in [27]. Yet another approach treats the problem as a denoising problem with a strong prior and impulse noise assumption. An example of this type of method would be to take the same depth prior we use, but solve a denoising problem with an \( L_1 \) fidelity term.

\[
\arg\min_z \int_x \| z - \tilde{z}_0 \| dx + \tilde{c}_1 \sum_j \psi(P_j z), \quad (12)
\]

where \( \tilde{z}_0 \) is the reconstruction results without a prior. This approach would be similar, in a sense, to the depth image denoising suggested in [35]. This approach is marked in Table 1 under the Sparse Denoise column. In addition, it would be interesting to try a weaker prior for reconstruction such as TV regularization as suggested in Section 3. This approach is shown in the table as column TV. For all of the methods, parameters were chosen so as to obtain optimal robust \( L_2 \) results, while preventing remaining depth outliers. The table demonstrates the effectiveness of the proposed algorithm. While the computational cost of our algorithm is quite high with current Matlab code, the algorithm is highly parallelizable and one future line of work involves fast parallel implementation of this algorithm.

In Figure 4 we demonstrate the results of our algorithm on artifacts caused by head motion in the vertical direction. Even though the assumption of constant \( a(x), b(x) \) breaks, the algorithm overcomes many of the errors caused by reconstruction followed by outlier removal. The size of the median filter is chosen to be the smallest size that filters the motion artifacts over the eyes and mouth regions, a 7 \( \times \) 7 filter in this case. We note that at this filter size, the mouth and nose areas merge, while artifacts remain on the eyelids.

5. Conclusions

In this paper we presented a novel model for regularized structured light reconstruction. Incorporating a sparse surface prior into a physically-motivated probabilistic outlook...
on structured light decoding, we demonstrate accurate results in scenarios where the usual approach for decoding structured light tends to fail.

The results obtained merit the coupling of a strong surface prior with a probabilistic model for structured light reconstruction, and motivate further exploration of the benefits of the proposed method as well as investigating the use of this approach for different types of depth scanners. An additional line of work involves implementing the current algorithm in an efficient manner, exploiting the high level of parallelism available in each phase.

Acknowledgment

This research was supported by the European Community’s FP7-ERC program, grant agreement no. 267414. The authors would like to thank Mr. Alon Zvirin for work with the structured light scanner.

References


Figure 3. First row, left-to-right: An example textured pattern, reconstruction results, reconstruction with median filtering, reconstruction with sparse prior, where camera images were added Gaussian noise with standard deviation of 5, with close-up on the right eye region and the nose and mouth region. Second row, left-to-right: ground-truth reconstruction obtained from noiseless reconstruction, same sequence of results, where camera images were added Gaussian noise with standard deviation of 10. Third row, left-to-right: 3D raw reconstruction results, reconstruction with median post-processing and with a sparse prior for the case of $\sigma = 5$ noise. Fourth row, left-to-right: (3D raw reconstruction omitted since it was too noisy), reconstruction with median post-processing and with a sparse prior for the case of $\sigma = 10$ noise. In order to view the range images, color and/or online viewing is suggested.
Figure 4. Left-to-right: An example with artifacts caused by vertical head motion, a median-filtered result, the result of the proposed method. Note the merging of the mouth and nose area in the median filter, and the remaining artifacts around the left eye and nose area.