Multi-Modal Depth Sensing

Robert Marcus
Carnegie Mellon University
rbm@cmu.edu

Abstract

There are multitude of ways to compute a target objects depth by imaging. These methods, taken on their own, typically suffer some variety of drawbacks, ranging from low accuracy to high cost. Taken together, however, these methods can alleviate or greatly reduce the others deficiencies. We review an approach for combining LIDAR measured depth information with normals recovered by two methods - shape from polarization, and photometric stereo. Additional discussion is dedicated to a high level overview of methods for improving shape from polarization, and why these adjustments are merited.

1. Introduction

1.1. Problem Background

There are various methods for extracting the depth information of target objects. A number of methods have gained prominence in recent years. Some of these *Shape-From-x* methods can be summarized briefly as follows:

1. Structured light

   Pros: Fast, high fidelity, precise.

   Cons: Expensive, useful implementations may be impractically large.

2. Photometric stereo

   Pros: Allows re-rendering of an image, recovers additional useful properties.

   Cons: Requires many exposures, not useful for many materials, accuracy can be lacking.

3. Stereo-vision

   Pros: Fast, uses commercial hardware

   Cons: High touch setup, scales poorly, accuracy is constrained by size of setup.

4. Polarization

   Pros: Potentially fast, works with a variety of materials and lighting conditions

   Cons: Significant ambiguities during recovery lead to low accuracy.

On their own, each of these methods suffer from significant drawbacks. In general, it would be fair to say that *shape-from-* stereo and polarization methods would benefit from the high precision of the structured light approach, and the structured light method would benefit from the compact size and economy of scale of the former. It has been shown, in fact, that while the underlying properties of each method cannot be reduced, by combining the methods at various points in the process of reconstructing an objects shape, superior results can be achieved than by each method on their own.

That is, to say, we can use smaller, lower quality, and cheaper imaging solutions to achieve superior results by combining these multiple modes of sensing to build our reconstruction. This has the effect of mitigating specific drawbacks of individual methods on their own.

For example, say we wanted to construct a device for scanning vases of variable material. We could, for example, choose to use a structured light approach with a laser scanner, i.e., an Intel L515 LiDAR. However, at its best, the L515 has a 5mm spatial resolution, which may be insufficiently *accurate* for the task. Alternatively, we could try using photometric stereo to achieve our goal, however, due to material variability, lighting imperfections, etc., this approach may be insufficiently *precise* without post-processing. The same could be said, for example, by undertaking a shape from polarization approach.

It would be convenient if we could combine the methods in such a way that we could keep the high accuracy of the visual methods with the high precision of the structured light method.

1.2. Problem Approach

It has been shown by Nehab Et Al that we can perform this operation on aligned datasets [2]. Specifically, given we have the normals of an object, and a depth map aligned

to the perspective of those normals, we can combine them in such a way that combines the high-frequency details of the visual image with the precision of depth maps.

In other words, for a given scene, we can loosely state an approach to this problem as follows:

1. Collect our depth map by structured light (LIDAR, TOF, etc.)
2. Collect our images by some procedure[1].
3. Align the requisite data.
4. Construct our normals and pre-process to correct for ambiguities

   In the case of shape from polarization, we can involve the depth map to constrain the ambiguities in the normals, as shown by Kadambi Et Al [1].
5. Construct a new depth map by combining the depth map and normals following the procedure outlined by Nehab et al, giving the depth map weight λ, and the normals weight $1 - \lambda$.

We will discuss each step in the following methods sections, preceded by giving a minor recap of uncalibrated photometric stereo, as well as a gentle introduction to shape from polarization.

2. Method

2.1. Collecting and aligning the depth map and image frames.

This project was centered around the usage of an Intel L515 Realsense depth sensing camera. It features a LIDAR sensor which can be used to stream depth and color RGB information simultaneously. Interfacing with the unit is handled by the PyRealsense2 Python API[2].

The setup for this experiment featured placing the L515 on a tripod, approximately one meter away from the target scene, where the units spatial resolution is maximized (about 5mm.) We are able to mask out the scenes foreground and background using the devices depth range to build a binary mask of all pixels within a certain range.

Additionally, the device camera and LIDAR sensor were aligned using the cameras intrinsic information, which is made available through the PyRealsense2 API. See figure 1 at the end of the document for a reference of the setup.

One limitation of this method is that the L515 is not natively capable of capturing RAW image information [3]. This

[1] In this paper, we consider photometric stereo and shape from polarization.

[2] Note, as of writing this, the PyRealsense2 library is essentially only usable on Windows. Installing this on MacOS, and, especially ARM MacOS, is a futile exercise.

[3] It is, interestingly, capable of capturing raw depth map information.

Figure 1: The experimental setup. When dealing with SfP, a 55m circularly polarized filter is placed on the front surface of unit.

presents a challenge for photometric stereo, as preferably we would pass in linearized `tiff` images. We settle by using linearized PNG's, where each image has been gamma-corrected by the following map:

$$I_{\text{lin}}(I_{\text{nl}}) = \begin{cases} \frac{I_{\text{nl}}}{12.92} & I_{\text{nl}} \leq 0.0404482 \\ \left(\frac{I_{\text{nl}}+0.055}{1.055}\right)^{2.4} & \text{o.w} \end{cases}$$

2.2. Constructing normals

2.2.1 Constructing normals by uncalibrated photometric stereo

For context on this approach, see homework 5 from 15-463.

2.2.2 Constructing normals by SfP

There are several different approaches for constructing normals by SfP. We follow the approach laid out by Smith Et al in *Linear depth estimation from an uncalibrated, monocular polarisation image*, which makes the assumptions that the ambient light is unpolarized, the materials are dielectric, and surfaces are specular dominant. We have tried our best to follow these assumptions with our scene, using what we believe to be a polarized filter in between the scene and the sole point light source.

We additionally assume a fixed refraction index of $n = 1.5$ for simplicity, and because it is approximately average for the given restrictions.

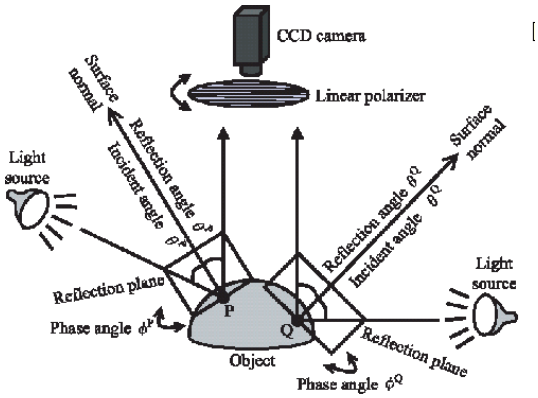

Figure 2: The basic scheme of SfP. It becomes clear how we can reconstruct the normals by going from spherical coordinates to cartesian.

We will now establish the math needed to go from raw images to normals, which is fairly straightforward to lay out from a high level[4] The two terms we would like are the zenith (θ, the reflection angle off the surface) and azimuth (ϕ, the phase angle of the reflected ray.)

We can parametrize every pixel in the image by three quantities, its phase ϕ, the degree of polarization ρ, and the unpolarized intensity, i_{un}.

Principally, we must capture multiple images of the scene to parameterize the sinusoid characterizing the intensity of the pixels across filter rotations. We calculated the phase angle using a non-linear curve fitting approach for a given pixel $p_{x,y}$ across the various filtering rotations.

Additionally, the degree of polarization ϕ can be taken as $\phi = \frac{I_{\max} - I_{min}}{I_{\max} + I_{\min}}$, and our unpolarized intensity i_{un} can be taken as $i_{\text{un}} = \frac{I_{\max} + I_{\min}}{2}$[5]

We can relate the zenith θ to the degree of polarization by Fresnels equations, namely,

$$\rho = \frac{(n - \frac{1}{n})^2 \sin\theta}{4\cos\theta\sqrt{n^2 - \sin^2\theta} - \sin^2\theta(n + \frac{1}{n})^2 + 2n^2 + 2}$$

Where ρ, n are both known, and we can solve for θ by any number of manners. We choose to solve for it directly using a procedure outlined by Kadambi [1]. Note though that this is where a key ambiguity is introduced in SfP - both θ, and $\theta + \pi$ are solution to the above. We can correct for this ambiguity in practice, either by disamb

By recovering ϕ, θ, we can then translate our parametrization into a cartesian normals by

$$n = \begin{bmatrix} \sin\phi\sin\theta \\ \cos\phi\sin\theta \\ \cos\theta \end{bmatrix}$$

The above procedure should theoretically yield the relevant surface normals. However, in practice, these efforts are beguiled by three physical constraints, as noted in Kadambi [1]:

1. Azimuthal ambiguity (as noted above)
2. Refractive unknowns

   the $n = 1.5$ assumption is inexact and leads to inaccuracies. It is rarely practical to know the refractive index at every pixel in the scene.
3. Fronto parallel surfaces

   When the zenith angle is near zero, the obtained normals are noisy[6]

2.3. Combining positions and normals

Nehab et al formulated the question of finding an optimal surface in terms of a minimization problem regarding two error terms - that of the depth measurements, E^p, and the normal measurements, E^n: $\text{argmin}_Z \lambda E^p + (1-\lambda)E^n$. Here, $\lambda \in [0, 1]$ is a weight controlling how much of the fused output should consist of the normals versus the depth.

Nehab provides a formulation of the minimization problem that relies on surface tangents of the depth map. These tangents can be written as linear in terms of the depth. We rely on Yu et al [5] for an explicit formulation of this minimization problem, whereby our tangent vectors can rather artfully be laid out as follows:

$$T_x = \begin{bmatrix} -\frac{1}{f_x}X & -\frac{1}{f_x}I \\ -\frac{1}{f_y}Y & 0 \\ I & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial Z}{\partial x} \\ I \end{bmatrix} \quad T_Y = \begin{bmatrix} -\frac{1}{f_x}X & 0 \\ -\frac{1}{f_y}Y & -\frac{1}{f_y}I \\ I & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial Z}{\partial y} \\ I \end{bmatrix}$$

Where X, Y form a meshgrid over the $N \times M$ size image, and $I, 0$ are the relevant matrices of the same size. As constructed in code, these form a highly sparse matrix of size $NM \times NM$. Additionally, f_x, f_y are the respective axis focal lengths of the camera in pixels, retrieved with the other camera intrinsics during the alignment step.

Each pixel in the image yields 3 equations or less: one equation for the position error. This allows us to solve the system by least squares as it is highly overconstrained. To that end, we similarly build a very sparse representation of the normals so we can incorporate them into the minimization, building them as the $NM \times 3NM$ matrix containing

[4] Which is not to say it is easy to implement in practical terms.

[5] The unpolarized intensities can be used to enforce additional surface normal constraints by assuming light is reflected according to the Lambertian model. We do not implement this constraint, but the value is explicated for consistency with general SfP explanations.

[6] This is evident in my data, in figure 3 and 4, with the foam roller.

the channelwise normals. We are left to solve the over-constrained least squares problem of $\begin{bmatrix} \lambda I \\ NT_x \\ NT_y \end{bmatrix} \hat{z} = \begin{bmatrix} z \\ 0 \\ 0 \end{bmatrix}$ where the left hand side has shape $3NM \times NM$ z is our flattened depth map of size $1 \times NM$, and each 0 is the $1 \times NM$-sized 0-vector. This can be solved rather easily, and gives us the NM-length vector $\hat{z}$ containing the optimal depth surface. This procedure works as the surface is effectively minimized by the tangents being made perpendicular to the corrected normals, i.e., their dot product is zero.

3. Results

We captured four sets of data of two Lambertian objects for this data. Two sets of polarized data, where 4 frames were captured of the object with a linear polarizer in front of the L515 unit, rotated at the angles $0°, 45°, 90°, 135°$[7].

See figure 3 and 4 at the end of the document for reference.

Also provided are the recovered depolarized light intensities, the azimuth angles, and the zenith angles. Note these have not been processed for ambiguities as described by Kadambi [1] (with respect to correcting azimuthal ambiguity, refractive distortions, or enforcing physics constraints through integration) or Atkinson [4] (with respect to azimuthal disambiguation.)

4. Discussion

The results from the photometric stereo data was excellent. The improvements are visually discernible on both datasets, with no obvious artifacts or degradations on the resultant depth maps. While the specific improvements are impossible to qualify because we lack a true ground source truth for the improvements, the results are certainly obvious in this context - in both cases, information that was entirely missing from the orignal depth map has now been incorporated, without distorting the map or introducing additional ambiguities. The results are impressive.

However, this project was originally not focused on photometric stereo, and it's now obvious to the reader as to why the project has been extended to include it - the results with SfP are less than stellar. While the initial plan for this project included going far beyond naive SfP, unexpected friction was encountered from the very beginning. These headwinds meant that instead of delivering an optimized SfP with ambiguity improvements, instead we primarily have a slightly buggy naive SfP that isn't quite right. Notice the white contour lines on the conch or foam roller, for example.

These contour lines correspond to sharp drops in the depth map that are certainly not present in the real scene.

While the photometric stereo approach improved the depth map without introducing artifacts, the SfP approach introduced artifacting to the depth map, for seemingly every value of λ. This is indicative of a normal map that is wrong in a kind of way.

There are several possible explanations for this. One is that the data was improperly collected, which is possible - the angle calibrations on the circular filter were not exact, and, while we tried our best to place the filter at the right angular orientation for each capture, we could not have been exact. Additionally, the L515 was not designed for use with a polarizing filter, so it's possible the additional glass between the built in camera introduced too much additional distortion for the results to ever be meaningful. Finally, there is always the possibility for buggy code. Part of the issue here is that there are many different implements of "basic" normal recovery by SfP, each with their own minute differences in material and lighting assumption. This paper took the approach of Smith et al 2016 with *Linear Depth Estimation from an Uncalibrated, Monocular Polarisation Image* [3] as discussed in the method section.

It's possible that we would have been better served by an approach that made fewer assumptions about the scene. However, we likely would have been stymied by the above physical limitations with respect to equipment either way.

Additionally, even with a more successful SfP implementation, forward progress on resolving SfP ambiguities was somewhat unlikely. Kadambi undersells the difficulty of his approach, at least for those replicating his approach in modern Python - while there have historical been graphical model libraries in python, they are not readily available anymore, or we have not found the right libraries[8]. Similarly, when we tried implementing Atkinson's [4] disambiguation protocol in an attempt to improve the normals, we were stymied by debugging difficulties. Implementing the algorithm was difficult for all the normal reasons: the paper was less explicit than desired, the algorithm was quite slow, and debugging was difficult because it was hard to tell if the problem was bad data or bad design. In general, experiments with SfP are not something the author would recommend undertaking in the future, and, perhaps they should have heeded Kadambi's warning on the practicality of SfP without additional processing steps: "Due to the limitations of SfP, SfP has never been considered as a robust alternative to SfS or PS."

In other words, the originally scoped project had a far greater all-or-nothing element to it than anticipated.

[7] The polarizer angle was calibrated to $0°$ by rotating it in front of a polarizing light source until the light passing through the filter was at its brightest. A ring of paper was taped around the filter and the relevant points were marked on the filter in this manner.

[8] For example, we came across old sources for the OpenGM library while designing an approach to handle section 4 of Kadambi, but were unable to build the library successfully after a day of trying. This meant large parts of Kadambi were essentially impractical to re-implement without significant additional effort.

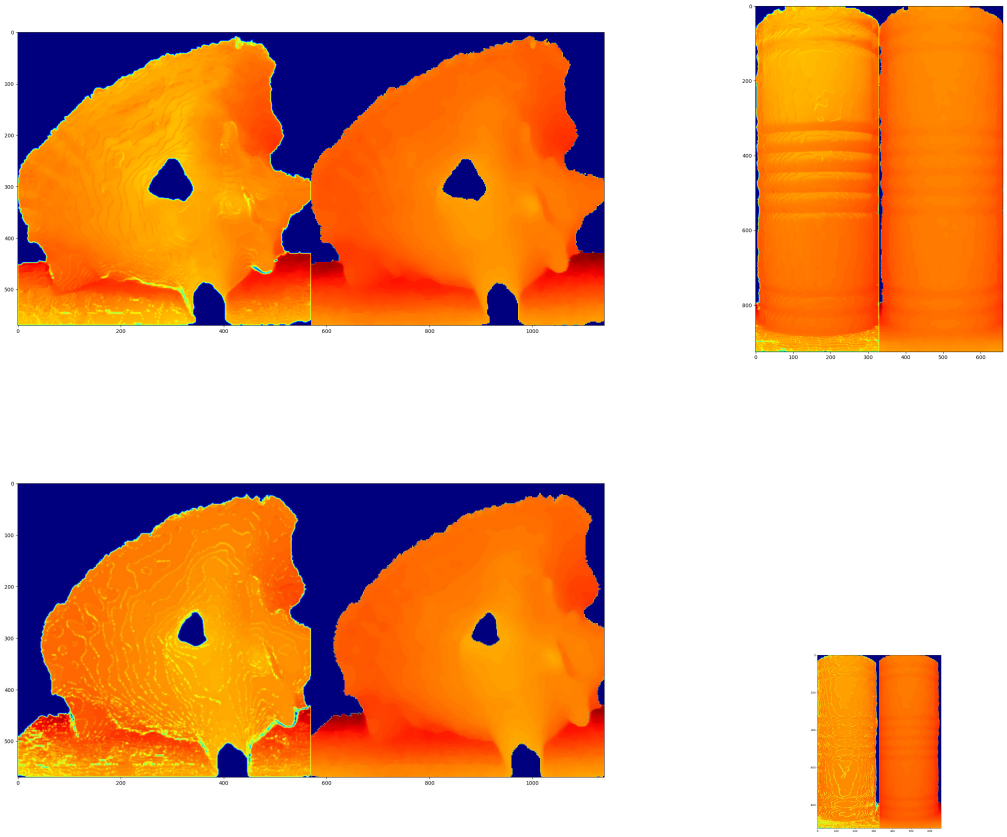

Figure 3: The improvements from the merging algorithm on the left versus the original depth maps on the right. On the top, results from the photometric stereo normals. On the bottom, results from the polarization normals.

Regardless, the results from re-implementing Nehab were quite good, and, given the limitations of naive SfP as discussed in methods section, the given results were fairly OK.

5. Material

Data for this project is available at the following github repository. Additionally, almost all of the code for this project is available there as well, however, the uncalibrated photometric stereo code has not been released. The code largely contains citations where relevant, but some may be missing due to accidental oversight. See the relevant `*_runner.py` code for usage and examples.

References

[1] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1-3):34–51, June 2017. 2, 3, 4

[2] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH 2005)*, 24(3), Aug. 2005. 1

[3] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. Linear depth estimation from an uncalibrated, monocular polarisation image. pages 109–125, 2016. 4

[4] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2875–2888, 2019. 4

[5] Ye Yu and William Smith. Depth estimation meets inverse rendering for single image novel view synthesis. *CVMP '19: European Conference on Visual Media Production*, pages 1–7, 12 2019. 3

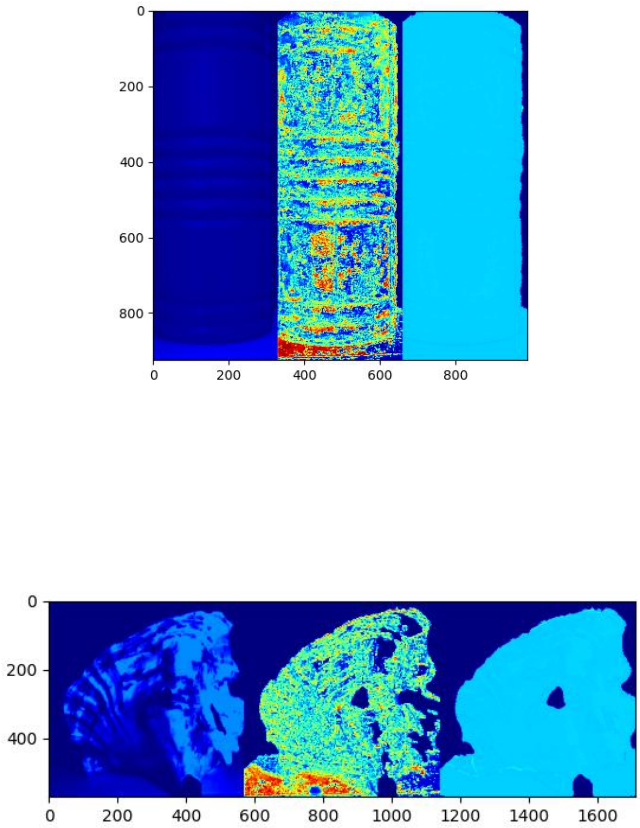

Figure 4: The intensities, the azimuth angles, and the zenith angles of the polarized photos.

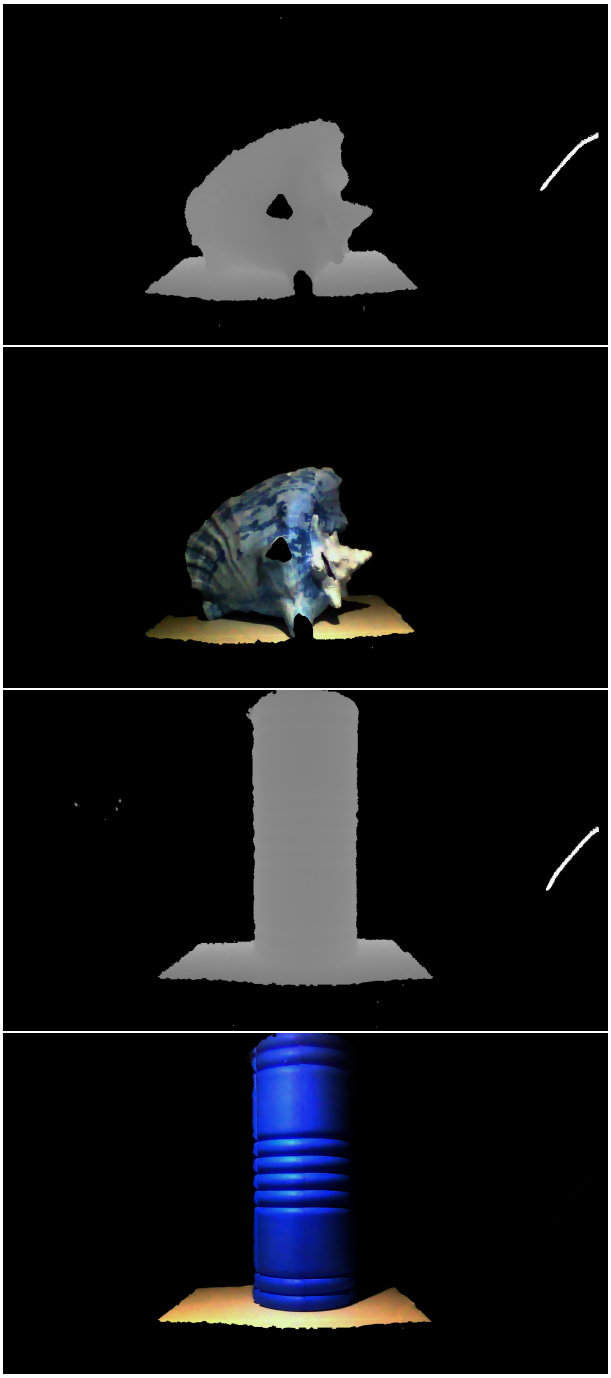

Figure 5: Some sample raw frames retrieved from the intel L515.