Title

Fast Single-Frequency Time-of-Flight Range Imaging

Authors

Crabb, Ryan
Manduchi, Roberto

Publication Date

2015-06-01

Peer reviewed

Fast Single-Frequency Time-of-Flight Range Imaging

Ryan Crabb
Dept. of Computer Engineering
University of California, Santa Cruz
rcrabb@soe.ucsc.edu

Roberto Manduchi
Dept. of Computer Engineering
University of California, Santa Cruz
manduchi@soe.ucsc.edu

Abstract

This paper proposes a solution to the 2-D phase unwrapping problem, inherent to time-of-flight range sensing technology due to the cyclic nature of phase. Our method uses a single frequency capture period to improve frame rate and decrease the presence of motion artifacts encountered in multiple frequency solutions. We present an illumination model that considers intensity image and estimates of the surface normal in addition to the phase image. Considering the number of phase wrap as the 'label', the likelihood of each label is estimated at each pixel, and support for the labeling is shared between pixels throughout the image by Non-Local Cost Aggregation. Comparative experimental results confirm the effectiveness of the proposed approach.

1. Introduction

Range imaging is becoming an essential component in applications such as automotive, augmented reality, natural user interface, biometric, computational photography, and more. There are currently three main methods for range imaging: stereo or multi-camera triangulation; triangulation from projected patterns (structured light); and time of flight (ToF) measurements. This contribution focuses on ToF technology, which has already been implemented in multiple products such as Microsoft Kinect 2, PMD, Intel Real Sense.

ToF cameras are comprised of an illuminator producing modulated infrared light, and an imaging sensor that is synchronized with the illuminator. The imaging sensor computes the phase difference (or shift) Θ between the transmitted and received wavefronts, along with the intensity (magnitude) B of the received light (irradiance)[1]. Light reflected by a surface at distance D has a phase shift equal to

$$\Phi = (4\pi f_m D/c) \bmod 2\pi \qquad (1)$$

where f_m is the modulation frequency of the illuminator, and c is the speed of light. The distance (range) to the surface can thus be recovered from Φ, but only up to a multiple of the *wrapping distance* $D_w = c/2f_m$ or *unambiguous range*. This is the well known *phase unwrapping* problem of ToF sensing. To deal with this problem, most commercial cameras use the so-called dual-frequency approach: two images are taken of the same scene, where the illuminator is modulated at different frequencies f_m in the two images. By analyzing the two phase shift images, the distance D to each visible surface element can be uniquely recovered.

Dual-frequency phase unwrapping assumes that the scene has not changed between the two images. If this is not the case (e.g., if the camera is moving), the result is not accurate. In fact, it has been shown that dual-frequency is *not* necessary for phase unwrapping. A number of techniques have been demonstrated that leverage knowledge of the measured intensity B to recover the "unwrapped" distance D from a *single* image. These techniques rely on the observation that the intensity B of light reflected by a Lambertian (opaque) surface is inversely proportional to the square of the surface distance D – suggesting that some information about D could be inferred from B. Unfortunately, the intensity B is also affected by the (unknown) albedo ρ and slant angle β of the surface. To reduce the uncertainty of inference, it is thus necessary to impose additional constraints, such as spatial smoothness priors, typically expressed in the form of a Markov Random Field (MRF). While this approach has produced impressive results [2], it requires use of techniques such as Belief Propagation or graph cuts, which are computational demanding and preclude frame-rate processing.

In this paper, we report progress toward the development of a computational single-frequency ToF camera that solves the phase unwrapping problem with accuracy comparable to or exceeding the state of the art, and speed that, at 0.5 seconds per frame, is orders of magnitude faster than previous approaches. Specifically, this work presents two main contributions. First, we show that the surface slant angle β can be computed to a good approximation even *before* unwrapping the phase shift. Knowledge of β reduce the uncertainty in the determination of D. Second, we impose a smoothness prior using the fast Non-Local Cost Aggregation algorithm, which was recently proposed for stereo matching. This non-iterative algorithm is extremely efficient, requiring only a few operations per pixel per wrap number.

This article is organized as follows: Section 2 provides an overview of related work and current solutions; in 3 we formally define the problem, present our contributions, and describe the method; Section 4 provides our results in comparison to similar methods and includes an analysis of the components of our method and the variations within.

2. Related Work

In time-of-flight imaging, each observation $\mathrm{O} = \{\Phi, \mathrm{I}\}$ is initially composed of four distinct intensity images $\{\mathrm{I}_0, \mathrm{I}_{\pi/2}, \mathrm{I}_\pi, \mathrm{I}_{3\pi/2}\}$ captured at specific time intervals with respect to the illumination modulation cycle ,each staggered by $\pi/2$ with respect to the modulation phase. As described in more complete detail in [2],we take the difference of image pairs offset by π, $\mathrm{X}=\mathrm{I}_\pi-\mathrm{I}_0$, and $\mathrm{Y}= \mathrm{I}_{3\pi/2}- \mathrm{I}_{\pi/2}$, and with simple computation can recover the phase offset from the illumination modulation $\Phi = \arctan(-\mathrm{Y}/X)$, and the active scene illumination $\mathrm{I} = \sqrt{X^2 + Y^2}/2$.

The classic phase unwrapping solutions [3] use path integration to recover the unwrapped phase. These rely solely on the notion of spatial coherence, that we expect local areas to be smooth. These solutions often have inconsistencies called *residues*, and may only be known up to a constant. These and other methods are discussed comprehensively in [3]. In the solution proposed in this paper, the wrap number at each pixel is estimated absolutely and directly, and the possibility of residues is eliminated.

As the limited unambiguous range is ultimately the problem that necessitates phase unwrapping, then the problem could be circumvented by simply increasing the unambiguous range. This could be accomplished by decreasing the modulation frequency, though at the cost of precision, as the measurement uncertainty increases proportionally with the unambiguous range [4]. The use of multiple frequencies in the signal modulation, as adapted from INSAR technology [5], is a clever way increase the unambiguous range. For time-of-flight technologies it is currently regarded as the de facto solution [6][7][8]. The principle is that if two or more frequencies are used in successive frames of a static scene, then the set of differently wrapped phases can be used to calculate a larger unambiguous range. The different frequencies act together as a single, lower *effective frequency*. Of course, capturing multiple successive signals requires more time and power for each single frame. Depending on the particular task, especially those involving rapidly moving objects (say, gesture recognition), this may be especially troublesome.

Other single frequency solutions, such as the one proposed in this paper, incorporate additional information to the wrapped phase. Beder et al. [9] combine ToF data with a traditional stereo camera pair to optimize local "patchlets" of the surface. Similarly, Gudmundsson et al. [10] bootstrap a stereo pair with ToF data. Choi and Lee [11] skip the traditional cameras and directly pair ToF cameras. Using a Markov random field, several others [2][12][13][14] fuse the data under a probabilistic framework.

In techniques more closely related to that presented here, the intensity image intrinsic to ToF cameras is exploited, along with the phase data. The shape from shading approach inspired Böhme et al. to smooth noisy ToF depth images [15], however an already correctly unwrapped phase is assumed. McClure et al. [16] use depth edges to initially segment the scene before analyzing each segment to determine whether it falls into the unambiguous range based on average intensity, using a manually set threshold. Choi et al. [17], using a corrected intensity image [18], apply an EM optimization to classify each pixel as being within or outside the unambiguous range, which feeds the data term for a Markov random field optimization by graph cuts.

3. Method

As mentioned in the Introduction, the sensor measures the "wrapped" phase shift Φ_p at each pixel, where the subscript indicates the pixel index. Our goal is to recover Θ_p, the actual ("unwrapped") phase difference, from which the distance to the surface D_p is obtained as by $D_p = c \cdot \Theta_p/4\pi f_m$. Θ_p and Φ_p are related as by $\Theta_p = \Phi_p + 2K_p\pi$, where K_p is the (unobservable) "wrap number". Our goal is to compute the wrap number K_p at each pixel, using the observed data: wrapped phase shift Φ and intensity B.

3.1. Intensity Model

Light that reaches the imaging sensor can be attributed to at least four sources: light from the illuminator reflected directly from the observed surface (*direct reflection*); light from ambient sources reflected off of the surface (*ambient*); light from the illuminator reflected indirectly from intermediate surfaces in the area (*multipath reflection*); and light from surfaces other than that directly observed, caused by lens defects or other unintended sources (*stray light*). In this paper we only consider the direct reflection, which gives the greatest contribution to the measured intensity. (Note that ambient light, which is usually approximately constant in time, is filtered out for the most part during phase shift computation.)

We assume that the visible surfaces are Lambertian (which also means that we expect incorrect results in areas with high specular reflection). The illuminator is mounted on the camera itself, as close to the lens as possible. It is modeled as a point light source, located at the camera's optical center. If the illuminator were an ideal isotropic point source, then the irradiance I_p at a pixel would be related to the (constant) radiant intensity L from the point source as by $B_p = L \cdot \rho_p \cdot \cos\beta_p/D_p^2$, where ρ_p is the albedo of the surface element imaged by pixel p, and β_p is its slant angle (the angle between the surface normal and the line of sight). In practice, the radiant intensity (that is, the light power emitted per solid angle) is not uniform, nor is the pixel sensitivity to light arriving from multiple directions (due to multiple reasons, including the effect of optical elements). This non-uniformity can be calibrated *a priori*, resulting in a distribution L_p of equivalent radiant intensity (or *light profile*, shown in Figure 1). This allows us to specify the general model of observed intensity B_p at pixel p as

$$B_p = \frac{L_p \cdot \rho_p \cdot \cos\beta_p}{D_p^2} \qquad (2)$$

This expression for the measured intensity was used in [2] to derive the conditional likelihood of B_p given the distance D_p under the assumption that the albedo and the surface orientation are uniformly distributed random variables. In fact, we observe that the surface normal could, to some approximation, be computed from the wrapped data Φ. More specifically, suppose one reconstructs a surface patch from distances computed from the measured wrapped phase Φ, assuming a constant wrap number K. A generic 3-D surface point $P_p^{(K)}$ in this surface has distance $D_p^{(K)}=c \cdot (\Phi_p + 2K\pi)/4\pi f_m$. The normal $N_p^{(K)}$ to the so computed surface patch at $P_p^{(K)}$ can, in first approximation, be considered constant with K (note that this approximation is only exact if p is the principal point, that is, the pixel along the optical axis). Hence, we can approximate the slant angle β_p of the actual surface patch imaged by p by the slant angle of the reconstructed patch for a fixed K (e.g., K=0). Note that this approximation fails catastrophically if the actual surface patch is at a distance that is a multiple of c/f_m, that is, exactly where the phase shift undergoes a wrap. An example of reconstructed surface orientation is shown in Figure 1.

Based on our estimation of the slant angle β_p, and neglecting sensor noise, we can use (2) to compute the conditional likelihood $P(B_p|D_p,\beta_p)$ with a proper prior distribution of the albedo ρ_p. We can then use Bayes' formula to compute the conditional posterior probability of the distance D_p given B_p and β_p:

$$\begin{aligned} P(D_p|B_p,\beta_p) &\propto P(B_p|D_p,\beta_p) \cdot P(D_p|\beta_p) \qquad (3) \\ &\propto P(B_p|D_p,\beta_p) \end{aligned}$$

where the last identity derives from the fact that the surface orientation is statistically independent of its distance, combined with a uniform prior distribution of surface distances. Note that this expression for the posterior distribution of D_p does *not* take into account the measured phase shift Φ_p. In fact, we can combine the two to obtain a (marginal) posterior probability distribution of the wrap number K_p as by:

$$P(K_p|\Phi_p, B_p, \beta_p) = P(D_p|, B_p, \beta_p) \qquad (4)$$

with $D_p = c \cdot (\Phi_p + 2K_p\pi)/4\pi f_m$.

If the slant angle β_p is assumed known, and assuming a prior uniform distribution for the unknown albedo ρ_p, it is easy to see that $P(B_p|D_p,\beta_p)$ is also uniformly distributed between 0 and $L_p \cdot \cos\beta_p/D_p^2$. However, we noted that the estimation of β_p is often inaccurate, in part because, as mentioned earlier, the actual surface normal depends on the (unknown) wrap number K_p, in part because surface normal estimation is notoriously noisy. Hence, rather than assuming a fixed value for β_p, we model the slant angle by means of a normal distribution centered at the estimated value $\hat{\beta}$. The resulting form for $P(B_p|D_p,\beta_p)$ becomes more complex and does not lend itself to a closed form expression. It can, however, be pre-computed and stored in a reasonably sized 2-D look-up table.

3.2. Enforcing Spatial Coherence

In the previous section, we derived an expression for the marginal probability of the wrap number K_p at each pixel. We now discuss how this knowledge can be used in a framework that also leverages spatial coherence priors.

Spatial coherence is traditionally modeled by means of the Markov Random Field (MRF) formalism. MRF and related techniques attempt to find an image labeling that maximizes the joint posterior probability of label assignment given the observables. In practice, this translates to defining a cost function that is the sum of *data cost*, that penalizes label assignment inconsistent with the observation, and *discontinuity cost*, that penalizes changes of label assignment across nearby pixels. Unfortunately, closed form expressions for these cost functions are available only for simple 1-D cases, whereas cost minimization for generic 2-D images requires computationally expensive operations such as simulated annealing, belief propagation, or graph cut [19].

In this contribution, we define a different cost function, one that, while enforcing spatial coherence, can be minimized very efficiently. Our approach is inspired by the Non-Local Cost Aggregation (NLCA) algorithm, originally proposed by Yang for stereo matching [20]. In order to make this contribution self-contained, we begin by providing a short introduction to NLCA; we then show how NLCA can be applied to our problem.

3.2.1 The NLCA Algorithm

Let us first introduce our notation[1]. O_p represents the observation at pixel p. (In our case, O_p comprises Φ_p, B_p, and β_p.) $C_p(K)$ is the *marginal data cost* of assigning label K to pixel p based on the observation O_p. For example, in Yang's paper, the marginal data cost is defined by $C_p(K) = |I_p^l - I_{p-K}^r|$; it represents the "matching cost" between the

[1] Please note that our notation is slightly different from the one in [20]

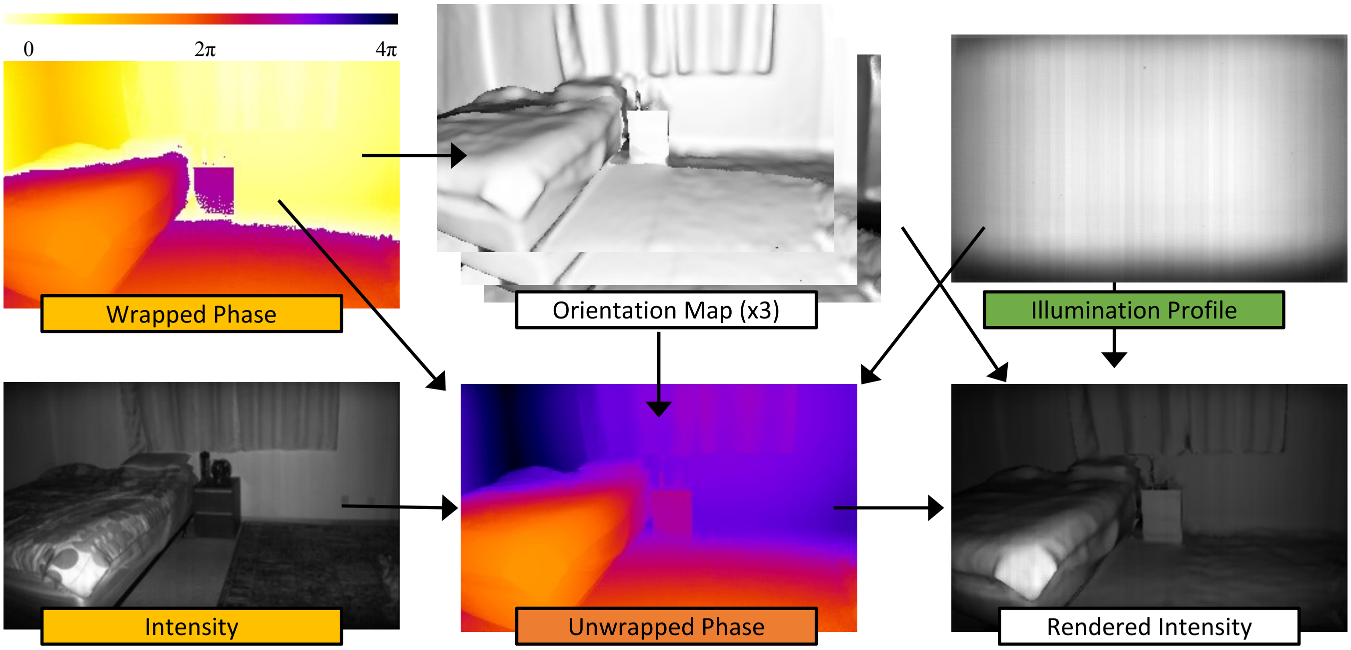


Figure 1: Visualization of the intensity model. Gold labels represent the observations: the wrapped phase and the intensity image. Green is measured a priori: the illumination profile is calibrated by making measurements of intensity against a surface with known albedo at known distances; embedded are electronic system artifacts (visible as stripes). White are rendered images: the surface normal is estimated from the wrapped phase, and this is done for each wrap value. Rendered here is the cos(β), the slant with respect to the camera. The rendered intensity is the image that would be predicted by our model. The simplifying assumption of uniform albedo is a chief cause in the differences between the predicted intensity image and the observed image: note the missing texture form the bed and carpet, and how the dark wooden dresser appears light. Orange is the final phase predicted by the algorithm.

left and the right image (I^l, I^r) if the disparity value K is assigned to pixel p in the left image.

The *aggregated cost* $C_p^A(K)$ is defined as follows:

$$C_p^A(K) = \sum\nolimits_q C_q(K) S_{p,q} \tag{5}$$

where the *similarity function* $S_{p,q}$ represents our belief, based on the observations, that pixels p and q should be assigned the same label. A small value of $C_p^A(k)$ (the aggregated cost at p for a certain label K) signifies that the set of *supporting pixels* (pixels that are *similar* to p) "agree" on this label. The NLCA algorithm simply assigns to each pixel the minimizer of its associated aggregated cost.

Similarity functions have been used extensively in computer vision. For example, the bilateral filter [21] uses a similarity function to define adaptive filter kernels, where two pixels are "similar" if they are at close distance *and* their colors are close in color space. Specifically, the bilateral filter defines the following:

$$S_{p,q} = \exp\left(-d\left(O_p, O_q\right)^2/\sigma_O^2\right) \tag{6}$$
$$\cdot \exp(\text{-}\|p-q\|^2/\sigma_D^2)$$

where $d\left(O_p, O_q\right)$ is a suitable distance between observables, and σ_O, σ_D are balancing constants[2]. The normalized cut algorithm [22] defines a similar metric for the edges of the graph to be clustered.

The NLCA algorithm defines the similarity function $S_{p,q}$ in a way that preserves the character of (6), while allowing for very fast computation. The algorithm first defines a planar (4- or 8-connected) graph on the image pixel grid, with edge cost between two neighboring pixels (r, s) equal to $d(O_r, O_s)$. Then, the minimum spanning tree of this graph is computed. The spanning tree, coupled with the edge costs, defines a *tree metric* on the image pixels, induced by the distance in the tree between the nodes representing any two pixels (where the tree distance is equal to the sum of the edge costs in the unique path between the two nodes). Let us define the tree distance between p and q as $d^T(p,q)$. One easily sees that two pixels have a small tree distance only if they have similar appearance *and* they are close in the tree (and thus close in the image grid). The NLCA algorithm defines the similarity function $S_{p,q}$ in (5) simply as:

$$S_{p,q} = \exp(-d^T(p,q)/\sigma) \tag{7}$$

A very useful characteristic of this similarity function is that, if pixel r is in the path in the tree between p and q, then

[2] Note that the similarity function $S_{p,q}(K)$ must be normalized for use as a kernel in the bilateral filter. Normalization is not necessary for NLCA.

$$S_{p,q} = S_{p,r} \cdot S_{r,q} \tag{8}$$

Yang uses this property to cleverly derive an extremely efficient algorithm for minimization of the aggregated cost $C_p^A(K)$ at each pixel. The computational cost of producing a labeling (in addition to the computation of the minimum spanning tree) is of 2 additions/subtractions and 3 multiplications for each label K in the set of labels. The maximum number of wraps in a given scene is a function of the maximum range of that scene and the modulation frequency of the illumination source. Though in practice, the ability to measure phase shift from the returning signal is dependent on a strong enough signal, so the illumination power should be chosen to sufficiently light the desired range. In our experiments, we use a maximum wrap value of 3, that is, 4 times the unambiguous range.

3.2.2 Phase Unwrapping Via NLCA

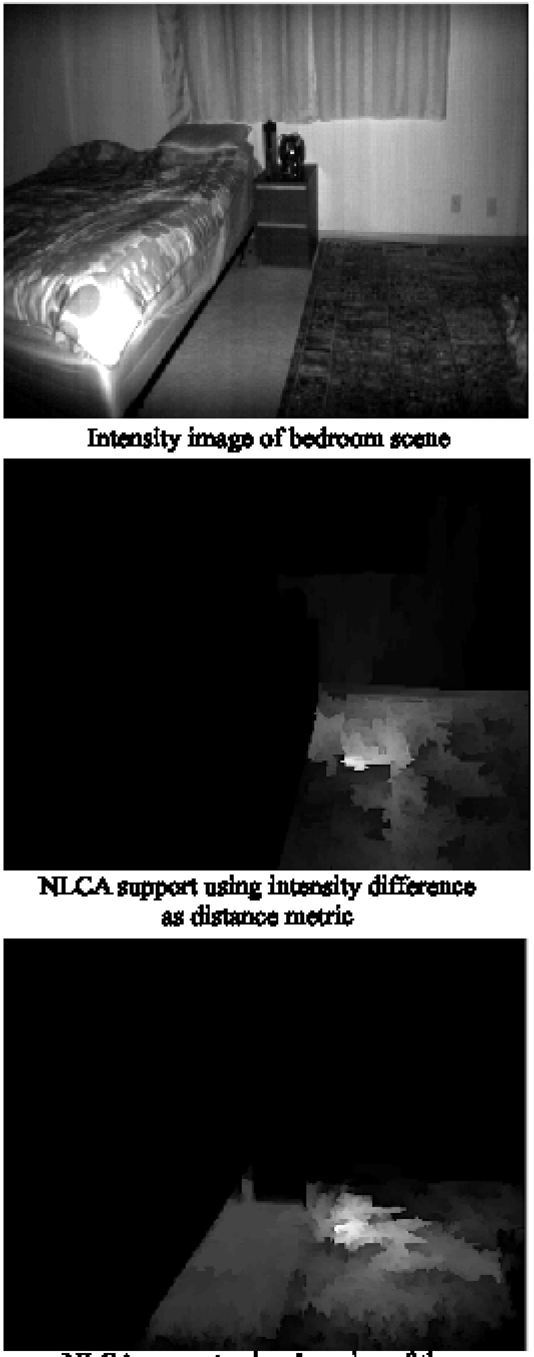
The NLCA algorithm is a generic labeling technique that can be easily extended to our wrap number estimation problem. Specifically, we define the marginal data cost as follows:

$$C_p(K) = -P\left(K_p | \Phi_p, B_p, \beta_p\right) \tag{9}$$

The cost is used to populate a cost volume, with each slice representing the wrap label. The volume is reweighted by the aggregated cost from (5), efficiently computed as described in [20].To encourage spatial coherence only between coherent areas, we choose a distance function that can reflect similarity of pixels beyond having merely similar phase measurements. The natural extensions involve using our other observable features: intensity B and surface normal N. We define distance functions for each feature type as: $d_\Phi\left(O_p, O_q\right) = \left|\Phi_p - \Phi_q\right|/2\pi$, $d_B\left(O_p, O_q\right) = \left|B_p - B_q\right|/\max(B)$, and $d_N\left(O_p, O_q\right) = 1 - \mathrm{dot}\left(N_p, N_q\right)$, where $\max(B)$ is the maximum intensity value over the image, and $\mathrm{dot}(\cdot,\cdot)$ is the dot product of normalized vectors. The functions can be assembled into multi-feature distance functions, as described in 4.4.

4. Experiments

A set of 45 indoor scenes (consisting of wrapped phase Φ, intensity B, and surface normal offset angle β) were used to test the proposed algorithm. Surface normal were estimated using the Point Cloud Library [23]. Data was collected from 3 different locations: a home, an office setting, and a computer lab. Only indoor scenes—the primary setting for ToF sensors—were chosen to be included as a known issue for active illumination sensors is excessive ambient light. We attempted to capture scenes with a variety depths, so to include a range of difficulties for which to test the algorithm. Data was captured using a prototype ToF camera with a resolution of 320x200 pixels, on loan from a commercial vendor.



Intensity image of bedroom scene

NLCA support using intensity difference as distance metric

NLCA support using 1-cosine of the angle between surface normals

Figure 2: Visualization of the support for one pixel. The top image shows the intensity image of a bedroom scene. The middle image shows the support provided to a single pixel using the absolute difference of intensity as the distance metric. The bottom shows the support using the 1-cos(β), the angle between the estimated surface normal.

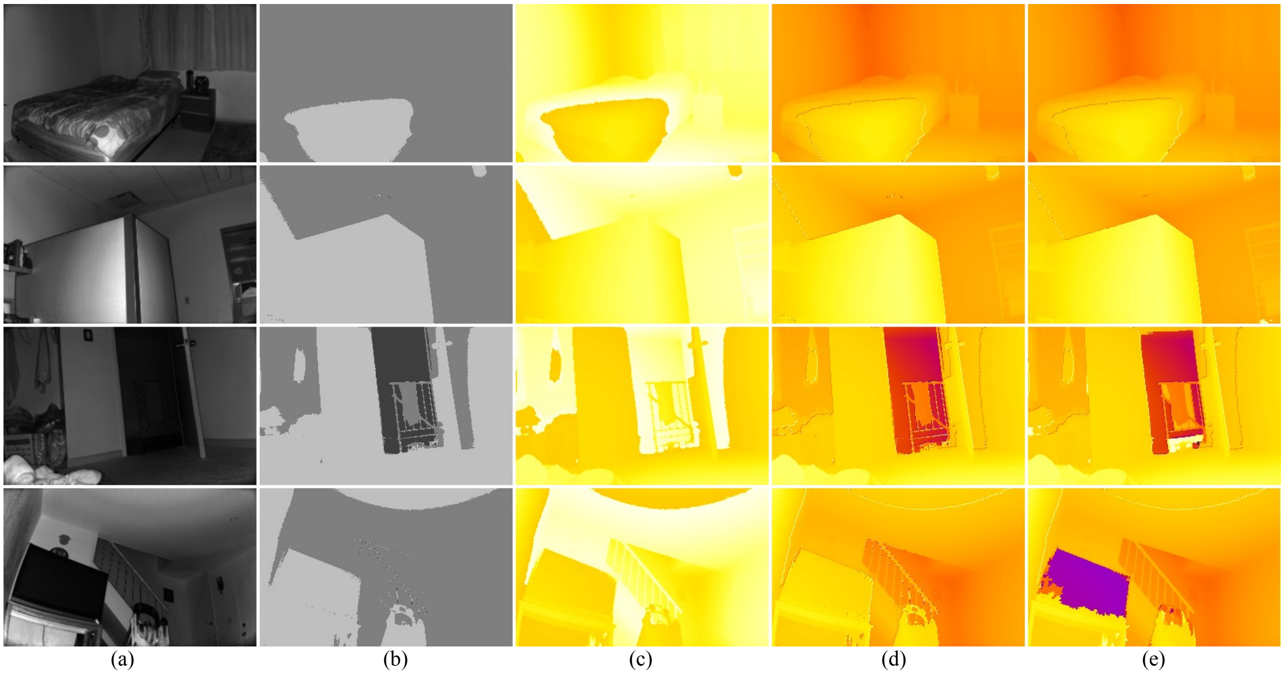


Figure 4: Selected scenes demonstrating our algorithms' performance. Column (a) shows the intensity image from the active illumination. Column (b) is the ground truth number of phase wraps, from 0 to 2 wraps with darker greys being more wraps. Columns (d) and (e) show phase, increasing from hot white to cool purple, while (c) shows the measured phase (wrapped at 2π). Column (d) is representation of the unwrapped phase or depth, as measured by a multifrequency ToF camera. The final column shows the reconstructed phase from our method. The top two rows taken from the *easy* group and the bottom two from the *medium*, chosen to highlight some specific difficulties. In the 3rd row, column (b) contains a thin railing challenging the spatial coherence assumption. While the very low albedo of the TV screen in the 4th row may lead the brightness model to assume it is distant.

Each static scene was captured at two frequencies, 51.4MHz and 68.6MHz, which determined unambiguous ranges of 2.2m and 2.9m respectively. Ground truth was determined by combining these pairs of captures using the approach of [2], which uses both frequencies to obtain an unambiguous phase measurement. We ran our phase unwrapping algorithm on both frequencies individually.

Our tests were run under the assumption that the maximum range of the scene is known ahead of time, thus limiting the number of phase wraps the algorithm can expect to encounter. The difficulty of the problem is compounded by a larger number of wraps, and we therefor classified our test scenes by difficulty in the maximum value of wrap label K, either 1 (14 cases), 2 (45 cases), or 3 (31 cases).

4.1. Comparison to Previous Methods

We tested the full proposed solution, intensity model complete with surface normal estimation, and spatial coherence enforced by nonlocal cost aggregation, using a distance metric utilizing the phase Φ and surface normal N (see 4.4 for details). We compared the results of our algorithm to that of Choi [17] and Crabb [2]. Over the entire data set, we observed an average of 94.1% pixels labeled correctly from our method, compared to 84.3% for Choi and 89.7% for Crabb. However, we noted that the performance was very much dependent on the difficulty. In the easiest cases of a single phase wrap, Crabb's MRF method slightly outperforms our proposed method 99.4% to 99.1%, in the harder cases we see that the proposed method more than makes up for it, achieving rates of up to 92.3% of pixels labeled correctly, more than 9 percentage points above the previous method.

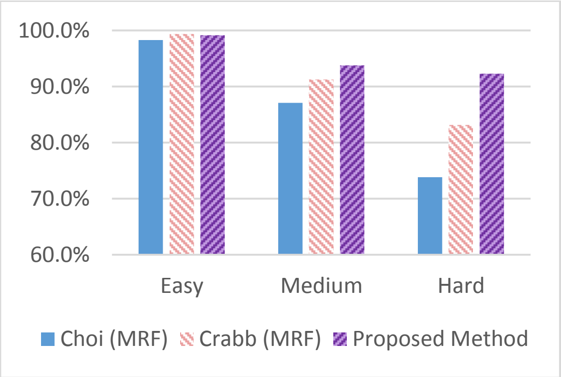
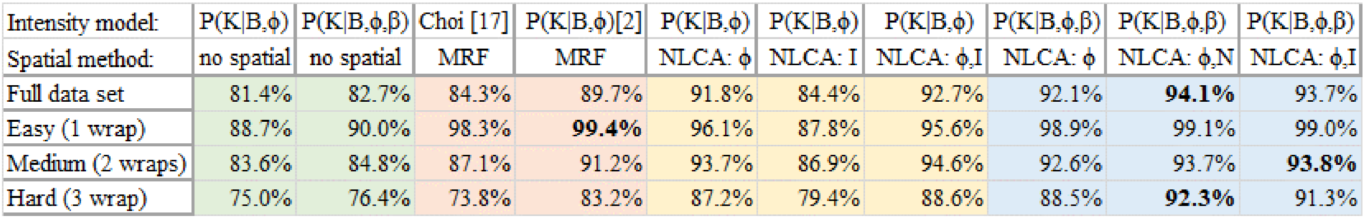


Figure 3: Comparison of the proposed method against prior methods of Choi and Crabb. Broken down into easy, medium, and hard cases, we can see that in the easy cases, all method perform nearly perfectly, but the advantages of the proposed method stand out in harder cases.



Intensity model:	$P(K\mid B,\phi)$	$P(K\mid B,\phi,\beta)$	Choi [17]	$P(K\mid B,\phi)$[2]	$P(K\mid B,\phi)$	$P(K\mid B,\phi)$	$P(K\mid B,\phi)$	$P(K\mid B,\phi,\beta)$	$P(K\mid B,\phi,\beta)$	$P(K\mid B,\phi,\beta)$
Spatial method:	no spatial	no spatial	MRF	MRF	NLCA: ϕ	NLCA: I	NLCA: ϕ,I	NLCA: ϕ	NLCA: ϕ,N	NLCA: ϕ,I
Full data set	81.4%	82.7%	84.3%	89.7%	91.8%	84.4%	92.7%	92.1%	**94.1%**	93.7%
Easy (1 wrap)	88.7%	90.0%	98.3%	**99.4%**	96.1%	87.8%	95.6%	98.9%	99.1%	99.0%
Medium (2 wraps)	83.6%	84.8%	87.1%	91.2%	93.7%	86.9%	94.6%	92.6%	93.7%	**93.8%**
Hard (3 wrap)	75.0%	76.4%	73.8%	83.2%	87.2%	79.4%	88.6%	88.5%	**92.3%**	91.3%

Table 1. This table presents a summary of performance from a handful of the experiments we ran, in which we analyzed the many components and variations of our proposed method. The first two columns (green) compare the best estimate using the intensity model by itself, without enforcement of spatial coherence. Columns 3-4 intensity (red) show the performance using the Markov random field approach of [2], the first uses an intensity model by Choi [17] and the next uses an intensity model without normal estimation and discontinuity cost considering only Φ. Columns 5-7 (orange) enforce spatial coherence by NLCA, shown with selected variations in distance metric. Using Φ only is the most direct comparison to the approach of column 1. The final columns 8-10 (blue) show the full proposed method: using an intensity model with normal estimates with spatial coherence enforced by NLCA, with a selection of distance metrics. Distance metric with both the phase measurement Φ with surface normal estimate N produces overall superior results.

4.2. Intensity Model Comparison

The first set of tests looked exclusively at the intensity models' ability to estimate the correct wrapping label. Simply, the likelihood of each wrapping label was computed as in eq. (4) or eq. (7) from [2], and the mostly likely option is selected. Surprisingly, the inclusion of an estimate of the surface normal had only a small impact on the ability to choose a wrap label from intensity alone.

The normal is estimated using the Point Cloud Library [23]. We found choice of window size effects the labeling accuracy by up to nearly 3%, when using intensity model alone. Choosing a neighborhood of 49 points showed good results, and our reported results use that parameter setting.

Compiling all scenes, we found a labeling accuracy of 81.4% for the intensity model without normal, and 82.7% with. Breaking the data up by difficulty we found a similar spread of just over 1%, as shown in table 1.

4.3. Comparison of Spatial Coherence Methods

We can compare the spatial coherence methods directly by using the same values from the 'data term' of the MRF to populate the cost volume described in 3.2.2. To make this comparison as similar as possible, we define the distance function as $d_\Phi(O_p, O_q) = |\Phi_p - \Phi_q|$.

Looking at the tests as a whole, we find a small but significant advantage to the NLCA approach, with 91.8% of pixels labeled correctly, over 89.7%. However, when we separate the cases by difficulty we find the advantage is not so clear cut. In the easiest cases, involving only 1 phase wrap, the MRF approach performs excellently, mislableling only 0.6% of the pixels, while NLCA misses almost 4%. However, as the difficulty increases, we find NLCA demonstrates an advantage, seen in table 1.

4.4. Exploring the Distance Function of Minimum Spanning Trees

A distinct advantage of the NLCA approach over MRF is that here we have more freedom in defining our similarity function beyond solely the measured phase. Examples of the difference in pixel support are visualized in Figure 2. Notice that when using the intensity, the support stays mostly limited to pixels on the carpet, as they are of a similar intensity. However, using the normal estimate N, support is spread over the floor, as it all shares the same orientation.

We experimented with these different distance functions by themselves, and combined with each other in a number of ways, such as the maximum, l1- and l2-norms: $\max(\alpha_B d_B, \alpha_\Phi d_\Phi, \alpha_N d_N)$, $\|\alpha_B d_B, \alpha_\Phi d_\Phi, \alpha_N d_N\|_1$, and $\|\alpha_B d_B, \alpha_\Phi d_\Phi, \alpha_N d_N\|_2$, with weighting coefficients α where $\sum_i \alpha_i = 1$ (and we have dropped the observations O_p simply to shorten the notation).

In our tests with the model excluding the surface normal estimate, we found that using the intensity alone as s distance metric produced quite poor results. This makes sense as it is important the support for labels is not shared across phase wrap borders, in which the difference in Φ will be quite high. However, using B jointly with Φ produces superior results, as seen in columns 6, 7 of table 1. Including the normal estimation, we show a handful of distance metrics composed of combinations of observed phase Φ, intensity B, and estimated normal N. We found the best performance with the distance function defined as $d_{\Phi,N}(O_p, O_q) = \alpha_\Phi |\Phi_p - \Phi_q|/2\pi + \alpha_N |1 - \mathrm{dot}(N_p, N_q)|$ with $\alpha_\Phi = .7, \alpha_N = .3$.

4.5. Algorithm efficiency

It is difficult to make a direct comparison of running times to the previous method using MRF, as that approach was implemented only in Matlab using a less than optimal message passing schedule (simultaneous message passing rather than sequential (e.g. left, right, up, down), while we have implemented our proposed in C++ with an aim to optimize for speed. However, we can report that the observed running time (about 0.35-.75 seconds per frame, depending on maximum wrap on an Intel i7 2.7Ghz core) is more than 2 orders of magnitude faster than the time reported in [2] (about 175 seconds per frame). We can break down this time into the various steps of the algorithm:

surface normal estimation (.2s per wrap label), cost volume construction (.005s using look-up table as a replacement for numerical integration on the fly), MST construction .02 s), NLCA support computation (.004 s).

5. References

[1] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, *Time-of-flight cameras and microsoft Kinect™*.: Springer Science & Business Media, 2012.

[2] Ryan Crabb and Roberto Manduchi, "Probabilistic Phase Unwrapping for Single-Frequency Time-of-Flight Range Cameras," in *Proc. International Conference on 3D Vision (3DV 14)*, Tokyo, 2014.

[3] Dennis Ghiglia and Mark Pritt, *Two-dimensional phase unwrapping: theory, algorithms, and software*. New York: Wiley, 1998.

[4] S Burak Göktürk, Hakan Yalcin, and Cyrus Bamji, "A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions," in *CVPRW '04*, Washington, DC, USA, 2004, p. 35.

[5] W. Xu et al., "Phase-unwrapping of SAR interferogram with multi-frequency or multi-baseline," in *Geoscience and Remote Sensing Symposium*, Pasadena, 1994.

[6] D. Droeschel, D. Holz, and S. Behnke, "Multi-frequency phase unwrapping for time-of-flight cameras," in *Intelligent Robots and Systems (IROS)*, Taipei, 2010.

[7] D. Falie and V. Buzuloiu, "Wide range time of flight camera for outdoor surveillance," in *Microwaves, Radar and Remote Sensing Symposium (MRRS)*, Kiev, 2008.

[8] O. Choi, S. Lee, and H Lim, "Interframe consistent multifrequency phase unwrapping for time-of-flight cameras," *Optical Engineering*, vol. 52, no. 5, 2013.

[9] C. Beder, B. Barzak, and R. Koch, "A combined approach for estimating patchlets afrom pmd depth images and stereo intensity images," in *German Association for Pattern Recognition (DAGM)*, Heidelberg, 2007.

[10] Sigurjon Arni Gudmundsson, Henrik Aanaes, and Rasmus Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3d estimation," *International Journal of Intelligent Systems Technologies and Applications*, vol. 5, no. 3, pp. 425-433, 2008.

[11] Ouk Choi and Seungkyu Lee, "Wide range stereo time-of-flight camera," in *International Conference on Image Processing*, Orlando, 2012.

[12] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *Computer Vision and Pattern Recognition*, Anchorage, 2008.

[13] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, "A probabilistic approach to ToF and stereo data fusion," in *3DPVT*, Paris, 2010.

[14] Ouk Choi and S. Lee, "Fusion of time-of-flight and stereo for disambiguation of depth measurements," in *Asian Conference on Computer Vision*, Daejeon, Korea, 2012, pp. 640-653.

[15] Martin Böhme, Martin Haker, Thomas Martinetz, and Erhardt Barth, "Shading constraint improves accuracy of time-of-flight measurements," *Computer vision and image understanding*, vol. 114, no. 12, pp. 329-1335, 2010.

[16] Shane H. McClure, Cree M.J., A Dorrington, and A Payne, "Resolving depth-measurement ambiguity with commercially available range imaging cameras," in *IS&T/SPIE Electronic Imaging*, San Jose, CA, USA, 2010.

[17] Ouk Choi et al., "Range unfolding for time-of-flight depth cameras," in *International Conference on Image Processing (ICIP)*, Hong Kong, 2010.

[18] Serban Oprisescu, Dragos Falie, Mihai Ciuc, and Vasile Buzuloiu, "Measurements with ToE cameras and their necessary corrections," in *International Symposium on Signals, Circuits and Systems*, Iasi, Romania, 2007.

[19] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222-1239, 2001.

[20] Qingxiong Yang, "A non-local cost aggregation method for stereo matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[21] Carlo Tomasi and Roberto Manduchi, "Bilateral Filtering for Gray and Color Images," in *Internation Conference on Computer Vision*, 1998, pp. 839-846.

[22] Jianbo Shi and Jitendra Malik, "Normalized Cut and Image Segmentation," *TPAMI*, vol. 22, no. 8, pp. 888-905, 2000.

[23] Radu Bogdan Rusu and Steve Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, 2011.