# Augmenting Depth Camera Output Using Photometric Stereo

Robert Anderson
Department of Engineering
University of Cambridge

Björn Stenger
Computer Vision Group
Toshiba Research Europe

Roberto Cipolla
Department of Engineering
University of Cambridge

## Abstract

*We present a system for augmenting depth camera output using multispectral photometric stereo. The technique is demonstrated using a Kinect sensor and is able to produce geometry independently for each frame. Improved reconstruction is demonstrated using the Kinect's inbuilt RGB camera and further improvements are achieved by introducing an additional high resolution camera. As well as qualitative improvements in reconstruction a quantitative reduction in temporal noise is shown. As part of the system an approach is presented for relaxing the assumption of multispectral photometric stereo that scenes are of constant chromaticity to the assumption that scenes contain multiple piecewise constant chromaticities.*

## 1 Introduction

Depth cameras are becoming increasingly accessible and their resolution and accuracy is continually improving. Examples are time-of-flight cameras and triangulation-based devices, such as the Kinect [1]. Both produce low resolution range images exhibiting high frequency noise. In this paper we propose the use of photometric stereo to augment the output from depth cameras with the two aims of (1) reducing high frequency spatial and temporal noise in the output, and (2) recovering fine detail that the depth camera alone cannot resolve. The choice of photometric stereo is motivated by its error characteristics, which are to provide accurate high frequency data while tending to introduce error in the global shape. Combining the accurate high frequency information from photometric stereo with the accurate low frequency information from the depth camera yields a high quality final result. In order to produce geometry from every frame multispectral photometric stereo is used. This allows three lighting directions to be captured in a single image by using three different colored lights to illuminate the scene from separate directions. This makes the proposed technique particularly suited to dynamic scenes.

## 2 Related work

Prior work has examined using multiview stereo to enhance the output from depth cameras, [5, 8, 17], but while improvements were demonstrated this approach is ultimately limited by the fact that both depth cameras and stereo methods exhibit high frequency noise. In [7] the heuristic that depth discontinuities tend to co-occur with color or brightness changes within an image is used, allowing information from a high resolution color image to be applied to a lower resolution depth image. This enables edges to be recovered more accurately and noise on smooth surfaces to be reduced. Registered high resolution images are also taken advantage of in [16] where an iterative bilateral filtering approach is used.

Photometric stereo produces accurate normal maps which are ideal for enhancing noisy range images. Nehab *et al.* [12] presented a method for producing a surface containing the high frequency information of a normal map and the low frequency information from a depth map. Photometric stereo has been used to produce reconstructions containing very fine detail [11, 13].

Photometric stereo uses at least three different lighting conditions of the same scene to estimate a dense normal map [14]. The necessary three lighting conditions can either be achieved by multiplexing in the time domain or the frequency domain. Multiplexing in the frequency domain, by using three different colored light sources, means that three lighting directions can be observed in a single image, allowing reconstruction at frame rate. This idea was first proposed over 15 years ago [2, 15] and recently has been demonstrated to give good performance on dynamic scenes [6, 9]. The disadvantage of multispectral photometric stereo over time multiplexed photometric stereo is that it assumes scenes are of constant chromaticity, although this assumption can be relaxed to piecewise constant chromaticity as shown later.

## 3 Theory

The proposed system carries out reconstruction at each frame in three steps. Firstly a dense normal map is constructed using multispectral photometric stereo. Secondly this normal map is aligned with the depth camera's output and finally the two sources of data are combined to give a 3D model.

Section 3.1 outlines the photometric calibration technique used and gives a brief description of the extension to scenes which contain multiple chromaticities. The key to this extension is using the low frequency geometry from the depth sensor along with a color image to segment the scene into regions of constant chromaticity. The subsequent sections detail the reconstruction process.

### 3.1 Multispectral photometric calibration

We assume a Lambertian reflectance model and no ambient lighting. Given three distant point light sources illuminating a surface with unit normal $\mathbf{n}$ and albedo $\alpha$, it has been shown, [2], that the observed intensity of the surface is given by

$$\mathbf{c} = \alpha \mathbf{V} \mathbf{L} \mathbf{n} = \alpha \, [\, \mathbf{v_0} \;\; \mathbf{v_1} \;\; \mathbf{v_2} \,] \, [\, \mathbf{l_0} \;\; \mathbf{l_1} \;\; \mathbf{l_2} \,]^\top \mathbf{n}, \qquad (1)$$

where $\mathbf{c}$, $\mathbf{l}_i$ and $\mathbf{v}_i$ are all 3-vectors. $\mathbf{c}$ denotes the RGB image intensity, $\mathbf{l}_i$ defines the direction of light $i$, and $\mathbf{v}_i$ is the combined response of surface and sensor to light $i$. The matrix $\mathbf{V}$ models the combination of the surface's chromaticity, the lights' spectral distributions and the camera sensors' spectral sensitivities and so varies for regions of different chromaticity. Given a new input image, a normal map can be reconstructed using

$$\mathbf{n} \propto (\mathbf{VL})^{-1}\,\mathbf{c}. \tag{2}$$

To calibrate the photometric setup, the matrix $\mathbf{M} = \mathbf{VL}$ must be found. We use the technique of [6] in which ($\mathbf{c}$,$\mathbf{n}$) pairs are used to robustly estimate $\mathbf{M}$ through a RANSAC-based algorithm. Three ($\mathbf{c}$,$\mathbf{n}$) pairs are sufficient to estimate $\mathbf{M}$ up to a scale factor by solving the linear system $\mathbf{c} = \alpha\mathbf{Mn}$, as long as $\alpha$ and $\mathbf{M}$ are the same for each pair. This means that each pair must come from a point with both the same albedo and the same chromaticity, however these surface properties are not known *a priori*. Three points are chosen at random and used to hypothesize $\mathbf{M}_e$, then support is measured from all other pairs by testing for each pair whether

$$|\,\mathbf{c} - \mathbf{M}_e\mathbf{n}\,| < \tau, \tag{3}$$

where $\tau$ is a threshold value. If the three points have the same albedo and chromaticity then $\mathbf{M}_e$ will be a good estimate and correctly predict the image color from the normal map. More sets of three pairs are chosen at random, and the $\mathbf{M}_e$ with the greatest support kept, until the RANSAC stopping criteria are met. Subsequently an estimate of $\mathbf{M}_e$ is made using a least squares approach from all pairs which supported the last accepted $\mathbf{M}_e$.

In [6] a structure-from-motion approach is used to estimate coarse geometry to provide estimates of $\mathbf{n}$ throughout a calibration sequence. Taking advantage of the depth camera this coarse geometry is already available. The calibration technique can be applied to a single frame or to a sequence, but it must be guaranteed that a wide range of normal directions are present to make the estimation of $\mathbf{M}$ well posed.

To extend this calibration procedure to scenes containing multiple chromaticities we estimate more than one $\mathbf{M}$ matrix. After the first $\mathbf{M}$ has been estimated, corresponding to the dominant chromaticity in the scene, we remove from consideration all points for which

$$\left|\mathbf{c} - \frac{|\mathbf{c}|}{|\mathbf{Mn}|}\mathbf{Mn}\right| < \tau, \tag{4}$$

indicating that this point is well modeled by $\mathbf{M}$. This condition differs from (3) in that the albedo of the surface no longer plays a role, only the chromaticity. All points with the same chromaticity are removed independent of their albedo. To find a calibration matrix for the second most dominant chromaticity in the scene the same calibration procedure is carried out for the remaining points.

### 3.2 Multispectral photometric reconstruction

During reconstruction of multichromatic scenes the input image must be segmented to decide which $\mathbf{M}$ matrix to use to estimate $\mathbf{n}$ at each pixel. The segmentation is set up in a Markov Random Field (MRF) framework with unary and pairwise terms described below. There is one node for each non-shadowed pixel in the input image and edges are added between nodes corresponding to neighboring pixels within a 4-neighborhood.

The depth image from the depth camera is smoothed and a normal $\mathbf{n}_s$ calculated from it at each pixel. The input image is smoothed by the same amount to give a smoothed image value $\mathbf{c}_s$ at each pixel and the unary term for using matrix $\mathbf{M}_i$ is given by

$$\left|\mathbf{c}_s - \frac{|\mathbf{c}_s|}{|\mathbf{M}_i\mathbf{n}_s|}\mathbf{M}_i\mathbf{n}_s\right|^2. \tag{5}$$

This encourages calibration matrices to be chosen which produce a normal map with similar low frequency characteristics to the recovered depth image. We use the Potts model for the pairwise terms [4] in which no cost is assigned to neighboring pixels sharing a label and a cost $\gamma$ is assigned if they have different labels. The cost $\gamma$ is modulated by an edge map of the input image. For pixels on an edge $\gamma$ is small while for pixels not on an edge $\gamma$ is large. The maximum a posteriori (MAP) solution to the MRF is estimated using the tree reweighted message passing algorithm [10] and reconstruction is based upon the labels assigned.

### 3.3 Registration of photometric and depth data

Depending upon the type of depth camera used the depth image and color image may be inherently aligned, however if this is not the case then registration between the two must be carried out. Stereo calibration is a well studied topic and we wish to make use of the available tools to calibrate between color images and depth images. One popular approach for standard stereo calibration is to image a chessboard pattern at several orientations allowing for a homography to be estimated followed by a full optimization of both intrinsic and extrinsic camera parameters. To extend this to include depth images a chessboard pattern that shows up in both color and depth images is required. This can be achieved, for example, by cutting out the black squares on every other row in a chessboard pattern and ensuring that there is a dark background behind the board during capture. The depth images produced by this method do not always have clean corners at the edges of the cut out squares which results in corner localization approaches failing. More robust results are obtained by fitting lines to the pattern and then using the intersection of these lines to estimate corner locations.

### 3.4 Combining photometric and depth data

Given a depth image a 3D mesh is created. Using the calibration information obtained in the previous section this mesh is transformed into the color camera's coordinate system and ray tracing is used to produce a depth image from the point of view of the color camera.

Once values from the depth image have been projected onto the color image, the method of [12] is used to combine the two types of information. First the low frequency bias in the normal field is removed using the depth map. Geometry is then estimated by optimizing an energy function which forces the surface to fit both the observed depths and the observed normals (see [12] for details).

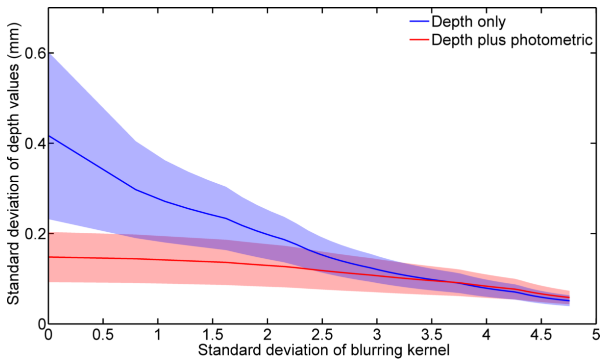


Figure 1. Temporal noise as a function of spatial smoothing of the depth images. Shaded areas equal to half standard deviation of results. The use of photometric data allows for a reduction in temporal noise without the need for smoothing.

## 4 Experimental setup

The Kinect sensor used in these experiments has a resolution of 640 × 480, both in the depth as well as the color domain. In addition three light sources of different colors were arranged in a triangular configuration around the sensor. A Grasshopper camera with a resolution of 1600 × 1200 was used as a secondary high resolution camera (Section 5.4). Depth and color images were captured from the Kinect at an average rate of 30 fps. Average processing time was 6 seconds per frame using single threaded C++. The majority of this processing time is taken by the integration from normal map to 3D surface. There exist Fourier Transform methods for this computation [3], which can run in real time, however they require a regular grid of samples and so may introduce deformations when reconstructing scenes such as faces.

## 5 Experiments

### 5.1 Reduction of temporal flicker

To measure temporal flicker thirty reconstructions of a static scene were carried out and the standard deviation of the estimated depth value at each pixel that was successfully reconstructed in all thirty frames was calculated. This was carried out on three scenes and the average results computed. Figure 1 shows the temporal noise values as a function of spatial smoothing introduced by blurring each depth image independently with a Gaussian kernel. When there is little spatial smoothing the additional photometric information significantly reduces the magnitude of the temporal noise. While sufficient smoothing does reduce flicker, it also removes fine details, while the addition of photometric data reduces noise as well as enhancing detail. Note that no temporal smoothing is used.

### 5.2 Reconstruction of a plane

The previous experiment only investigates variations in reconstruction over time and does not give any information about the absolute accuracy of the system. To investigate this a known planar surface was reconstructed and a plane was fitted to the resulting point cloud using least squares estimation. The average deviation of each reconstructed point from this plane was measured and used as a metric for reconstruction accuracy. Using only the depth information, the average absolute error to the best fit plane was 1.6 mm. By including photometric information this error was reduced to 1.2 mm. From figure 2 it can be seen that the photometric information reduces high frequency noise, but cannot correct the overall low frequency deformation present in the depth image.

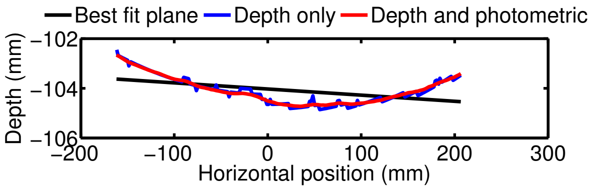


Figure 2. Slice though a reconstructed plane. Photometric stereo can smooth the high frequency noise but does not remove the low frequency deformation. Note different axis scales.

### 5.3 Reconstruction of dynamic scenes

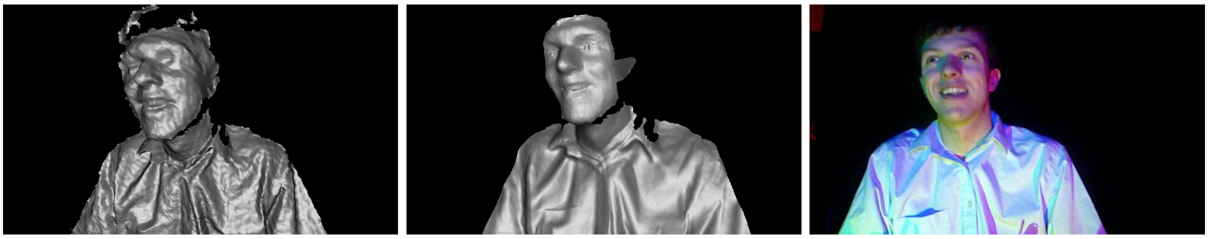
To demonstrate the qualitative improvement that incorporating photometric stereo brings to reconstructions of dynamic scenes several face sequences were captured. Figure 3 shows an example reconstruction and it can be seen that the addition of photometric stereo both reduces noise and resolves finer details than are visible in the initial reconstruction. Two calibration matrices were found, one that modeled the shirt and one that modeled the skin.

### 5.4 Addition of a second camera

Figure 4 shows the quality of reconstruction achieved when a high resolution camera is used to augment the depth camera. The depth camera information is still important as the photometric reconstruction on its own results in low frequency deformations, as seen in figure 4(d). Currently the high resolution camera is not synchronized with the Kinect camera, so reconstruction can only be performed on individual images, which are temporally aligned.

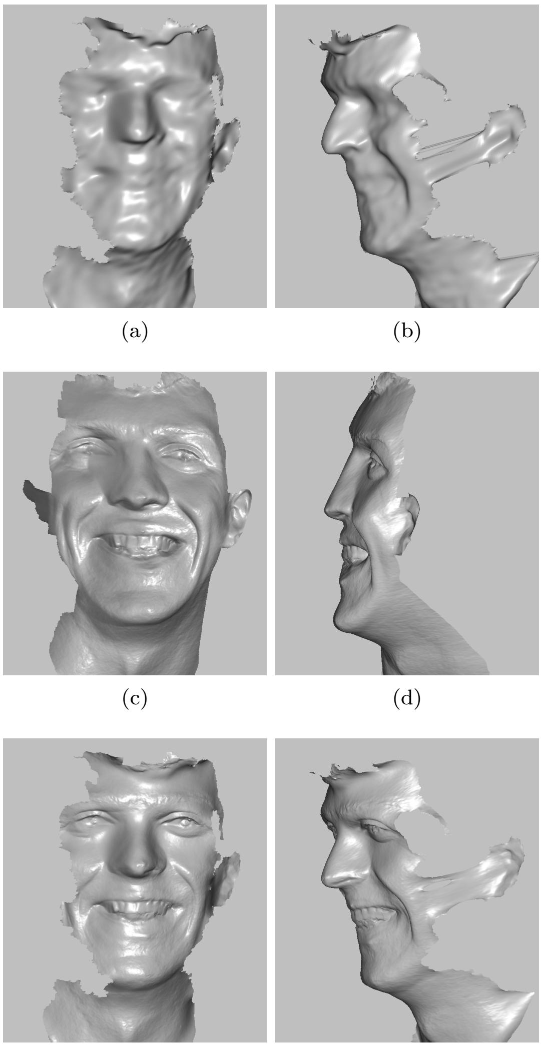
## 6 Conclusions

We have demonstrated the effectiveness of photometric stereo for both reducing the noise present in depth camera outputs and allowing for the resolution of fine detail. The modest additional equipment requirements for this approach are three different colored lights. The further addition of a high resolution camera allows for more detail to be recovered. We have also demonstrated that the assumption of constant chromaticity imposed by multispectral photometric stereo can be relaxed by taking advantage of a depth camera.



(a) (b) (c)

Figure 3. Reconstruction of a dynamic sequence using (a) the depth camera only and (b) the depth camera combined with photometric stereo applied to the image from the inbuilt Kinect color camera (c).



(a) (b) (c) (d) (e) (f)

Figure 4. (a,b) Reconstructions using depth data only, from (a) the original viewpoint and (b) a novel viewpoint. (c,d) Reconstructions from the same viewpoints using photometric data only, note the low frequency deformation. (e,f) Final result from the same viewpoints, showing high frequency detail and correct overall shape.

## References

[1] Kinect depth sensor, www.xbox.com/en-us/kinect.

[2] M. Drew and L. Kontsevich. Closed-form attitude determination under spectrally varying illumination. In *Proc. CVPR*, pages 985–990, June 1994.

[3] R. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *PAMI*, 10(4):439–451, July 1988.

[4] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51:271–279, 1989.

[5] U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. In *Int. J. of Intelligent Systems Technology and Applications*, volume 5, pages 325–333, 2008.

[6] C. Hernandez and G. Vogiatzis. Self-calibrating a real-time monocular 3D facial capture system. In *Proc. 3DPVT*, 2010.

[7] B. Huhle, S. Fleck, and A. Schilling. Integrating 3D time-of-flight camera data and high resolution images for 3DTV applications. In *3DTV Conference, 2007*, pages 1–4, May 2007.

[8] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3D reconstruction. In *ICCV Workshops*, pages 1542–1549, 2009.

[9] M. Klaudiny, A. Hilton, and J. Edge. High-detail 3D capture of facial performance. In *Proc. 3DPVT*, 2010.

[10] V. Kolmogorov. Convergent tree-rewieghted message passing for energy minimization. In *PAMI*, volume 28, pages 1568–1583, 2006.

[11] Z. Lu, Y. Tai, M. Ben-Ezra, and M. Brown. A Framework for Ultra High Resolution 3D Imaging. In *Proc. CVPR*, 2010.

[12] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *In Proc. of the ACM SIGGRAPH*, pages 536–543, August 2005.

[13] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. In *SIGGRAPH Asia*, pages 1–11, 2009.

[14] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

[15] R. J. Woodham. Gradient and Curvature from Photometric Stereo Including Local Confidence Estimation. *Journal of the Optical Society of America*, (11):3050–3068, 1994.

[16] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *Proc. CVPR*, pages 1–8, 2007.

[17] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. CVPR*, pages 1–8, 2008.