A Case Study Evaluation: Perceptually Accurate Textured Surface Models

Greg Ward, *Dolby Canada*, greg.ward@acm.org
Mashhuda Glencross, *University of Manchester*, mashhuda@manchester.ac.uk

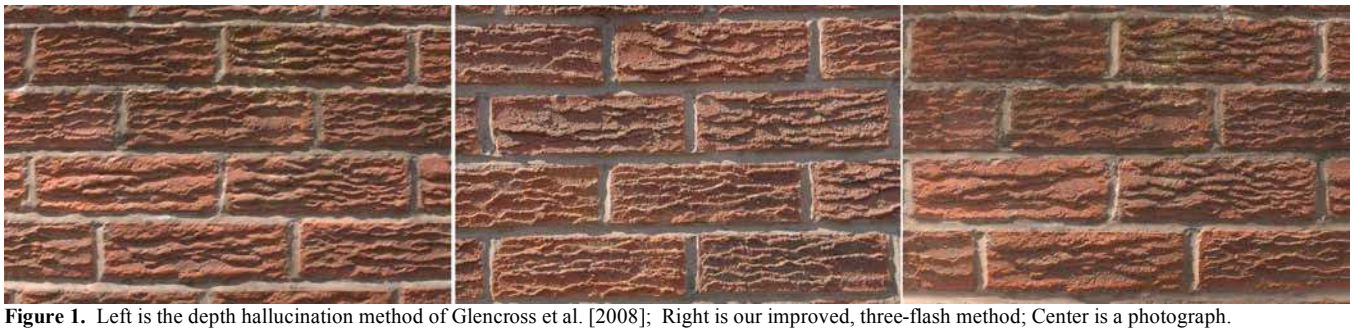

Figure 1. Left is the depth hallucination method of Glencross et al. [2008]; Right is our improved, three-flash method; Center is a photograph.

ABSTRACT
This paper evaluates a new method for capturing surfaces with variations in albedo, height, and local orientation using a standard digital camera with three flash units. Similar to other approaches, captured areas are assumed to be globally flat and largely diffuse. Fortunately, this encompasses a wide array of interesting surfaces, including most materials found in the built environment, e.g., masonry, fabrics, floor coverings, and textured paints. We present a case study of naïve subjects who found that surfaces captured with our method, when rendered under novel lighting and view conditions, were statistically indistinguishable from photographs. This is a significant improvement over previous methods, to which our results are also compared.

Index Terms—Lighting, shading and textures, Perceptual validation, Computer vision, Texture.

1 INTRODUCTION

Photographic textures have been applied to geometric models to enhance realism for decades, and are an integral part of every modern rendering engine. However, two-dimensional textures have a tendency to resemble wallpaper at oblique angles, and are unable to produce realistic silhouettes or change appearance under different lighting. Displacement mapping or relief mapping methods [Oliveira 2000] can overcome these limitations, but full reflectance and geometry model data are difficult to capture from real surfaces, requiring expensive scanning equipment and subsequent manual alignment with photographically acquired textures [Rushmeier et al. 2003; Lensch et al. 2003], a large set of data, and/or complicated rigs [Dana et al. 1999; Marschner et al. 1999]. Games companies often employ skilled artists to create texture model data for displacement mapping using 3D modeling packages, which is a laborious process. Glencross et al. introduced a simple and inexpensive shape-from-shading technique for "hallucinating" depth information from a pair of photographs taken from the same viewpoint, one captured with diffuse lighting and another taken with a flash [Glencross et al. 2008].

Since the method captures albedo simultaneously, no alignment steps are needed. Although the authors do not claim absolute accuracy in terms of reproducing depth values, user studies showed that subjects found it difficult to distinguish the plausibility of hallucinated depth relative to ground truth data, adequately demonstrating the technique's value for realistic computer graphics.

In this paper, we ask the question "what level of additional captured model accuracy will result in synthetic images that are indistinguishable from photographs?" To answer this question, we extend the depth hallucination method to include photometrically measured surface orientation [8]. By adding two additional flash units to the one employed by [Glencross et al. 2008], we are able to derive accurate surface orientations at most pixels in our captures. Our validation studies demonstrate that the addition of measured surface orientation results in no statistically significant differences in perception between photographs and captured, re-rendered images.

The entire process has been automated, with capture taking a few seconds and model extraction less than a minute.

Since the focus of this work is the evaluation of captured model fidelity and its impact on the visual accuracy of the results, we begin by first briefly discussing related work, and then give an overview of the photometric method used. We evaluate the visual impact of measured surface orientation on computer-generated imagery through an experimental study. Finally, we conclude by discussing the limitations and suggesting future directions.

2 RELATED WORK

Besides the aforementioned work of Glencross et al. [2008], our method is closely related to that of Rushmeier and Bernardini, who used a comparable multi-source arrangement to recover surface normal information [Rushmeier and Bernardini 1999]. This is similarly built on the photometric stereo technique of Woodham [1980]. Rushmeier and Bernardini also employ a separate shape camera with a structured light source to obtain large-scale geometry, which they went to considerable effort to align with the captured texture information. Their system employed 5 tungsten-halogen sources, so they could dismiss up to 2 lights that were shadowed or caused specular reflection and still have enough information to recover the surface normal at a pixel. Ours is not so much an improvement on their method, as a simplified approach for a different application. Since our goal is local depth and surface normal variations, we do not require the 3-D geometry capture equipment or registration software, and our single-perspective diffuse plus flash images are sufficient for us to

hallucinate depth at each pixel. To avoid specular highlights, we employ crossed polarizers as suggested by [Glencross et al. 2008], and interpolate normals over pixels that are shadowed in one or more captures.

Our technique also bears close resemblance to the material capture work of Paterson et al. [2005]. Using photometric stereo in combination with surface normal integration and multiple view captures, these researchers were able to recover displacement maps plus inhomogeneous BRDFs over nearly planar sample surfaces using a simple flash plus camera arrangement. Their method incorporates a physical calibration frame around the captured surface to recover camera pose and flash calibration data. In contrast, our method uses only single-view capture, and flash/lens calibration is performed in advance, thus avoiding any restrictions of surface dimensions. Since we do not rely on surface normal integration to derive height information, our method is more robust to flash shadowing and irregular or spiky terrain. Similar to their technique, we assume a nearly planar surface with primarily diffuse reflection, and capture under ambient conditions. However, we make no attempt to recover specular characteristics in our method, which would be difficult from a single view.

Figure 2. Three-flash capture system mounted on a tripod with a digital SLR camera.

Multiple flashes have also been used to produce non-photorealistic imagery. Specifically, Raskar et al. developed a method for enhancing photographic illustrations exploiting the shadows cast by multiple flashes [Raskar et al. 2004]. Toler-Franklin et al. employed photometric stereo to capture surface normals, then applied these to enhance and annotate photo-based renderings [Toler-Franklin et al. 2007]. With the additional depth information our technique provides from the same data, it could be applied in a similar way to the problem of non-photorealistic rendering, though that is not our focus.

3 METHOD

Our technique borrows from and improves upon previous methods by employing a digital camera with three external flash units. We build on the flash/no-flash depth hallucination method of Glencross et al. [2008] by capturing two additional flash images to derive surface normal information and overcome limitations in their original albedo estimation. Employing three flashes virtually guarantees that every point on the surface will be illuminated in at least one image, and for points lit by all three flashes, we can accurately measure the surface normal as well. This normal map is used to correct the albedo estimate and further enhance re-rendering under different lighting conditions.

We begin by describing our three-flash capture system, followed by a description of the capture process and how the images are processed into a detailed surface model.

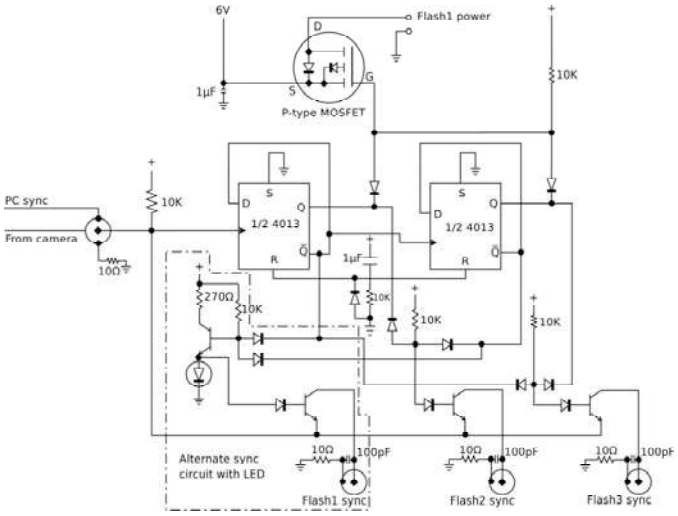

Figure 3. Circuit diagram for our three-flash controller.

3.1 Three-Flash Controller

To automatically sequence each flash, we built the simple controller circuit shown in Figure 3 to fire each flash in sequence, followed by a no-flash capture. In our configuration, we cycle the power to a shoe-mounted flash to force the camera into ambient exposure mode for the no-flash capture. This avoids having to touch the camera or control it via a USB tether – a tripod and a remote release cable are the only additional equipment required. The hot-shoe flash sync is controlled by the camera, so it fires while it has power. Therefore, some additional image processing is required for this set-up, which we explain in Section C, below.

A full cycle is achieved after 4 shutter releases. The first shutter release fires Flash 1 mounted on the hot-shoe only. The second shutter release fires Flash 2 as well, and the third shutter release fires Flashes 1 and 3. After three firings, power is turned off to Flash 1 mounted on the hot-shoe, thus putting the camera into ambient exposure mode, and none of the flashes fire. Once this final no-flash image has been captured, the cycle repeats.

Figure 2 shows our capture system mounted on a tripod. An amber LED indicates the controller is powered in its initial state, ready to begin a capture sequence. Linear polarizers are placed over each flash unit and aligned 90° out-of-phase with a polarizer filter mounted on the lens in order to reduce specular reflections as suggested in [Glencross et al. 2008].

3.2 Capture Process

The hot-shoe mounted flash is set to half its maximum output in manual mode, while the other two flashes are set to maximum. Since the hot-shoe flash fires every time, setting its output to half prevents it from drowning out the other flashes when they fire. Sufficient time is allowed between shutter releases for the flashes to fully recharge, ensuring that they produce roughly the same output each time. A cable release is used to avoid any camera movement, which would make subsequent image processing more difficult. After the full sequence of 4 images is captured and the histograms are checked to ensure a good set of exposures, the capture process is complete.

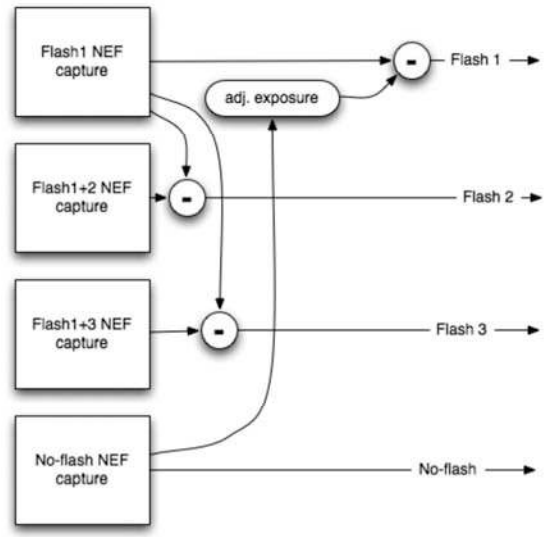

Figure 4. Diagram of RAW capture processing with dark subtraction used to obtain three separate flash no-flash images.

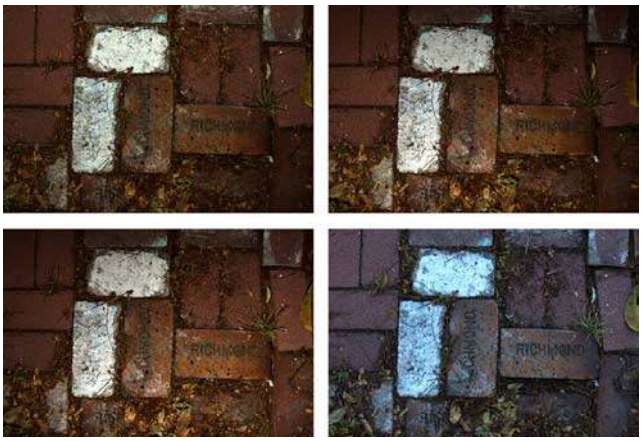

Figure 5. Our three separate flash images with the no-flash image in the lower right, all after RAW processing.

3.3 Image Processing

The first stage of our image-processing pipeline converts RAW captures to 16-bit/channel linear encoded TIFF. Taking advantage of the dark subtraction feature of **dcraw** [Coffin], we eliminate the effect of ambient lighting on our Flash 1 capture by subtracting the no-flash capture after applying the appropriate scale factor to account for differences in exposure time. We use this same trick to separate flash images by subtracting the Flash1-only capture from the Flash 1+2 and Flash 1+3 captures. Since Flash 1 also includes the ambient lighting, this takes care of the whole process for Flashes 2 and 3. This conversion is illustrated in Figure 4, with results shown in Figure 5.

The second stage of our image processing applies a calibration to the flash images to correct for vignetting and other uniformity issues. Since this correction varies with distance, aperture, lens and focal length, we capture a set of 50 to 100 reference flash images of a white, diffuse wall, then interpolate these calibration images to obtain a more accurate result. This interpolation process pulls out the six nearest flash triplets from our set and applies a weighted average to these. We then divide each flash image by its interpolated calibration image as in [Glencross et al. 2008] in preparation for the next processing stage.

In the third image processing stage, we simultaneously obtain local surface orientation (normals) and albedo (reflectance) by solving the following 3x3 matrix equation at each pixel illuminated by all three flashes [Rushmeier and Bernardini 1999]:

$$\mathbf{V}\vec{\mathbf{n}} = \vec{\mathbf{i}} \tag{1}$$

where:

$\mathbf{V}$ = illumination direction matrix

$\vec{\mathbf{n}}$ = normal vector times albedo

$\vec{\mathbf{i}}$ = adjusted flash pixel values

The computed adjusted flash pixel values are the corrected luminance values for each flash capture, multiplied again by the cosine of the incident angle, which was undone by our flash calibration. We compute the illumination direction matrix $\mathbf{V}$ by subtracting the estimated 3-D pixel positions given by our lens focal length and focus distance (recorded in the image metadata) from the known flash positions. We normalize each of these vectors, thus our measured pixels in $\vec{\mathbf{i}}$ are proportional to the dot product of the illumination vectors with the surface normal, times albedo. Solving for $\vec{\mathbf{n}}$ at each pixel, we take this vector length as our local variation in albedo. In shadow regions where only two flashes illuminate the surface, a technique such as [Hernández et al. 2008] could be used to resolve normals via an integrability constraint. We found that a simple hole-filling algorithm that averaged the four closest neighbors worked well enough in shadow regions, thanks to the masking from texture complexity that hides small artifacts.

A global scale factor may be applied to ensure an expected range of albedo values as a final step if necessary. Similarly, we found that applying a global flattening of the derived surface normals improves later rendering. This is accomplished by subtracting a low-frequency (blurred) version of the normal map from the high-resolution original, providing local detail while suppressing systematic errors due to imperfect calibration.

The fourth and final stage exactly follows the method laid out by [Glencross et al. 2008] to hallucinate depth using a multi-scale model based on the no-flash image divided by the albedo image. The important differences here are that we have a better estimate of albedo based on our knowledge of local surface orientation, and our multiple flashes avoid areas of complete shadow.

Figure 6. Left image contains depth hallucinated from a single flash/no-flash pair. Right image shows results of 3-flash system. Center is a photograph.

Figure 7. Comparison of hallucination and rendering methods showing the original diffuse photo, single-flash re-rendered result, three-flash depth result, and finally the three-flash result with derived normals.

4 RESULTS

4.1 Comparison to Single-flash Method

Figure 7 shows a side-by-side comparison between our three-flash method and the previous method of Glencross et al. [2008]. The upper-left image shows the original no-flash (diffusely lit) photograph. The upper-right image shows a rendering under simulated daylight using depth hallucinated with a single flash image and this diffuse photo. The lower-left image shows the same rendering using depth hallucinated from all three flashes, but without taking advantage of the derived surface normals. The final image on the lower-right shows the same improved depth map with derived normal information.

While we expected some slight improvements to the depth hallucination using three flashes, we found that most of the visible differences in the result came when we applied the derived surface

normals during rendering. Figures 1 and 6 show two additional comparisons. In Figure 6, specularity has been added to highlight the improvement to surface normals.

Since the extra flashes also eliminated problems with shadows, there were some cases where we noted missing features in the single-flash method that were recovered with three flashes. The close-up comparison of a wicker fence in Figure 8 shows one such example. Even though these renderings are created without using the normal information, the definition of the twigs is much finer in the three-flash result, thanks to the reduction of flash shadows.

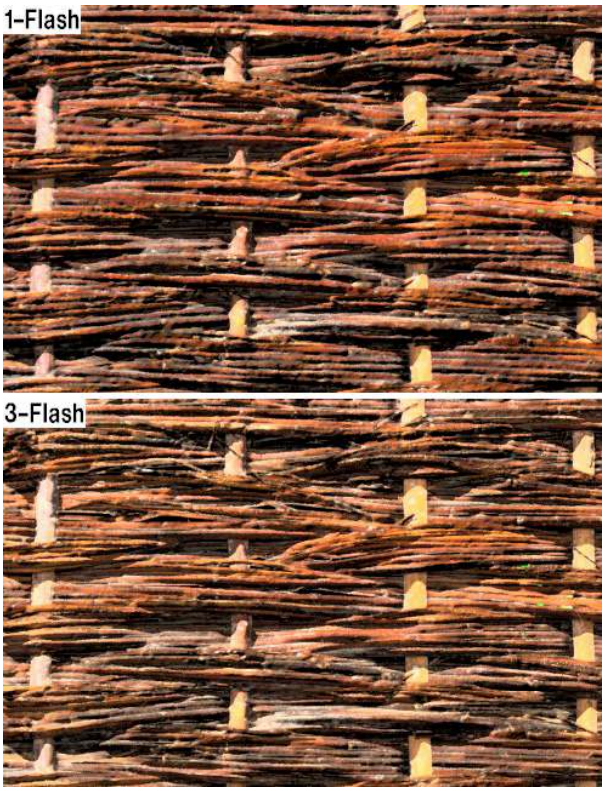

Figure 8. Depth hallucination on a wicker fence with single-flash and three-flash method, neither rendered with surface normals.

4.2 User Validation Study

We replicated the experimental validation study undertaken by Glencross et al. [2008]. The objective of this study was to assess the immediate impression of users when asked if the presented image was a photograph or a rendered image. In our case, we were seeking to identify how much the addition of measured surface orientation improves user perception of synthetically rendered images.

Users were seated in front of a color-calibrated display and sequentially shown a randomized series of synthetically re-rendered models and equivalent day-lit photographs. As in [Glencross et al. 2008], the *Radiance* physically-based rendering system [Ward 1994] was employed. Each user was asked to rate the presented images based on their certainty that the image was an un-touched photograph. This rating was performed on an integer scale ranging from 1 – 5. Participants were instructed in advance of the session to give a low rating if they were sure the image was computer generated. On the other hand, if they were certain the image was a photograph then they were instructed to choose a high number. A rating of 3 corresponded to the user being unable to decide if the image was computer generated or a photograph. All image stimuli were displayed on a 13.3-inch (diagonal) LED-backlit glossy screen MacBook at a resolution of 1280 x 800.

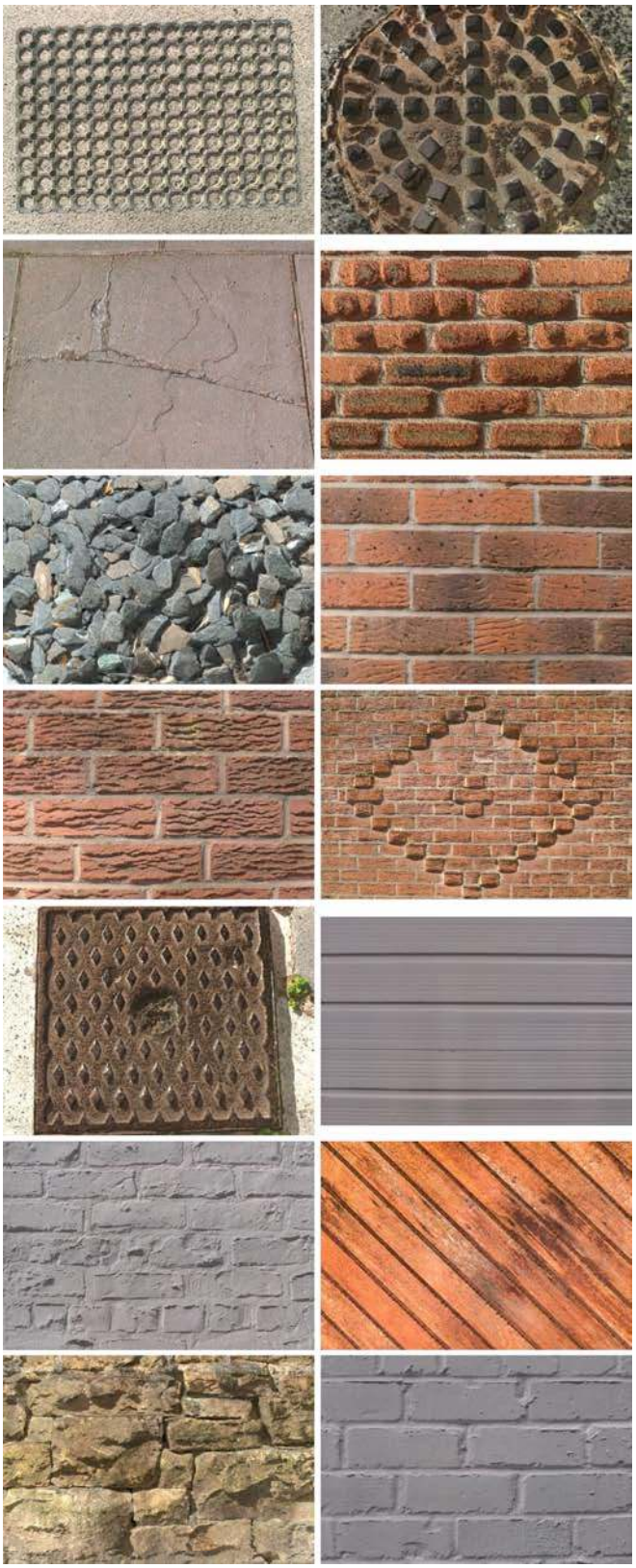

Figure 9. Three-flash hallucinations showing our 14 test scenes.

Figure 9 shows our fourteen distinct sets of textures (scenes) covering a range of materials from brick and stone to woven willow hurdles and wood. Within each set we included three conditions consisting of one day-lit photograph, one synthetically re-lit model captured with the single-flash method, and one synthetically re-lit model combining photometrically measured surface orientation using the three-flash method. For each test scene, the lighting in the synthetic images was matched visually to the day-lit equivalent photographs. Our hypothesis was that the three-flash and photo conditions would be perceptually equivalent.

A total of forty-two stimuli (images) were presented in randomized order for exactly three seconds each. After each image was presented, the user was automatically taken to a rating screen and asked to enter their choice of numeric rating before the next image was displayed, and in turn rated. Each user's image ratings were saved to a file with a unique anonymous ID that could be correlated to a post study questionnaire. We collected study data from a total of fifteen participants aged between eighteen and sixty-five. None of the study subjects were expert in computer graphics, photography, nor had any prior knowledge of this work.

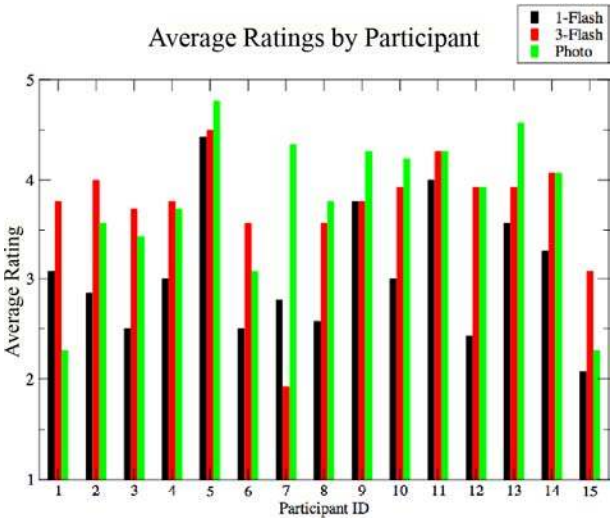

Figure 10. Graph showing average ratings per participant for each class of stimulus.

All participants were asked to fill in a post study questionnaire. In this, we collected a number of demographic metrics. These were collected in order to profile participants and account for any specific biases in the study results. The post study questionnaire asked about participants' expertise, gender, whether or not they had any problems with their vision, how regularly they played computer games, level of interest in photography and other similar questions. In addition, we also asked participants to report on their perceived difficulty in distinguishing between the synthetic images and photographs (on a scale between easy, difficult and could not tell).

As shown in Figure 10, nine out of fifteen participants gave higher or equivalent mean ratings to images generated using our three-flash method. Thirteen of the fourteen textured scenes achieved average ratings above 3.0, for synthetically re-lit images using models captured with our method, as shown in Figure 11. Our synthetically re-lit images were rated on average higher than the equivalent day-lit photographs for seven of our tested scenes, and outperformed the single-flash hallucination results for all the scenes that we tested.

Overall mean values for each class of stimuli were: single-flash = 3.0571, three-flash = 3.7238 and photos = 3.7762. We tested our results for statistical significance. A repeated measures ANOVA with pair-wise comparisons showed a significant main effect of image type $F(2,28) = 10.1$, $p<0.001$, with the pair-wise comparisons showing that the difference is significant between the single- and three-flash conditions, and the single-flash images and photos, but not significant between the three-flash images and photos.

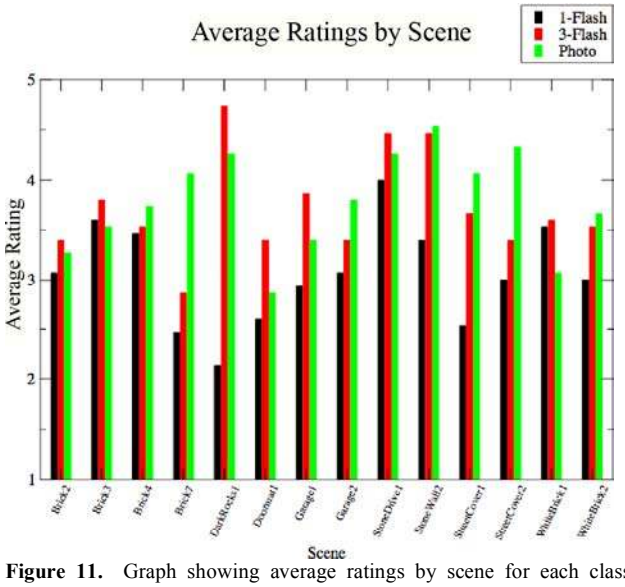

Figure 11. Graph showing average ratings by scene for each class of stimulus.

In addition to the metrics measured and reported by [Glencross et al. 2008], recall that we asked participants to rate the difficulty in choosing the images they believed to be un-touched photographs. Thirteen of the fifteen participants stated the task was difficult, only two reported they found the task easy (participants number 7 and 11) and no one stated they could not tell. The graph in Figure 10 shows that participant number 7 did in fact find it easy to distinguish the photographs from the synthetically re-lit images. In the post, study questionnaire participant 7 reported a keen interest in art. An explanation for this participant's ability to reliably distinguish the photos could be a heightened awareness of lighting gained from experience of painting. On the other hand, participant number 11's mean ratings indicate each class of stimuli to be more likely to be un-touched photographs and so could not reliably tell despite a contradictory perception of the ease of the task. At least 9 of the 15 participants' ratings strongly suggest they could not distinguish between the three-flash synthetically re-lit images and photographs, and all but one participant (number 7) showed no reliable ability to identify our synthetic results.

4.3 Failure Case

Naturally, we found failure cases during our capture trials. The most interesting were certain natural wood finishes, which exhibited optical diffraction that foiled our polarization filters. Figure 12 shows one such example, where specularities show through the polarizers and produce bogus depth, albedo, and surface normal maps. Such obvious failures were excluded from our user study, as we consider this an area for future work. In the case of optical diffraction, it may be necessary to revert to the original single-flash method with no polarizers.

Figure 12. Optical diffraction causing unexpected results.

5 CONCLUSIONS

We have demonstrated a significant perceptual improvement over the single-flash capture method of Glencross et al. [2008], obtaining a better estimate of albedo as well as providing detailed surface normal information. Ensuring every pixel is lit by at least one of two extra flash units reduces the appearance of perceptually important errors associated with filled in detail, and permits cosine correction of the albedo estimates. This in turn provides more stable information for surface depth hallucination. The main visual benefit to the final rendered results comes from the derived surface orientation. This better captures important edge cues when combined with our improved depth estimates. The results of our experimental study illustrate the importance of capturing this edge information in surface models obtained from photographs. Using an appropriate physically-based rendering method, the resulting synthetic imagery is statistically indistinguishable from equivalent photographs. This is an important result, because it gives insight into the level of 3D model fidelity we must recover from photos. Although our depth estimates are crude approximates, they spatially match the visual information conveyed in the texture image. Combined with estimates of surface orientation, this leads to entirely plausible self-shadowing.

In contrast to physically similar capture devices that rely purely on photometric stereo, we hallucinate depth from the diffuse lighting condition rather than integrating surface normals to obtain topography. The multiple flashes augment our results with normals and improved albedo estimates where available, but are not critical at every point on the surface.

While we employed three flashes, benefits may be drawn from additional flash units, which ameliorate most shadow-born defects. Similar to Rushmeier and Bernardini [1999] who used five sources, more flash units could provide complete surface normal data and might improve accuracy via a least-squares solution to an overdetermined version of Eq. (1).

Finally, we would like to develop a technique similar to the histogram matching method of [Glencross et al. 2008] to derive surface orientation without the benefits of flash photography, which is impractical beyond a five or six meters. Capturing an entire building would be very difficult using any flash-based method, and we believe there is a way to correlate surface normals with local depth and shading as obtained by the histogram matching method on exemplar data.

ACKNOWLEDGMENTS

The authors give generous thanks to our collaborators Napper Architects and Insight Digital. Thanks also to Prof. Roger Hubbold and Francisco Melendez for their invaluable support and Dr. Caroline Jay for assistance with data analysis. Funding for this work from the UK Engineering and Physical Sciences Research Council, grant number EP/D069734/1 is gratefully acknowledged.

REFERENCES

OLIVEIRA, M., BISHOP, G., AND MCALLISTER, D., 2000. Relief texture mapping. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 359–368.

RUSHMEIER, H., GOMES, J., GIORDANO, F., EL-SHISHINY, H., MAGERLEIN, K., AND BERNARDINI, F., 2003. Design and use of an in-museum system for artifact capture. *IEEE/CVPR Workshop on Applications of Computer Vision in Archaeology*, 8–8.

LENSCH, H., KAUTZ, J., GOESELE, M., HEIDRICH, W., AND SEIDEL, H.-P., 2003. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics (TOG)*, 22, 2, 234–257.

DANA, K., VAN GINNEKEN, B., NAYAR, S., AND KOENDERINK, J., 1999. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, 18, 1, 1–34.

MARSCHNER, S., WESTIN, S., LAFORTUNE, E., TORRANCE, K., AND GREENBERG, D., 1999. Image-based BRDF measurement including human skin. *Eurographics Workshop on Rendering*, 139–152.

GLENCROSS, M., WARD, G., JAY, C., LIU, J., MELENDEZ, F., AND HUBBOLD, R., 2008. A Perceptually Validated Model for Surface Depth Hallucination. *ACM Transactions on Graphics (TOG)*, 27, 3, 59:1 - 59:8.

RUSHMEIER., AND BERNARDINI, F., 1999. Computing consistent normals and colors from photometric data. *Proceeding of the Second International Conference on 3-D Digital Imaging and Modeling 3DIM*, IEEE, 99–108.

WOODHAM, R. J., 1980. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19, 1, 139–44.

PATERSON, J. A., CLAUS, D., AND FITZGIBBON, A. W., 2005. BRDF and geometry capture from extended inhomogeneous samples using flash photography. *Computer Graphics Forum (Eurographics)*, 24, 383 – 391.

RASKAR, R., TAN, K-H., FERIS, R., YU, J., AND TURK, M., 2004. Non-photorealistic camera: depth edge detection and stylized rendering using multi-flash imaging. *ACM Transactions on Graphics (TOG)*, 23, 3, 676–685.

TOLER-FRANKLIN, C., FINKELSTEIN, A., AND RUISINKIEWICZ, S., 2007. llustration of complex real-world objects using images with normals. *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, 111– 119.

COFFIN, D., 2008. DCRAW. *Wikipedia online document*, Active July 2008, en.wikipedia.org/wiki/Dcraw

HERNÁNDEZ, C., VOGIATZIS, G., AND CIPOLLA, R., 2008. Shadows in three-source photometric stereo. *IEEE European Conference on Computer Vision*, Marseille, Poster session.

WARD, G., 1994. The RADIANCE Lighting Simulation and Rendering System. *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 459-472.