1998 IEEE. Reprinted, with permission, from Proc. International Conference on Computer Vision
Bombay, India - January 1998 - pp. 43-50

3D photography on your desk

Jean-Yves Bouguet† and Pietro Perona†‡

† California Institute of Technology, 136-93, Pasadena, CA 91125, USA
‡ Università di Padova, Italy
{bouguetj,perona}@vision.caltech.edu

Abstract

A simple and inexpensive approach for extracting the three-dimensional shape of objects is presented. It is based on 'weak structured lighting'; it differs from other conventional structured lighting approaches in that it requires very little hardware besides the camera: a desk-lamp, a pencil and a checkerboard. The camera faces the object, which is illuminated by the desk-lamp. The user moves a pencil in front of the light source casting a moving shadow on the object. The 3D shape of the object is extracted from the spatial and temporal location of the observed shadow. Experimental results are presented on three different scenes demonstrating that the error in reconstructing the surface is less than 1%.

1 Introduction and Motivation

One of the most valuable functions of our visual system is informing us about the shape of the objects that surround us. Manipulation, recognition, and navigation are amongst the tasks that we can better accomplish by seeing shape. Ever-faster computers, progress in computer graphics, and the widespread expansion of the Internet have recently generated much interest in systems that may be used for imaging both the geometry and surface texture of object. The applications are numerous. Perhaps the most important ones are animation and entertainment, industrial design, archiving, virtual visits to museums and commercial on-line catalogues.

In designing a system for recovering shape, different engineering tradeoffs are proposed by each application. The main parameters to be considered are: cost, accuracy, ease of use and speed of acquisition. So far, the commercial 3D scanners (e.g. the Cyberware scanner) have emphasized accuracy over the other parameters. These systems use motorized transport of the object, and active (laser, LCD projector) lighting of the scene, which makes them very accurate, but expensive and bulky [1, 15, 16, 12, 2].

An interesting challenge for computer vision researchers is to take the opposite point of view: emphasize cost and simplicity, perhaps sacrificing some amount of accuracy, and design 3D scanners that demand little more hardware than a PC and a video camera, by now almost standard equipment both in offices and at home, by making better use of the data that is available in the images.

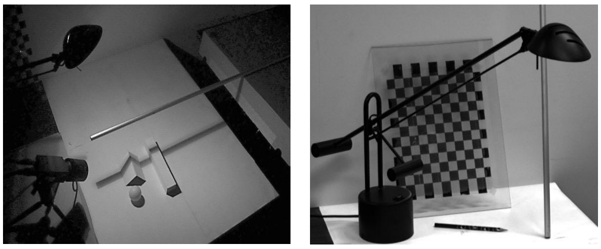

Figure 1: **The general setup of the proposed method:** The camera is facing the scene illuminated by a halogen desk lamp (left). The scene consists of objects on a plane (the desk). When an operator freely moves a stick in front of the lamp (over the desk), a shadow is cast on the scene. The camera acquires a sequence of images $I(x, y, t)$ as the operator moves the stick so that the shadow scans the entire scene. This constitutes the input data to the 3D reconstruction system. The variables x and y are the pixel coordinates (also referred to as spatial coordinates), and t the time (or frame number). The three dimensional shape of the scene is reconstructed using the spatial and temporal properties of the shadow boundary throughout the input sequence. The right-hand figure shows the necessary equipment besides the camera: a desk lamp, a calibration grid and a pencil for calibration, and a stick for the shadow. One could use the pencil instead of the stick.

A number of passive cues have long been known to contain information on 3D shape: stereoscopic disparity, texture, motion parallax, (de)focus, shadows, shading and specularities, occluding contours and other surface discontinuities amongst them. At the current state of vision research stereoscopic disparity is the single passive cue that gives reasonable accuracy. Unfortunately it has two major drawbacks: (a) it requires two cameras thus increasing complexity and cost, (b) it cannot be used on untextured surfaces (which are common for industrially manufactured objects).

We propose a method for capturing 3D surfaces that is based on 'weak structured lighting'. It yields good accuracy and requires minimal equipment besides a computer and a camera: a pencil (two uses), a checkerboard and a desk-lamp – all readily available in most homes; some intervention by a human operator, acting as a low precision motor, is also required.

We start with a description of the method in Sec. 2, followed in Sec. 3 by a noise sensitivity analysis, and in Sec. 4 by a number of experiments that assess the convenience and accuracy of the system. We end with a discussion and conclusions in Sec. 5.

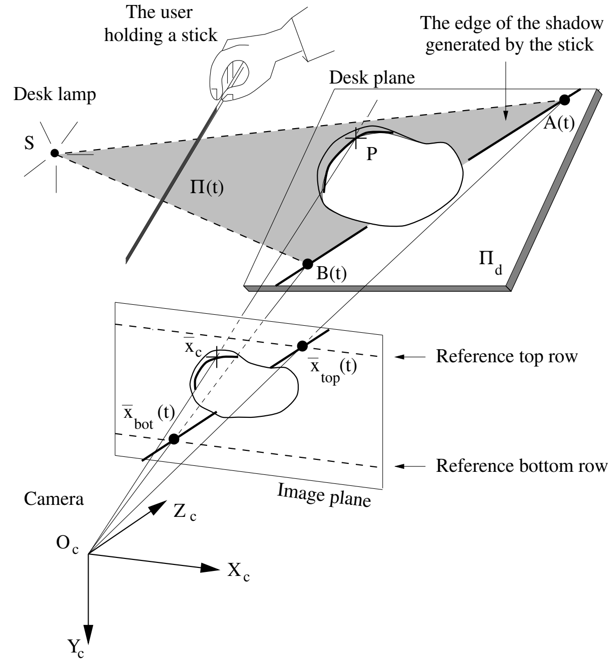

Figure 2: **Geometrical principle of the method:** Approximate the light source with a point S, and denote by Π_d the desk plane. Assume that the positions of the light source S and the plane Π_d in the camera reference frame are known from calibration. The goal is to estimate the 3D location of the point P in space corresponding to every pixel $\overline{x}_c$ in the image. Call t the time at which a given pixel $\overline{x}_c$ 'sees' the shadow boundary (later referred to as the shadow time). Denote by $\Pi(t)$ the corresponding shadow plane at that time t. Assume that two portions of the shadow projected on the desk plane are visible on two given rows of the image (top and bottom rows in the figure). After extracting the shadow boundary along those rows $\overline{x}_{\text{top}}(t)$ and $\overline{x}_{\text{bot}}(t)$, we find two points on the shadow plane $A(t)$ and $B(t)$ by intersecting Π_d with the optical rays $(O_c, \overline{x}_{\text{top}}(t))$ and $(O_c, \overline{x}_{\text{bot}}(t))$ respectively. The shadow plane $\Pi(t)$ is then inferred from the three points in space S, $A(t)$ and $B(t)$. Finally, the point P corresponding to $\overline{x}_c$ is retrieved by intersecting $\Pi(t)$ with the optical ray $(O_c, \overline{x}_c)$. This final stage is called triangulation. Notice that the key steps in the whole scheme are: (a) estimate the shadow time $t_s(\overline{x}_c)$ at every pixel $\overline{x}_c$ (*temporal processing*), and (b) locate the reference points $\overline{x}_{\text{top}}(t)$ and $\overline{x}_{\text{bot}}(t)$ at every time instant t (*spatial processing*). These two are discussed in detail in section 2.2.

2 Description of the method

The general principle consists of casting a shadow onto the scene with a pencil or another stick, and using the image of the deformed shadow to estimate the three dimensional shape of the scene. Figure 1 shows the required hardware and the setup of the system. The objective is to extract scene depth at every pixel in the image. Figure 2 gives a geometrical description of the method that we propose to achieve that goal.

2.1 Calibration

The goal of calibration is to recover the geometry of the setup (that is, the location of the desk plane Π_d and that of the light source S) as well as the *intrinsic parameters* of the camera (focal length, optical center and radial distortion factor). We decompose the procedure into two successive stages: first camera calibration and then lamp calibration.

Camera calibration: Estimate the intrinsic camera parameters and the location of the desk plane Π_d (tabletop) with respect to the camera. The procedure consists of first placing a planar checkerboard pattern (see figure 1) on the desk in the location of the objects to scan. From the image captured by the camera, we infer the *intrinsic* and *extrinsic* (rigid motion between camera and desk reference frame) parameters of the camera, by matching the projections onto the image plane of the known grid corners with the expected projection directly measured on the image (extracted corners of the grid). This method is very much inspired by the algorithm proposed by Tsai [13]. Note that since our calibration rig is planar, the optical center cannot be recovered through that process, and therefore is assumed to be fixed at the center of the image. A description of the whole procedure can be found in [3]. The reader can also refer to Faugeras [6] for further insights on camera calibration. Notice that the extrinsic parameters directly lead to the position of the tabletop Π_d in the camera reference frame.

Lamp calibration: After camera calibration, estimate the 3D location of the point light source S. Figure 3 gives a description of our method.

2.2 Spatial and temporal shadow edge localization

A fundamental stage of the method is the detection of the line of intersection of the shadow plane $\Pi(t)$ with the desktop Π_d; a simple approach may be used if we make sure that the top and bottom edges of the image are free from objects. Then the two tasks to accomplish are: **(a)** Localize the edge of the shadow that is directly projected on the tabletop $(\overline{x}_{\text{top}}(t), \overline{x}_{\text{bot}}(t))$ at every time instant t (every frame), leading to the set of all shadow planes $\Pi(t)$, **(b)** Estimate the time $t_s(\overline{x}_c)$ (*shadow time*) where the edge of the shadow passes through any given pixel $\overline{x}_c = (x_c, y_c)$ in the image. Curless and Levoy demonstrated in [4] that such a spatio-temporal approach is appropriate to preserve sharp discontinuities in the scene. Details of our implementation are given in figure 4. Notice that the shadow was scanned from the left to the right side of the scene. This explains why the right edge of the shadow corresponds to the front edge of the temporal profile in figure 4.

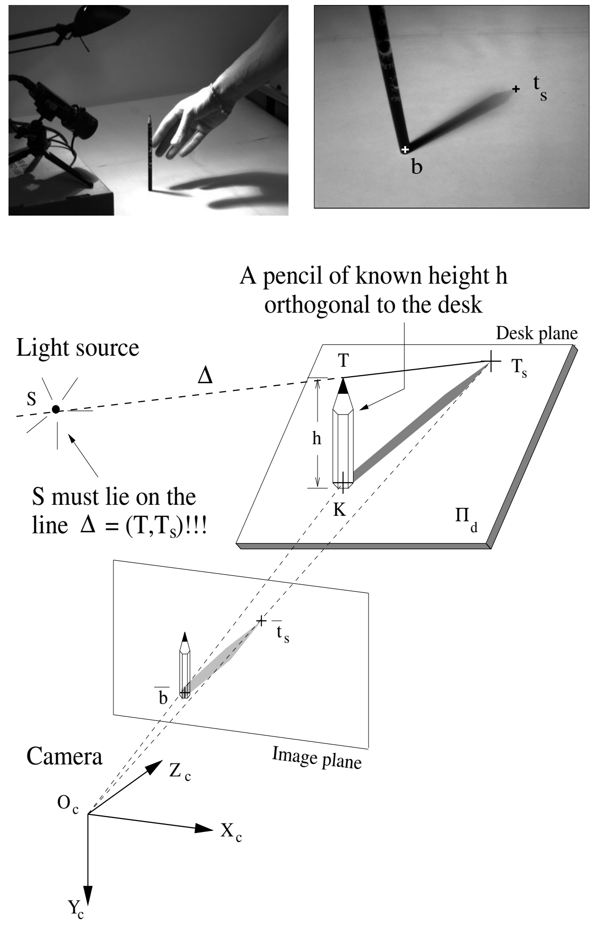

Figure 3: **Lamp calibration:** The operator places a pencil on the desk plane Π_d, orthogonal to it (top-left). The camera observes the shadow of the pencil projected on the tabletop. The acquired image is shown on the top-right. From the two points $\bar{b}$ and $\bar{t}_s$ on this image, one can infer the positions in space of K and T_s, respectively the base of the pencil, and the tip of the pencil shadow (see bottom figure). This is done by intersecting the optical rays $(O_c, \bar{b})$ and $(O_c, \bar{t}_s)$ with Π_d (known from camera calibration). In addition, given that the height of the pencil h is known, the coordinates of its tip T can be directly inferred from K. Then, the light source point S has to lie on the line $\Delta = (T, T_s)$ in space. This yields one linear constraint on the light source position. By taking a second view, with the pencil at a different location on the desk, one can retrieve a second independent constraint with another line Δ'. A closed form solution for the 3D coordinate of S is then derived by intersecting the two lines Δ and Δ' (in the least squares sense). Notice that since the problem is linear, one can easily integrate the information from more than 2 views and then make the estimation more accurate. If $N > 2$ images are used, one can obtain a closed form solution for the best intersection point $\tilde{S}$ of the N inferred lines (in the least squares sense). We also estimate the uncertainty on that estimate from the distance of $\tilde{S}$ from each one of the Δ lines. That indicates how consistently the lines intersect a single point in space. Refer to [3] for the complete derivations.

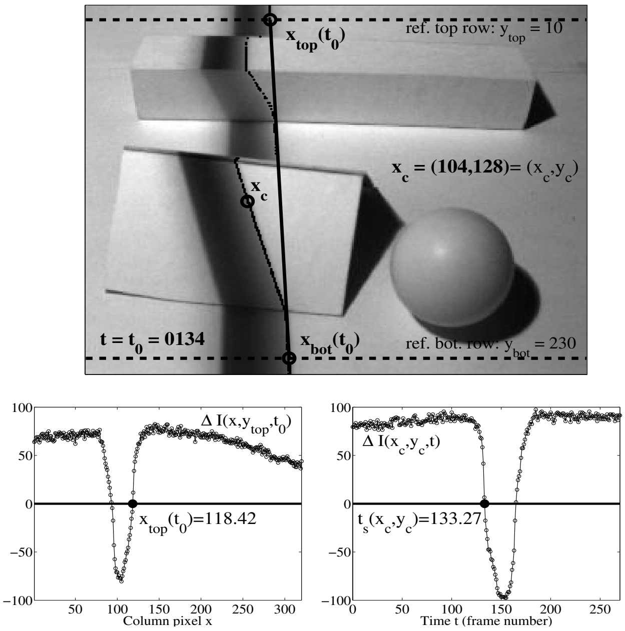

Figure 4: **Spatial and temporal shadow location:** The first step consists of localizing spatially the shadow edge $(\bar{x}_{\rm top}(t_0), \bar{x}_{\rm bot}(t_0))$ at every integer time t_0 (i.e. every frame). The top and bottom rows are $y_{\rm top} = 10$ and $y_{\rm bot} = 230$ on the top figure. This leads to an estimate of the shadow plane $\Pi(t_0)$ at every frame. The second processing step consists of extracting at every pixel $\bar{x}_c$, the time $t_s(\bar{x}_c)$ of passage of the shadow edge. For any given pixel $\bar{x}_c = (x, y)$, define $I_{\min}(x, y) \doteq \min_t (I(x, y, t))$ and $I_{\max}(x, y) \doteq \max_t (I(x, y, t))$ as its minimum and maximum brightness throughout the entire sequence. We then define the shadow edge to be the locations (in space-time) where the image $I(x, y, t)$ intersects with the threshold image $I_{\rm shadow}(x, y) \doteq (I_{\min}(x, y) + I_{\max}(x, y))/2$. This may be also regarded as the zero crossings of the difference image $\Delta I(x, y, t) \doteq I(x, y, t) - I_{\rm shadow}(x, y)$. The two bottom plots illustrate the shadow edge detection in the spatial domain (to find $\bar{x}_{\rm top}$ and $\bar{x}_{\rm bot}$) and in the temporal domain (to find $t_s(\bar{x}_c)$). The bottom-left figure shows the profile of $\Delta I(x, y, t)$ along the top reference row $y = y_{\rm top} = 10$ at time $t = t_0 = 134$ versus the column pixel coordinate x. The second zero crossing of that profile corresponds to the top reference point $\bar{x}_{\rm top}(t_0) = (118.42, 10)$ (computed at sub-pixel accuracy). Identical processing is applied on the bottom row to obtain $\bar{x}_{\rm bot}(t_0) = (130.6, 230)$. Similarly, the bottom-right figure shows the temporal profile $\Delta I(x_c, y_c, t)$ at the pixel $\bar{x}_c = (x_c, y_c) = (104, 128)$ versus time t (or frame number). The shadow time at that pixel is defined as the first zero crossing location of that profile: $t_s(104, 128) = 133.27$ (computed at sub-frame accuracy).

Notice that the pixels corresponding to regions in the scene that are not illuminated by the lamp (shadows due to occlusions) do not provide any relevant depth information. For this reason we can restrict the processing to pixels that have sufficient swing between maximum and minimum brightness. Therefore, we only process pixels with contrast value $I_{\rm contrast}(x, y) \doteq I_{\max}(x, y) - I_{\min}(x, y)$ larger than a pre-defined threshold $I_{\rm thresh}$. This threshold was 70 in all experiments reported in this paper (recall that the intensity values are encoded from 0 for black to 255 for white).

We do not apply any spatial filtering on the images; that would generate undesired blending in the final depth estimates, especially noticeable at depth discontinuities (at occlusions for example). However, it would be acceptable to low-pass filter the brightness profiles of the top and bottom rows (there is no depth discontinuity on the tabletop) and low-pass filter the temporal brightness profiles at every pixel. These operations would preserve sharp spatial discontinuities, and might decrease the effect of local processing noise by accounting for smoothness in the motion of the stick.

Experimentally, we found that this thresholding approach for shadow edge detection allow for some internal reflections in the scene [9, 8, 14]. However, if the light source is not close to an ideal point source, the mean value between maximum and minimum brightness may not always constitute the optimal value for the threshold image I_{shadow}. Indeed, the shadow edge profile becomes shallower as the distance between the stick and the surface increases. In addition, it deforms asymmetrically as the surface normal changes. These effects could make the task of detecting the shadow boundary points challenging. In the future, we intend to develop a geometrical model of extended light sources and incorporate it in the system.

Although $I_{\min}$ and $I_{\max}$ are needed to compute I_{shadow}, there exists an implementation of that algorithm that does not require storage of the complete image sequence in memory and therefore leads itself to real-time implementations. All that one needs to do is update at each frame five different arrays $I_{\max}(x,y)$, $I_{\min}(x,y)$, $I_{\text{contrast}}(x,y)$, $I_{\text{shadow}}(x,y)$ and the shadow time $t_s(x,y)$, as the images $I(x,y,t)$ are acquired. For a given pixel (x,y), the maximum brightness $I_{\max}(x,y)$ is collected at the very beginning of the sequence (the first frame), and then, as time goes, the incoming images are used to update the minimum brightness $I_{\min}(x,y)$ and the contrast $I_{\text{contrast}}(x,y)$. Once $I_{\text{contrast}}(x,y)$ crosses I_{thresh}, the adaptive threshold $I_{\text{shadow}}(x,y)$ starts being computed and updated at every frame (and activated). This process goes on until the pixel brightness $I(x,y,t)$ crosses $I_{\text{shadow}}(x,y)$ for the first time (in the upwards direction). That time instant is registered as the shadow time $t_s(x,y)$. In that form of implementation, the left edge of the shadow is tracked instead of the right one, however the principle remains the same.

2.3 Triangulation

Once the shadow time $t_s(\overline{x}_c)$ is estimated at a given pixel $\overline{x}_c$, one can identify the corresponding shadow plane $\Pi(t_s(\overline{x}_c))$. Then, the 3D point P associated to $\overline{x}_c$ is retrieved by intersecting $\Pi(t_s(\overline{x}_c))$ with the optical ray $(O_c, \overline{x}_c)$ (see figure 2). Notice that the shadow time $t_s(\overline{x}_c)$ acts as an index to the shadow plane list $\Pi(t)$. Since $t_s(\overline{x}_c)$ is estimated at sub-frame accuracy, the final plane $\Pi(t_s(\overline{x}_c))$ actually results from linear interpolation between the two planes $\Pi(t_0 - 1)$ and $\Pi(t_0)$ if $t_0 - 1 < t_s(\overline{x}_c) < t_0$ and t_0 integer. Once the range data are recovered, a mesh may be generated by connecting neighboring points in triangles. Rendered views of three reconstructed surface structures can be seen in figures 6, 7 and 8.

3 Noise Sensitivity

The overall scheme is based on first extracting from every frame (i.e. every time instants t) the x coordinates of the two reference points $x_{\text{top}}(t)$ and $x_{\text{bot}}(t)$, and second estimating the shadow time $t_s(\overline{x}_c)$ at every pixel $\overline{x}_c$. Those input data are used to estimate the depth Z_c at every pixel. The purpose of the noise sensitivity analysis is to quantify the effect of the noise in the measurement data $\{x_{\text{top}}(t), x_{\text{bot}}(t), t_s(\overline{x}_c))\}$ on the final reconstructed scene depth map. One key step in the analysis is to transfer the noise affecting the shadow time $t_s(\overline{x}_c)$ into a scalar noise affecting the x coordinate of $\overline{x}_c$ after scaling by the local shadow speed on the image at that pixel. Let V be the volume of the parallelepiped formed by the three vectors $\overline{O_cA}$, $\overline{O_cB}$ and $\overline{O_cS}$, originating at O_c (see figure 2):

$$V = \overline{X}_S^T \cdot \left\{(\overline{X}_B - \overline{X}_S) \times (\overline{X}_A - \overline{X}_S)\right\}$$

where $\overline{X}_S = [X_S \ Y_S \ Z_S]^T$, $\overline{X}_A = [X_A \ Y_A \ Z_A]^T$ and $\overline{X}_B = [X_B \ Y_B \ Z_B]^T$ are the coordinate vectors of S, A and B in the camera reference frame (× is the standard outer product operator). Notice that V is computed at the triangulation stage, and therefore is always available (see [3]). Define $\overline{X}_c = [X_c \ Y_c \ Z_c]^T$ as the coordinate vector in the camera reference frame of the point in space corresponding to $\overline{x}_c$. Assume that the x coordinates of the top and bottom reference points (after normalization) are affected by additive white Gaussian noise with zero mean and variances σ_{t}^2 and σ_{b}^2 respectively. Assume in addition that the variance on the x coordinate of $\overline{x}_c$ is $\sigma_{x_c}^2$ (different at every pixel). The following expression for the variance $\sigma_{Z_c}^2$ of the induced noise on the depth estimate Z_c was derived by taking first order derivatives of Z_c with respect to the 'new' noisy input variables x_{top}, x_{bot} and $\overline{x}_c$ (notice that the time variable does not appear any longer in the analysis):

$$\sigma_{Z_c}^2 = \frac{Z_c^2}{V^2}\left\{W^2 h_S^2 Z_c^2 \sigma_{x_c}^2 + (\alpha_1 + \beta_1 Y_c + \gamma_1 Z_c)^2 \sigma_{\text{t}}^2 + (\alpha_2 + \beta_2 Y_c + \gamma_2 Z_c)^2 \sigma_{\text{b}}^2\right\} \quad (1)$$

where W, h_S, α_1, β_1, γ_1, α_2, β_2 and γ_2 are constants depending only on the geometry (see figure 5):

$$\begin{aligned}
\alpha_1 &= Z_A\,(Z_B\,Y_S - Y_B\,Z_S)\\
\beta_1 &= -Z_A\,(Z_B - Z_S)\\
\gamma_1 &= Z_A\,(Y_B - Y_S)\\
\alpha_2 &= Z_B\,(Y_A\,Z_S - Z_A\,Y_S)\\
\beta_2 &= Z_B\,(Z_A - Z_S)\\
\gamma_2 &= -Z_B\,(Y_A - Y_S)
\end{aligned}$$

The first term in equation 1 comes from the temporal noise (on $t_s(\overline{x}_c)$ transferred to $\overline{x}_c$); the second and third terms from the spatial noise (on $\overline{x}_{\text{top}}$ and $\overline{x}_{\text{bot}}$). Let σ_I be the standard deviation of the image brightness noise. Given that we use linear interpolation of the temporal brightness profile to calculate the shadow time $t_s(\overline{x}_c)$, we can write σ_{x_c} as a function of the horizontal spatial image gradient $I_x(\overline{x}_c)$ at $\overline{x}_c$ at time $t = t_s(\overline{x}_c)$:

$$\sigma_{x_c} = \frac{\sigma_I}{|I_x(\overline{x}_c)|} \tag{2}$$

Since σ_{x_c} in inversely proportional to the image gradient, the accuracy improves with shadow edge sharpness. This justifies the improvement in experiment 3 after removing the lamp reflector (thereby significantly increasing sharpness). In addition, observe that σ_{x_c} does not depend on the local shadow speed. Therefore, decreasing the scanning speed would not increase accuracy. However, for the analysis leading to equation 2 to remain valid, the temporal pixel profile must be sufficiently sampled within the transition area of the shadow edge (the penumbra). Therefore, if the shadow edge were sharper, the scanning should also be slower so that the temporal profile at every pixel would be properly sampled. Decreasing further the scanning speed would benefit the accuracy only if the temporal profile were appropriately low-pass filtered before extraction of $t_s(\overline{x}_c)$. This is an issue for future research.

Notice that σ_{Z_c}, aside from quantifying the uncertainties on the depth estimate Z_c at every pixel $\overline{x}_c$, it also constitutes a good indicator of the overall accuracies in reconstruction, since most of the errors are located along the Z direction of the camera frame. In addition, we found numerically that most of the variations in the variance $\sigma^2_{Z_c}$ are due to the variation of volume V within a single scan. This explains why the reconstruction noise is systematically larger in portions of the scene further away from the lamp (see figures 6, 7 and 8). Indeed, it can be shown that, as the shadow moves into the opposite direction of the lamp (e.g. to the right if the lamp is on the left of the camera), the absolute value of the volume $|V|$ strictly decreases, making $\sigma^2_{Z_c}$ larger (see [3] for details).

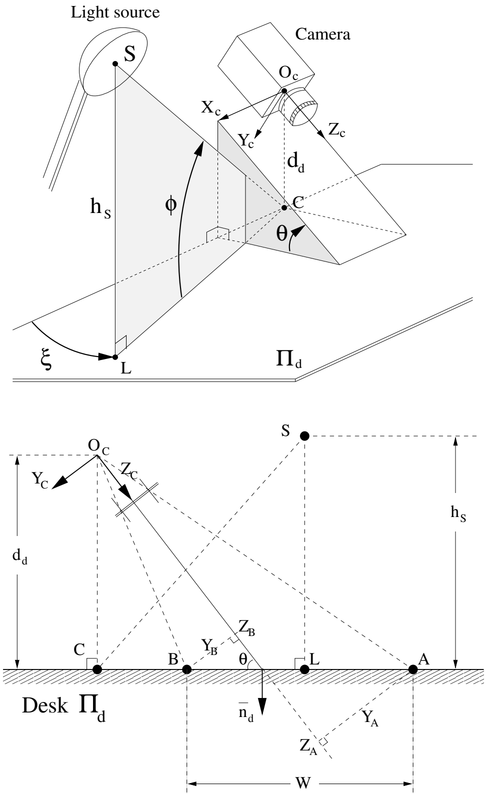

Figure 5: **Geometric setup:** The camera is positioned at a distance d_d away from the desk plane Π_d and tilted down towards it at an angle θ. The light source is located at a height h_S, with its direction defined by the azimuth and elevation angles ξ and ϕ. Notice the sign of $\cos\xi$ directly relates to which side of the camera the lamp is standing: positive on the right, and negative on the left. The bottom figure is a side view of the system (in the (O_c, Y_c, Z_c) plane). The points A and B are the reference points on the desk plane (see figure 2).

In order to obtain a uniformly accurate reconstruction of the entire scene, one may take two scans of the same scene with the lamp at two different locations (on the left (L) and on the right (R) of the camera), and merge them together using at each pixel the estimated reliability of the two measurements. Assume that the camera position, as well as the height h_S of the lamp, are kept identical for both scans. Suppose in addition that the scanning speeds were approximately the same. Then, at every pixel $\overline{x}_c$ in the image, the two scan data sets provide two estimates Z_c^L and Z_c^R of the same depth Z_c with respective reliabilities $\sigma^2_{Z_L}$ and $\sigma^2_{Z_R}$ given by equation 1. In addition, if we call V_L and V_R the two respective volumes, then the relative uncertainty between Z_c^L and Z_c^R reduces to a function of the volumes: $\sigma^2_{Z_R}/\sigma^2_{Z_L} = (V_L/V_R)^2$. Notice that calculating that relative uncertainty does not require any extra computation, since V_L and V_R are available from the two triangulations. The final depth is computed by weighted average of Z_c^L and Z_c^R: $Z_c \doteq \omega_L\, Z_c^L + \omega_R\, Z_c^R$. If Z_c^R and Z_c^L were Gaussian distributed, and independent, they would be optimally averaged using the inverse of the variances as weights [10]: $\omega_L = \sigma^2_{Z_R}/(\sigma^2_{Z_L} + \sigma^2_{Z_R}) = \alpha^2/(1+\alpha^2)$ and $\omega_R = \sigma^2_{Z_L}/(\sigma^2_{Z_L} + \sigma^2_{Z_R}) = 1/(1+\alpha^2)$, where $\alpha = V_L/V_R$. Experimentally, we found that this choice does not yield very good merged surfaces. It makes the noisy areas of one view interact too significantly with the clean corresponding areas in the other view, degrading the overall final reconstruction. This happens possibly because the random variables Z_c^L and Z_c^R are not Gaussian. A heuristic solution to that problem is to use sigmoid functions to calculate the weights: $\omega_L = (1 + \exp\{-\beta \Delta V\})^{-1}$, and $\omega_R = (1 + \exp\{\beta \Delta V\})^{-1}$ with $\Delta V = (V_L^2 - V_R^2)/(V_L^2 + V_R^2) = (\alpha^2 - 1)/(\alpha^2 + 1)$. The positive coefficient β controls the amount of diffusion between the left and the right regions, and should be determined experimentally. In the limit, as β tends to infinity, merging reduces to a hard decision: $Z_c = Z_c^L$ if $V_L > V_R$, and $Z_c = Z_c^R$ otherwise. Our merging technique presents two advantages: (a) obtaining more coverage of the scene and (b) reducing the estimation noise. Moreover, since we do not move the camera between scans, we do not have to solve for the difficult problem of view alignment [11, 7, 5]. One merging example is presented in experiment 3.

Independently from local variations in accuracy within one scan, one would also wish to maximize the global (or average) accuracy of reconstruction throughout the entire scene. In this paper, scanning is vertical (shadow parallel to the y axis of the image). Therefore, the average relative depth error $|\sigma_{Z_c}/Z_c|$ is inversely proportional to $|\cos\xi|$ (see [3]). The two best values for the azimuth angle are then $\xi = 0$ and $\xi = \pi$ corresponding to the lamp standing either to the right ($\xi = 0$) or to the left ($\xi = \pi$) of the camera (see figure 5-top).

4 Experimental Results

4.1 Calibration accuracies

Camera calibration. For a given setup, we acquired 10 images of the checkerboard (see figure 1), and performed independent calibrations on them. The checkerboard consisted of approximately 90 visible corners on a 8×9 grid. Then, we computed both mean values and standard deviations of all the parameters independently: the focal length f_c, radial distortion factor k_c and desk plane position Π_d. Regarding the desk plane position, it is convenient to look at the height d_d and the surface normal vector $\overline{n}_d$ of Π_d expressed in the camera reference frame. An additional geometrical quantity related to $\overline{n}_d$ is the tilt angle θ (see figure 5). The following table summarizes the calibration results (notice that the relative error on the angle θ is computed referring to 360 degrees):

Parameters	Estimates	Relative errors
f_c (pixels)	857.3 ± 1.3	0.2%
k_c	-0.199 ± 0.002	1%
d_d (cm)	16.69 ± 0.02	0.1%
$\overline{n}_d$	$\begin{pmatrix} -0.0427 \pm 0.0003 \\ 0.7515 \pm 0.0003 \\ 0.6594 \pm 0.0004 \end{pmatrix}$	0.06%
θ (degrees)	41.27 ± 0.02	0.006%

Lamp calibration. Similarly, we collected 10 images of the pencil shadow (like figure 3-top-right) and performed calibration of the light source on them. See section 2.1. Notice that the points $\overline{b}$ and $\underline{t_s}$ were manually extracted from the images. Define $\overline{S}_c$ as the coordinate vector of the light source in the camera frame. The following table summarizes the calibration results (refer to figure 5 for notation):

Parameters	Estimates	Relative errors
$\overline{S}_c$ (cm)	$\begin{pmatrix} -13.7 \pm 0.1 \\ -17.2 \pm 0.3 \\ -2.9 \pm 0.1 \end{pmatrix}$	$\approx 2\%$
h_S (cm)	34.04 ± 0.15	0.5%
ξ (degrees)	146.0 ± 0.8	0.2%
ϕ (degrees)	64.6 ± 0.2	0.06%

The estimated lamp height agrees with the manual measure (with a ruler) of 34 ± 0.5 cm.

Our method yields an accuracy of approximately 3 mm (in standard deviation) in localizing the light source. This accuracy is sufficient for final shape recovery without significant deformation, as we discuss in the next section.

4.2 Scene reconstructions

On the first scene (figure 6), we evaluated the accuracy of reconstruction based on the sizes and shapes of the plane at the bottom left corner and the corner object on the top of the scene (see figure 4-top).

Planarity of the plane: We fit a plane across the points lying on the planar patch and estimated the standard deviation of the set of residual distances of the points to the plane to 0.23 mm. This corresponds to the granularity (or roughness) noise on the planar surface. The fit was done over a surface patch of approximate size 4 cm × 6 cm. This leads to a relative non planarity of approximately 0.23mm/5cm = 0.4%. To check for possible global deformations due to errors in calibration, we also fit a quadratic patch across those points. We noticed a decrease of approximately 6% in residual standard deviation after quadratic warping. This leads us to believe that global geometric deformations are negligible compared to local surface noise. In other words, one may assume that the errors of calibration do not induce significant global deformations on the final reconstruction.

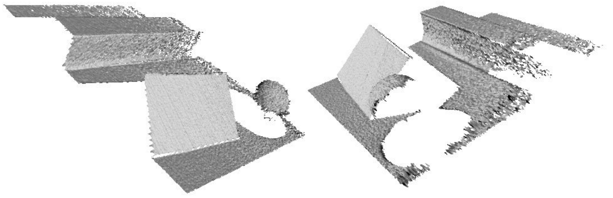

Figure 6: **Experiment 1 - The plane/ball/corner scene:** Two views of the mesh generated from the cloud of points obtained after triangulation. The original sequence was 270 frames long, the images being 320 × 240 pixels each. At 60 Hz acquisition frequency, the entire scanning take 5 seconds. The camera was positioned at distance $d_d = 16.7$ cm from the desk plane, tilted down by $\theta = 41.3$ degrees. The light source was at height $h_S = 37.7$ cm, on the left of the camera at angles $\xi = 157.1$ degrees and $\phi = 64.8$ degrees. From the right-hand figure we notice that the right-hand side of the reconstructed scene is more noisy than the left-hand side. This was expected since the lamp was standing on the left of the camera (refer to section 3 for details).

Geometry of the corner: We fit 2 planes to the corner structure, one corresponding to the top surface (the horizontal plane) and the other one to the frontal surface (vertical plane). We estimated the surface noise of the top surface to 0.125 mm, and that of the frontal face to 0.8 mm (almost 7 times larger). This noise difference between the two planes can be observed on figure 6. Once again, after fitting quadratic patches to the two planar portions, we did not notice any significant global geometric distortion in the scene (from planar to quadratic warping, the residual noise decreased by only 5% in standard deviation). From the reconstruction, we estimated the height H and width D of the right angle structure, as well as the angle ψ between the two reconstructed planes, and compared them to their true values:

Parameters	Estimates	True values	Relative errors
H (cm)	2.57 ± 0.02	2.65 ± 0.02	3%
D (cm)	3.06 ± 0.02	3.02 ± 0.02	1.3%
ψ (degrees)	86.21	90	1%

The overall reconstructed structure does not have any major noticeable global deformation (it seems that the calibration process gives good enough estimates). The most noticeable source of errors is the surface noise due to local image processing. A figure of merit to keep in mind is a surface noise between 0.1 mm (for planes roughly parallel to the desk) and 0.8 mm (for frontal plane in the right corner). In most portions of the scene, the errors are of the order of 0.3 mm, i.e. less than 1%. Notice that these figures may very well vary from experiment to experiment, especially depending on how fast the scanning is performed. In all the presented experiments, we kept the speed of the shadow approximately uniform.

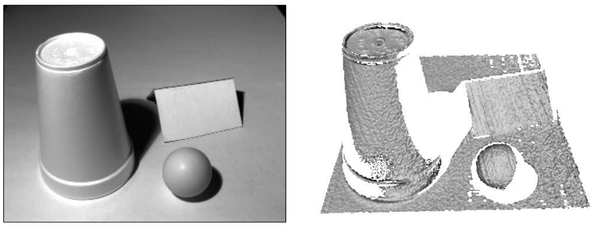

Figure 7: **Experiment 2 - The cup/plane/ball scene:** The scanned objects were a cup, the plane and the ball. The initial image of the scene is shown on the left, and the final reconstructed mesh on the right. We found agreement between the estimated height of the cup from the 3D reconstruction, 11.04 ± 0.09 cm, and the measured height (obtained using a ruler), 10.95 ± 0.05 cm. Once again the right portion on the reconstructed scene is noisier than the left portion. This was expected since the light source was, once again, standing to the left of the camera. Geometrical parameters: $d_d = 22.6$ cm, $\theta = 38.2$ degrees, $h_S = 43.2$ cm, $\xi = 155.9$ degrees, and $\phi = 69$ degrees.

Figures 7 and 8 report the reconstruction results achieved on two other scenes.

5 Conclusion and future work

We have presented a simple, low cost system for extracting surface shape of objects. The method requires very little processing and image storage so that it can be implemented in real time. The accuracies we obtained on the final reconstructions are reasonable (at most 1% or 0.5 mm noise error) considering the little hardware requirement. In addition, the final outcome is a dense coverage of the surface (one point in space for each pixel in the image) allowing for direct texture mapping.

An error analysis was presented together with the description of a simple technique for merging multiple 3D scans together in order to (a) obtain a better coverage of the scene, and (b) reduce the estimation noise. The overall calibration procedure, even in the case of multiple scans, is very intuitive, simple, and sufficiently accurate.

Another advantage of our approach is that it easily scales to larger scenarios indoors – using more powerful lamps like photo-floods – and outdoors where the sun may be used as a calibrated light source (given latitude, longitude, and time of day). These are experiments that we wish to carry out in the future.

Other extensions of this work relate to multiple view integration. We wish to extend the alignment technique to a method allowing the user to move freely the object in front of the camera and the lamp between scans in order to achieve a full coverage. That is necessary to construct complete 3D models.

It is also part of future work to incorporate a geometrical model of extended light source to the shadow edge detection process, in addition to developing an uncalibrated (or projective) version of the method.

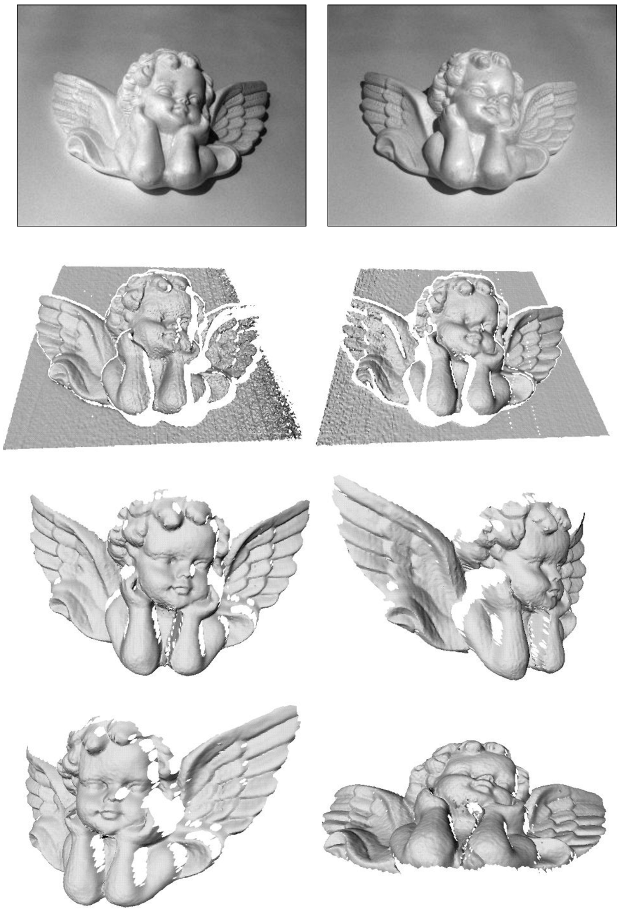

Figure 8: **Experiment 3 - The angel scene:** We took two scans of the angel with the lamp first on the left side (top-left) and then on the right side (top-right) of the camera. The two resulting meshes are shown on the second row, left and right. As expected, the portions further away from the light source are noisier. The two meshes were then merged together following the technique described in section 3, with diffusion coefficient $\beta = 15$. Four different views of the final mesh (47076 triangles) are presented. Notice the small surface noise: we estimated it to 0.09 mm throughout the entire reconstructed surface. Over a depth variation of approximately 10 cm, this means a relative error of 0.1%. The few white holes correspond to the occluded portions of the scene (not observed from the camera or not illuminated). Most of the geometrical constants in the setup were kept roughly identical in both scans: $d_d = 22$ cm, $\theta = 40$ degrees, $h_S = 62$ cm, $\phi \approx 70$ degrees; we only changed the azimuth angle ξ from π (lamp on the left) to 0 (lamp on the right). In this experiment we took the lamp reflector off, leaving the bulb naked. Consequently, we noticed a significant improvement in the sharpness of the projected shadow compared to the two first experiments. We believe that this operation was the main reason for the noticeable improvement in reconstruction quality. Once again, there was no significant global deformation in the final structured surface: we fit a quadratic model through the reconstructed set of points on the desk plane and noticed from planar to quadratic warping a decrease of only 2% on the standard deviation of surface noise.

Acknowledgments

This work is supported in part by the California Institute of Technology; an NSF National Young Investigator Award to P.P.; the Center for Neuromorphic Systems Engineering funded by the National Science Foundation at the California Institute of Technology; and by the California Trade and Commerce Agency, Office of Strategic Technology. We wish to thank all the colleagues that helped us throughout this work, especially Luis Goncalves, George Barbastathis, Mario Munich, and Arrigo Benedetti for very useful discussions. Comments from the anonymous reviewers were very helpful in improving a previous version of the paper.

References

[1] Paul Besl, *Advances in Machine Vision*, chapter 1 - Active optical range imaging sensors, pages 1–63, Springer-Verlag, 1989.

[2] P.J. Besl and N.D. McKay, “A method for registration of 3-d shapes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[3] Jean-Yves Bouguet and Pietro Perona, “3D photography on your desk”, Technical report, California Institute of Technology, 1997, available at: http://www.vision.caltech.edu/bouguetj/ICCV98.

[4] Brian Curless and Marc Levoy, “Better optical triangulation through spacetime analysis”, *Proc. 5^{th} Int. Conf. Computer Vision*, pages 987–993, 1995.

[5] Brian Curless and Marc Levoy, “A volumetric method for building complex models from range images”, *SIGGRAPH96, Computer Graphics Proceedings*, 1996.

[6] O.D. Faugeras, *Three dimensional vision, a geometric viewpoint*, MIT Press, 1993.

[7] Berthold K.P. Horn, “Closed-form solution of absolute orientation using unit quaternions”, *J. Opt. Soc. Am. A*, 4(4):629–642, 1987.

[8] Jurgen R. Meyer-Arendt, “Radiometry and photometry: Units and conversion factors”, *Applied Optics*, 7(10):2081–2084, October 1968.

[9] Shree K. Nayar, Katsushi Ikeuchi, and Takeo Kanade, “Shape from interreflections”, *Int. J. of Computer Vision*, 6(3):173–195, 1991.

[10] Athanasios Papoulis, *Probability, Random Variables and Stochastic Processes*, Mac Graw Hill, 1991, Third Edition.

[11] A.J. Stoddart and A. Hilton, “Registration of multiple point sets”, *Proceedings of the 13th Int. Conf. of Pattern Recognition*, 1996.

[12] Marjan Trobina, “Error model of a coded-light range sensor”, Technical Report BIWI-TR-164, ETH-Zentrum, 1995.

[13] R. Y. Tsai, “A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses”, *IEEE J. Robotics Automat.*, RA-3(4):323–344, 1987.

[14] John W. T. Walsh, *Photometry*, Dover, NY, 1965.

[15] Y.F. Wang, “Characterizing three-dimensional surface structures from visual images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(1):52–60, 1991.

[16] Z. Yang and Y.F. Wang, “Error analysis of 3D shape construction from structured lighting”, *Pattern Recognition*, 29(2):189–206, 1996.