Department of Computer Science

Technical Report

Combining Shape from Shading and Stereo Using Human Vision Model

James Edwin Cryer, Ping-Sing Tsai and Mubarak Shah

CS-TR-92-25

University of Central Florida

Orlando, FL 32816

Combining Shape from Shading and Stereo Using Human Vision Model

James Edwin Cryer, Ping-Sing Tsai and Mubarak Shah

Computer Science Department

University of Central Florida

Orlando, FL 32816

Abstract

Stereo algorithms suffer from the lack of local surface texture due to smoothness of depth constraint, or local miss-matches in disparity estimates. Thus, the stereo methods only provide a coarse depth map which can be associated with a low pass image of the depth map. On the other hand, shape from shading algorithms produce better estimates of local surface areas, but some of them have problems with variable albedo and spherical surfaces. Thus, shape from shading methods produce better detailed depth information, and can be associated with the high pass image of the depth map image. In order to compute a better depth map, we present a method for integrating the high frequency information from the shape from shading and the low frequency information from stereo. Our method is motivated by the human vision system, and follows Hall and Hall's model. The proposed algorithm is very simple, takes about .7 seconds for a 128×128 *image on a Sun SparcStation-1, is non-iterative, and does not use any thresholds. The results obtained with a variety of synthetic and real images are discussed. The quality of depth obtained by integrating shading and stereo is compared with the ground truth (range image) using average surface gradient error measure, and improvement ranging from 30% to 50% over stereo, and from 65% to 98% over shading is demonstrated.*

Key Words Shape from Shading, Shape from stereo, Integration of shape from X modules, Human Visual System.

Contents

1 Introduction

Modern Computer Vision research follows the Marr paradigm [13] that treats vision as a large, complex information processing system. Individual perceptual modules can be identified in the system which are responsible for the computation of shape from shading, stereo, motion, texture and contour, as well as processes for determining the location and nature of illumination sources, three dimensional motion, and others. During the last two decades there has been significant interest in these individual modules, which are termed *shape from X*. Interesting results have been reported, in particular, in motion, stereo and shading. Marr envisioned that the output from the individual modules will ultimately be integrated into a single representation called *2.5 D sketch*. However, this integration was never accomplished by Marr. Vision inherently is an ill-posed problem and the solutions for *shape from X* obtained by considering each module individually may not necessarily *exist*, may not be *unique* and may not be *stable*. Therefore, in order to tackle these problems we need more information. In particular, if we combine information from different image cues like stereo, shading and motion, the solution may be significantly improved. Surprisingly, it is only recently that researchers in Computer Vision have started realizing the benefits of integrating information from separate modules. These are the work of Horn [10] on combining shading with contour, Grimson's [6] use of shading in determining the surface orientation of feature-point contours obtained from stereo, Aloimonos's methods [1] for combining shading and motion, texture and motion, and motion and contour, and Waxman's approach for combining stereo and motion [17].

The objective of this research is to work on the integration of *shape from X* modules. In particular, we are interested in combining the depth information (3-D shape) from two very important cues, stereo and shading. Shape from shading is the estimation of 3-D shape from a 2-D image given the light source and surface reflectance information. Stereo is the estimation of 3-D shape from two images taken by cameras which are slightly shifted along the x axis. The image of the object in the left stereo image is shifted with respect to the image in the right stereo image. This shift, which is also called the disparity, is inversely proportional to the 3-D distance (depth) from the camera to the object. The problem with the stereo is that it produces a coarse depth map. On the other hand, the shape from shading methods perform better on the small details of the surface. It is important to come up with a method that works well for both the coarse areas and detailed areas in the depth map. In the frequency domain the low spatial frequencies are related to the coarse shape information, and the high spatial frequencies are related to the details of the shape. So our criterion for combining shape from shading and stereo is very simple. We want to keep the low frequency information from the stereo, and add it with the high frequency information from the shape from shading results. Our solution is motivated by the human vision system, and follows Hall and Hall's model [7].

There are many applications for the use of depth maps. For instance, an unmanned spacecraft attempting to land on a lunar surface needs to estimate depth in details such as small rocks and holes, as well as coarse depth such as large mountains or craters, in order to land safely.

The organization of the rest of the paper is as follows. The next section deals with the problems in the shape from stereo and shading. First, we briefly describe stereo and shading, and summerize Barnard's stereo algorithm and Pentland's shading algorithm which will be used in this work. Then, we present two examples to highlight the problems in shading and stereo. Section three is devoted to the human visual system. We report the hypothesis about the presence of the spatial channels tuned in orientation and spatial frequency in the human visual system. Next we summerize Hall and Hall's model, which will be used for integration of stereo and shading. Section four is the main thrust of the paper, where we describe our method for integration of stereo and shading. Section five deals with the related works, and the comparison of proposed method with the previous methods. Finally, the results for synthetic and real images are presented in section six.

2 Shape from X

The human visual system responds to light reflecting from objects. The visual system is able to determine depth from monocular scenes such as 2D images from smooth changes along the surface of the object, from former knowledge of size relationships of objects in the image, from occlusion, from size and from texture gradients. In this section we will focus on the shape from shading and stereo.

2.1 Shape from Shading

Shape from shading deals with the recovery of 3D shape from a *single* monocular image which is only one cue that humans use to determine shape. There are two main classes of algorithms for computing shape from a shaded image: global methods and local methods. The global methods, in general, are very complex and slow. Sometimes they (e.g., variational calculus methods) require more than thousands of iterations to converge. A representative method of global approaches is used by Horn [9]. The local methods [15, 12], on the other hand, are simple, fast and give accurate local details within each homogeneous area, but are not accurate enough globally.

In shape from shading algorithms it is assumed that the surface reflectance map is given, or its form is known. However, surfaces of most objects in the real world have mixed reflectance forms, such as Lambertian with Specular, or Lambertian with various albedo values. We will not be able to recover the accurate 3D shape information with the shape from shading method alone.

Pentland [15] proposed a local algorithm based on the linearity of the reflectance map in the surface gradient (p, q), which greatly simplifies the shape from shading problem, and is very suitable for our purpose. The reflectance function for the Lambertian surfaces is modeled as follows:

$$\begin{aligned} E(x,y) &= R(p,q) && (1)\\ &= \frac{1 + pp_s + qq_s}{\sqrt{1+p^2+q^2}\sqrt{1+p_s^2+q_s^2}} && (2)\\ &= \frac{\cos\sigma + p\cos\tau\sin\sigma + q\sin\tau\sin\sigma}{\sqrt{1+p^2+q^2}} && (3) \end{aligned}$$

where $E(x,y)$ is the gray level at pixel (x,y), Z is the depth map, $p = \frac{\partial Z}{\partial x}$, $q = \frac{\partial Z}{\partial y}$, $p_s = \frac{\cos\tau\sin\sigma}{\cos\sigma}$, $q_s = \frac{\sin\tau\sin\sigma}{\cos\sigma}$, τ is the tilt of the illuminant and σ is the slant of the illuminant. By taking the Taylor series expansion of the reflectance function, equation (1), about $p = p_0$, $q = q_0$, up through the first order terms, we have

$$E(x,y) = R(p_0,q_0) + (p - p_0)\frac{\partial R}{\partial p}(p_0,q_0) + (q - q_0)\frac{\partial R}{\partial q}(p_0,q_0). \qquad (4)$$

For Lambertian reflectance (equation (3)), the above equation at $p_0 = q_0 = 0$, reduces to

$$E(x,y) = \cos\sigma + p\cos\tau\sin\sigma + q\sin\tau\sin\sigma.$$

Next, Pentland takes the Fourier transform of both sides of this equation. Since the first term on the right is a DC term, it can be dropped. Using the identities:

$$\frac{\partial}{\partial x} Z(x,y) \longleftrightarrow F_Z(\omega_1,\omega_2)(-i\omega_1) \tag{5}$$

$$\frac{\partial}{\partial y} Z(x,y) \longleftrightarrow F_Z(\omega_1,\omega_2)(-i\omega_2), \tag{6}$$

where F_Z is the Fourier transform of $Z(x,y)$, we get,

$$F_E = F_Z(\omega_1,\omega_2)(-i\omega_1)\cos\tau\sin\sigma + F_Z(\omega_1,\omega_2)(-i\omega_2)\sin\tau\sin\sigma,$$

where F_E is the Fourier transform of the image $E(x,y)$. The depth map $Z(x,y)$ can be computed by rearranging the terms in the above equation, and then taking the inverse Fourier transform. Other shape from shading algorithms use more complex models such as interreflections [14], changing albedos and specular reflectance [8]. These methods are more difficult and require more computational time to solve.

2.2 Shape from Stereo

Stereo vision is present in the human visual system and does not depend on the complex cues like occlusion, shadows, texture gradients and size of objects used by a monocular visual system. Shape from stereo uses two images from which the depth map can be calculated. There are three main classes of algorithms for computing shape from stereo pairs: feature-based approaches, area-based approaches, and miscellaneous approaches. In the feature-based methods, the depth is computed at feature locations (mostly edges); In this case it is necessary to perform interpolation between features in order to get the dense depth maps. In the area-based approaches depth is computed for each pixel. The area-based methods have difficulties with the areas of nearly homogeneous image intensity which lack spatial structure. There are several miscellaneous approaches, such as Barnard's stochastic stereo algorithm [2], which can run very fast on a suitable parallel machine, but the computed depth map suffers from the lack of local surface texture. In general, most stereo methods only provide coarse depth maps, and the fine details are missing in these depth maps.

In Barnard's stereo approach the problem is to find an assignment of disparities, $D(i,j)$, such that two criteria, *similar intensity* and *smoothness*, are satisfied:

$$E = \sum_{j=1}^{n}\sum_{i=1}^{n} \|I_L(i,j) - I_R(i,j) + D(i,j))\| + \lambda\| \bigtriangledown D(i,j)\|, \tag{7}$$

where I_L and I_R are the left and right images, $D(i,j)$ is the disparity map, the $\bigtriangledown$ operator computes the sum of the absolute differences between disparity $D(i,j)$ and its eight neighbors, and λ is a constant. For a 128 × 128 image, and a disparity range of 10 pixels, there are 10^{16384} possible disparity assignments, which results in combinatorial explosion. Barnard uses a simulated annealing to solve this problem. The algorithm is as follows:

```
1. Select a random state S.
2. Select high temperature T.
3. While T > 0.
   (a) Select $S'$
           $\Delta E \leftarrow E(S') - E(S)$.
   (b) if $\Delta E \leq 0$ then $S \leftarrow S'$
   (c) else $P \leftarrow \exp^{\frac{-\Delta E}{T}}$, $X \leftarrow rand(0,1)$,
                                   if $X < P$ then $S \leftarrow S'$
   (d) if no decrease in $E$ for several iterations then lower T.
```

2.3 Problems

Some problems in shading and stereo can be demonstrated by the following examples. In the first example, we generate a stereo pair with a planar object (as shown in Figure 1.(a)-(b)). Since there is no texture on the object, and the surface has homogeneous intensity, the stereo matching algorithm computes very poor results (as shown in Figure 1.(e)), only some matches on the left and right edges are found. However, since the image is of constant albedo, the shape from shading algorithm produces a reasonable good result corresponding to the homogeneous areas on the object (as shown in Figure 1.(f)). In the next example, we use the random dot stereo pair, (as shown in Figure 1.(c)-(d)). Now, since the image contains texture the stereo computes reasonable depth for the object areas (as shown in Figure 1.(g)). The random dots on the object can be considered due to variable albedo. Since the shape

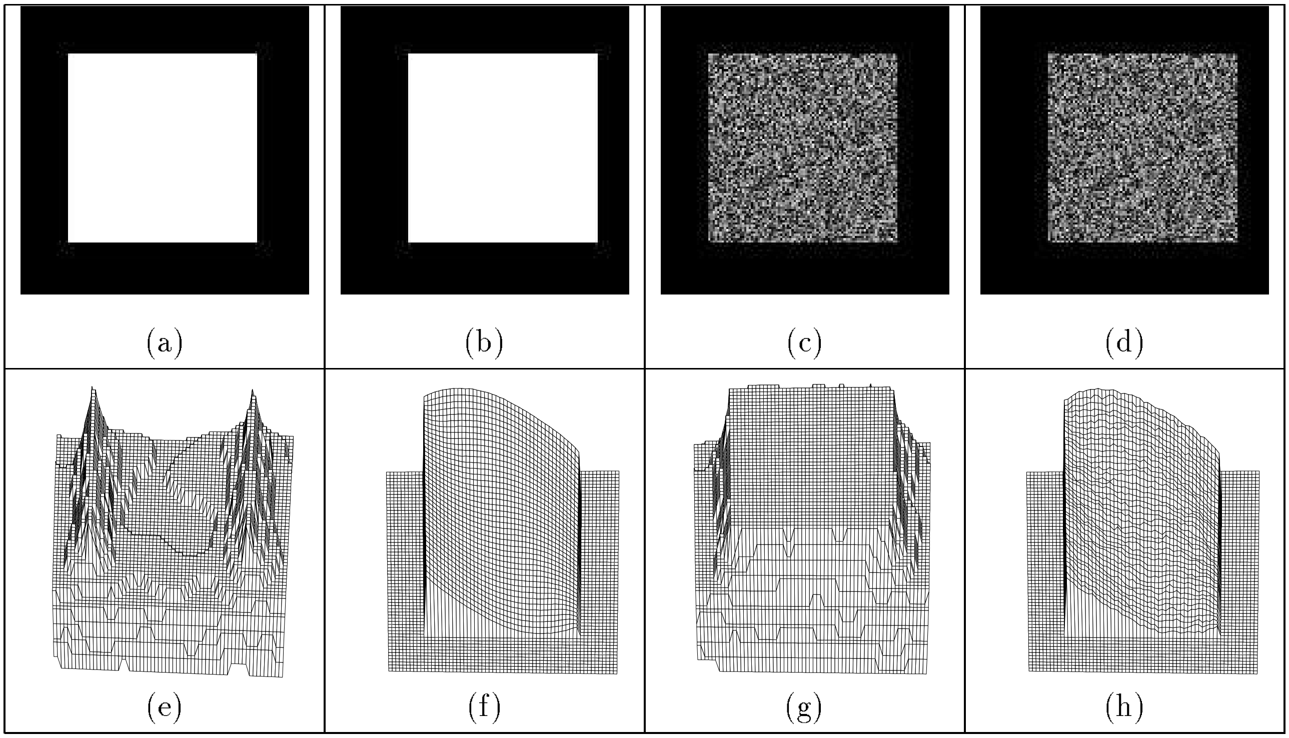


Figure 1: Examples for shape from shading and stereo. (a) and (b) are the left and right noise-free stereo images. (c) and (d) are the left and right stereo images with no texture. (e) A 3D plot of the estimated depth map by stereo matching algorithm for the stereo images (a) and (b). (f) A 3D plot of the estimated depth map by shape from shading algorithm for image (a). (g) A 3D plot of the estimated depth map by stereo matching algorithm for the stereo images (c) and (d). (f) A 3D plot of the estimated depth map by shape from shading algorithm for image (c).

from shading algorithm assumes constant albedo, the results for this example, as shown in Figure 1.(h), are very poor; we can see a bumpy surface on the object.

The two examples discussed above are extreme cases. For the first case, we do not need the stereo at all because the shape from shading can provide enough depth information. For the second case, we only need the stereo, since the shape from shading cannot provide any useful information. In more realistic situations, we need to combine the information from stereo and shading in order to get better results.

3 Human Visual System and Hall and Hall's model

3.1 Human Visual System

$$\boxed{\text{Optical System}} \longrightarrow \boxed{\text{Retina}} \longrightarrow \boxed{\text{Visual Pathway}}$$

Figure 2: Model of the human visual system.

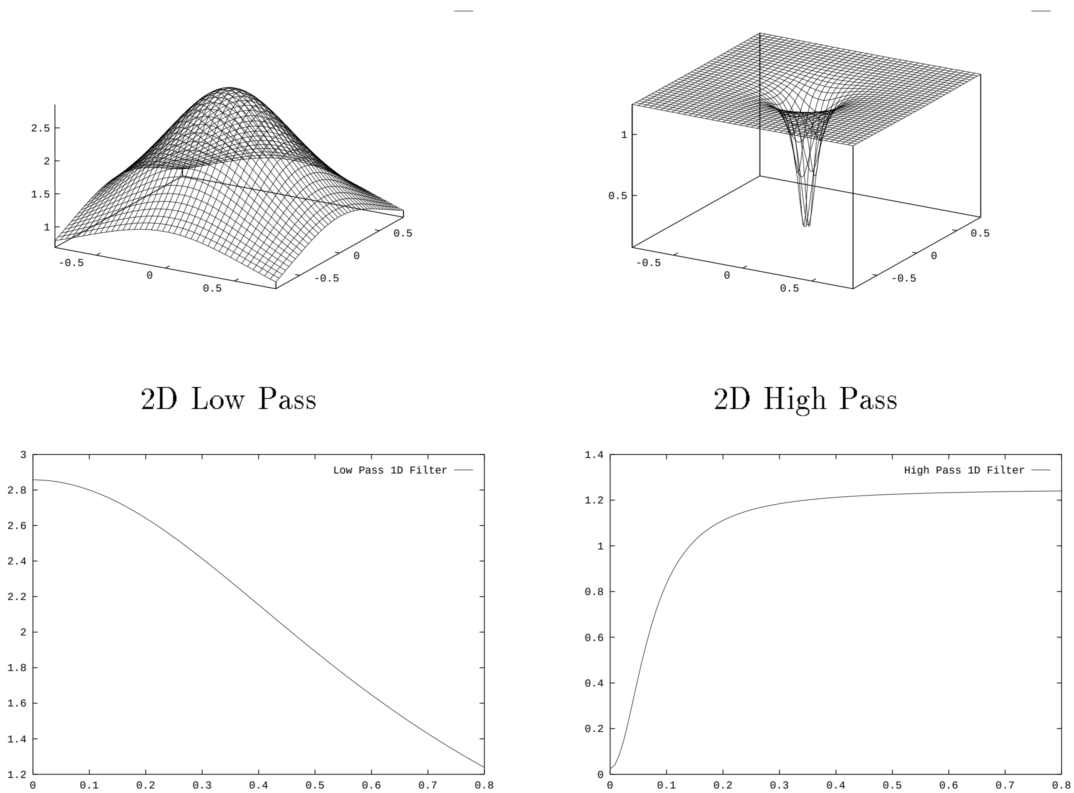
There is experimental data that supports the hypothesis that the Human Visual System (HVS) is composed of spatial frequency channels tuned in orientation and spatial frequency. The Human vision model is shown in Figure 2. The first box, optical system, is considered as a low pass filter. This is due to high frequency information being lost by the optical system. The amount of loss depends on the quality of the lens. The second box, retina, performs a log operation on the light entering the lens. This produces an upper bound on the intensities observed. Contrast consistency is the visual effect produced at edges where the step change depends on the differences in intensities. This step change represents contrast and visually does not change in proportion to the step sizes. This is due to the log operators in the visual system at the rods and cones, which compress the visual responses at edges or changes in intensities. The third box, visual pathway, is a high pass filter operation produced by the visual nerve. The photoreceptors of the human eye respond to light intensity and are modeled as neurons with inhibitory and excitatory responses to light intensity. This dual response accounts for the high pass filtering stage in the HVS and explains the Mach band effect.

3.2 Hall and Hall's model

$$\boxed{\text{Low Pass Filter}} \longrightarrow \boxed{\text{Log}} \longrightarrow \boxed{\text{High Pass Filter}}$$

Figure 3: Hall and Hall's model for human visual system.

Hall and Hall [7] describe a model to simulate the visual properties of the human eye by combining the low and high frequency information (Figure 3). The first box is a low pass filter which is derived from a lens of diameter 3 mm. The second box is the log operation



2D Low Pass

2D High Pass

1D Low Pass

1D High Pass

Figure 4: Hall and Hall's low and high pass filters.

performed by the retina, and the third box is the high pass filter derived from the neuron model. The low and high pass filters (as shown in Figure 4) in Hall and Hall's model of the HVS which work in the frequency domain are shown as follows:

$$Low(\omega) = \frac{2\alpha}{\alpha^2 + \omega^2} \quad and \tag{8}$$

$$High(\omega) = \frac{a^2 + \omega^2}{2a_0 + (1 - a_0)(a^2 + \omega^2)}, \tag{9}$$

where $\omega = \sqrt{u^2 + v^2}$, u and v represent the two dimensional frequencies in the Fourier domain. The term α is the spatial angular frequency. A typical value of α is 0.7 for a 3mm diameter of the iris opening. The term a_0 represents the distance factor, which is the amount of change between the low and high frequencies. The other term a represents the strength factor, which is the rate of the cutoff point change between the low and high frequencies. For the human visual system the normal values for a_0 and a are respectively 0.2 and 0.01.

3.3 Discussion

Consider an image which has good low frequency information, but bad high frequency information. Assume that the image has been passed through three stages in Hall and Hall's model, as shown in Figure 3. Since the high pass filter has amplified the high frequency information, which is not good, we can undo that by passing the image through an inverse high pass filter. Similarly, for an image with good high frequency and bad low frequency contents, which has passed through the three stages in Hall and Hall's model, we will recover the good information by passing the image through the inverse high pass filter and subtracting the result from the original image.

4 Integration of Shape from Shading and Stereo

The stereo methods provide the coarse shape information, and the shape from shading methods provide the detailed feature information. In the frequency domain the low spatial frequencies are related to the coarse shape information, and the high spatial frequencies are related to the details of the shape. Therefore, our criterion for combining shape from shading and stereo is very simple: *Keep the low frequency information from stereo, and add with the high frequency information from the shape from shading.*

Intuitively, one can use a high pass filter to separate the high frequency information from the shading result, and a low pass filter to separate the low frequency information from the stereo result. Then one can combine the low frequency information from stereo matching with the high frequency information from shading. However, choosing a good filter and the proper cut-off points is not an easy job.

We will use Hall and Hall's model for human vision to combine the stereo and shading depth maps. The high pass filter designed by Hall and Hall attenuates the low frequency information, and emphasizes the high frequency information. Therefore, by inverting Hall and Hall's high pass filter (it will serve as a low pass filter now) we can get the low frequency information from the image of an estimated depth map. This process is done to the images of the estimated depth map from both stereo matching and shading algorithms. Next we compute the high frequency information of the shading results by subtracting the low frequency information from the Fourier Transform of the original depth map. Finally, the low frequency information from stereo and the high frequency information from shading are combined in the frequency domain to produce a better depth map. The flowchart of our method is shown in Figure 5.

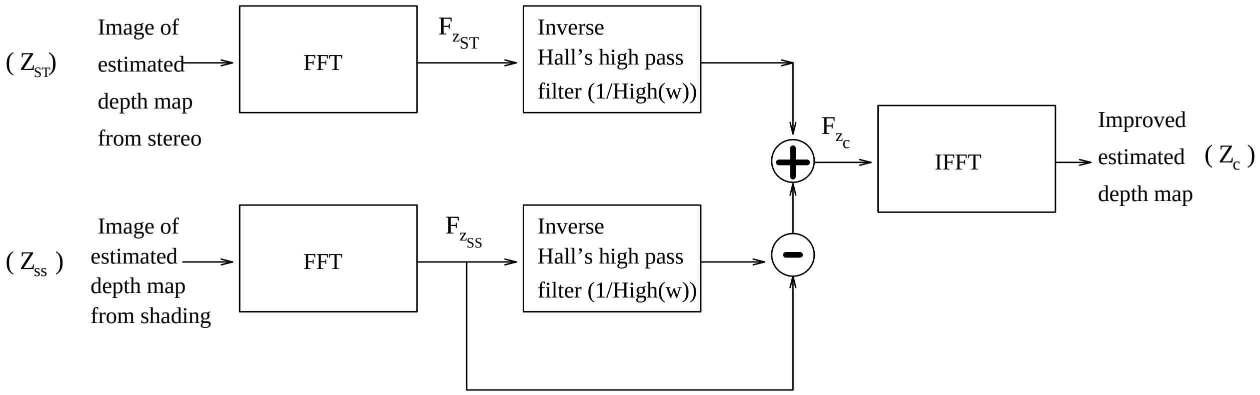

(Z_ST)
Image of estimated depth map from stereo
FFT
F_ZST
Inverse Hall's high pass filter (1/High(w))
F_Zc
IFFT
Improved estimated depth map
(Z_c)
(Z_ss)
Image of estimated depth map from shading
FFT
F_Zss
Inverse Hall's high pass filter (1/High(w))

Figure 5: Flowchart of proposed method for combining the stereo and shading.

Mathematically, the proposed method can be summarized as follows:

$$F_{Z_C}(\omega) = F_{Z_{ST}} \times High(\omega)^{-1} + (F_{Z_{SS}} - F_{Z_{SS}} \times High(\omega)^{-1}),$$

where $High(\omega)^{-1}$ is the inverse high pass filter, and F_{Z_C}, $F_{Z_{ST}}$ and $F_{Z_{SS}}$ respectively are the Fourier Transforms of the combined depth map, stereo depth map and shading depth map. The combined depth, Z_C, is computed by taking the inverse Fourier Transform of the above equation.

We want to emphasize here that it is important to apply the *inverse* Hall and Hall's high pass filter as discussed above. We have also experimented with some other filters, but have not obtained significant improvement in the depth map. For example, we applied the high pass Butterworth filter [5] to the shading depth map and the low pass Butterworth filter to the stereo depth map, and combined these filtered depth maps. However, the resultant depth map was not much better than stereo or shading alone.

The proposed method is very simple, and computationally inexpensive. Once the stereo and shading depth maps are available, the integration takes about 0.7 CPU seconds for a 128 × 128 image on a Sun SparcStation-1. The major computation occurs with the Fast Fourier Transform (FFT), and Inverse Fourier Transform (IFFT), which is known to be of order $n \log n$. Since we are using Pentland's shading algorithm which employs FFT, the Fourier Transform of the shading depth map is already available. Additional computation only involves the FFT of the stereo depth map. The computation for integration of stereo and shading consists only of the application of the filter, which is of order n^2 for an $n \times n$ image. The proposed algorithm does not have any parameters or hidden thresholds. All the parameters of the filter are fixed. The results presented in this paper were obtained by running the same program on different sets of data, with no change in any parameters. This is a very important feature of our method. Because many algorithms for the low level vision are fragile, in many cases it is almost necessary to fine-tune the thresholds for each and every set of data.

5 Discussion

There are several other possibilities for integrating stereo and shading. The stereo depth map can be used to improve the shape from shading algorithm. For instance, in Ikeuchi-Horn's shape from shading algorithm [10] it is assumed that the depth at the occluding contours is available. Their method iteratively computes the surface orientation at the remaining locations by propogating the depth at the occluding contours. The contour depth map computed by the feature-based stereo can be used for this purpose. In fact, Blake, Zisserman and Knowles [3] have shown that if the boundary information (depth) is available at the occluding contours then shape from shading converges to a unique solution.

In a recent shape from shading algorithm reported by Leclerc and Bobik [11], which uses the conjugate gradient method for minimizing the cost function, the depth is iteratively refined. Leclerc and Bobik assume that a good initial guess for depth at each pixel is available. In their case, the dense depth map computed by the area-based (correlation-based) stereo method can be used to obtain a good initial estimate.

The shape from shading can also be used to improve the depth map computed by stereo. For instance, Grimson [6] uses shading in determining the surface orientation of feature-point contours obtained from stereo.

Frankot and Chellapa [4] presented an elegant approach for enforcing the integrability constraint in shape from shading. They compute the orthogonal projection onto a vector subspace spanning the set of integrable slopes. This projection maps closed convex sets into closed convex sets, and hence, is attractive as a constraint in iterative algorithms. The authors noted that the low frequency information is lost in the process of image formation and due to regularization penalty and periodic boundary conditions. They showed improvements of their shape from shading results by incorporating the low frequency information obtained from another source (like the Digital Terrain model (DTM)).

Our approach for integrating stereo and shading is very different from the previous approaches. We assume that each module is working independently and in parallel, and can therefore be treated as a black box. In fact, our method for combining shading and stereo is not dependent upon any particular method for shading or stereo. In principle, any method can be used. We have used Pentland's shading method and Barnard's stereo method, because working implementations of those two methods are available in our lab. Comparing our approach with the other approaches described above, we find that none of the previous approaches can be generalized to intergrate stereo and shading. They are hardwired for the particular algorithms they use.

6 Experiments

6.1 Results for synthetic data

The proposed method was first tested on two synthetic images, Tomato and Mozart. The stereo images were created using the following formulas:

$$x' = \frac{(x-b)f}{f-z} \qquad and \tag{10}$$

$$x'' = \frac{(x+b)f}{f-z}, \tag{11}$$

where x' is the position in the left stereo pair, x'' is the position in the right stereo pair, $2b$ is the distance of shift of the camera in the left or right directions from the original x position in the image, z is the depth value at the x position, and f is the focal length of the camera. This can be seen in Figure 6. We first generate a gray level image using the Lambertian model, equation (2), based on the range data and the given light source direction. Then, for each pixel (x, y), we compute the corresponding positions in the left and right stereo pairs, (x', y) and (x'', y), using the above formulas, and set the intensity value $I(x, y) = I_L(x', y) = I_R(x'', y)$.

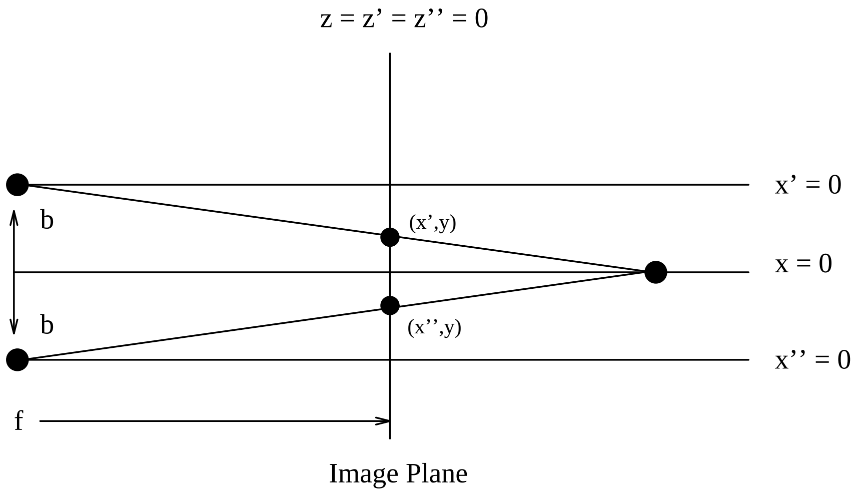


Figure 6: The stereo camera imaging system.

In order to evaluate the performance of our algorithm, we use the average surface gradient error proposed by Horn [9] as a error measure:

$$E = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} \sqrt{((p(i,j) - \hat{p}(i,j))^2 + (q(i,j) - \hat{q}(i,j))^2)}}{n^2}, \tag{12}$$

where $(p(i,j), q(i,j))$ is the actual surface normal (from the range image), and $(\hat{p}(i,j), \hat{q}(i,j))$ is the estimated surface normal (from the computed depth map), and n^2 is the number of pixels in the image.

The results for the Tomato images are shown in Figure 7. The gray level stereo pairs generated from the true depth map (shown in Figure 7.(a)) are shown in Figure 7.(b) and (c). The focal length and the distance between the two cameras were respectively assumed to be 400 and 60. The estimated depth map computed by Barnard's stereo matching algorithm is shown in Figure 7.(e). The average gradient square error is 0.46. This depth map is reasonably good. However, there are some noticeable errors in the depth map. For instance, the depth around outer portions of the tomato do not seem to be correct, and the surface patch around the upper left part is almost flat. The estimated depth map computed by Pentland's linear shape from shading algorithm (applied to the right stereo image) is shown in Figure 7.(f). Since the tomato is similar to a spherical object, and it is well known that the linear shape from shading method proposed by Pentland does not compute a good depth map for spherical surfaces [16], the average gradient square error is about 1.85. The result obtained by integrating stereo and shading using our method is shown in Figure 7.(d), the average gradient square error reduces to 0.24. This is approximately a 48% improvement over the stereo, and a 98% improvement over shading. We feel that achieving this great improvement by using a very simple algorithm is remarkable. Figure 7.(g)–(i) shows the reconstructed gray level images using the estimated depth maps in (d)–(f) with the light source direction $(0.01, 0.01, 1)$. The low frequency contents of stereo are shown in Figure 8.(a), and the high frequency contents of shading are shown in Figure 8.(b). The combination of (a) and (b) by our method results in a better depth map, which is shown in 8.(c).

Next, we tested our method for the Mozart image. The results are shown in Figure 9. The true depth map is shown in Figure 9.(a). The gray level stereo images generated from the true depth map are shown in Figure 9.(b) and (c). In this case also, the focal length and the distance between the two cameras were respectively assumed to be 400 and 60. The estimated depth map computed by Barnard's stereo matching algorithm is shown in Figure 9.(e). The average gradient square error is 0.77. In this case , it is also obvious that the stereo does a poor job on details. The surface is not that smooth, and the surface patches around the nose and eyes are not correct. The estimated depth map computed by Pentland's linear shape from shading algorithm (applied to the right stereo image) is shown in Figure 9.(f). The average gradient square error is 1.5. Pentland's method does a very poor job on this image. It is almost impossible to perceive a face from this depth map. For instance, the areas corresponding to the center of the face have incorrect dips in the surface. The results obtained by integrating stereo and shading using our method are shown in Figure 9.(d). The average gradient square error in this case reduces to 0.55. There is about a 30% improvement over stereo, and a 63% improvement over shading. This depth map is much closer to the original range image. The detailed surface patches around the nose and eyes are noticeable. Figure 9.(g)–(i) shows the reconstructed gray level images using the estimated depth maps in (d)–(f) with the light source direction $(0.01, 0.01, 1)$. It is very interesting to note that, even though the shading depth map in Figure 9.(f) appears to be very poor, the reconstructed gray level image in Figure 9.(i) looks much better than the reconstructed gray level image form the stereo depth map (shown in Figure 9.(h)). Some obvious problems around the nose (e.g., a line throughout the image) are noticeable in Figure 9.(i). The reconstructed gray level image (shown in Figure 9.(g)) is much closer to the gray level image (shown in Figure 9.(b)) generated using the true depth map. The low frequency contents of stereo are shown

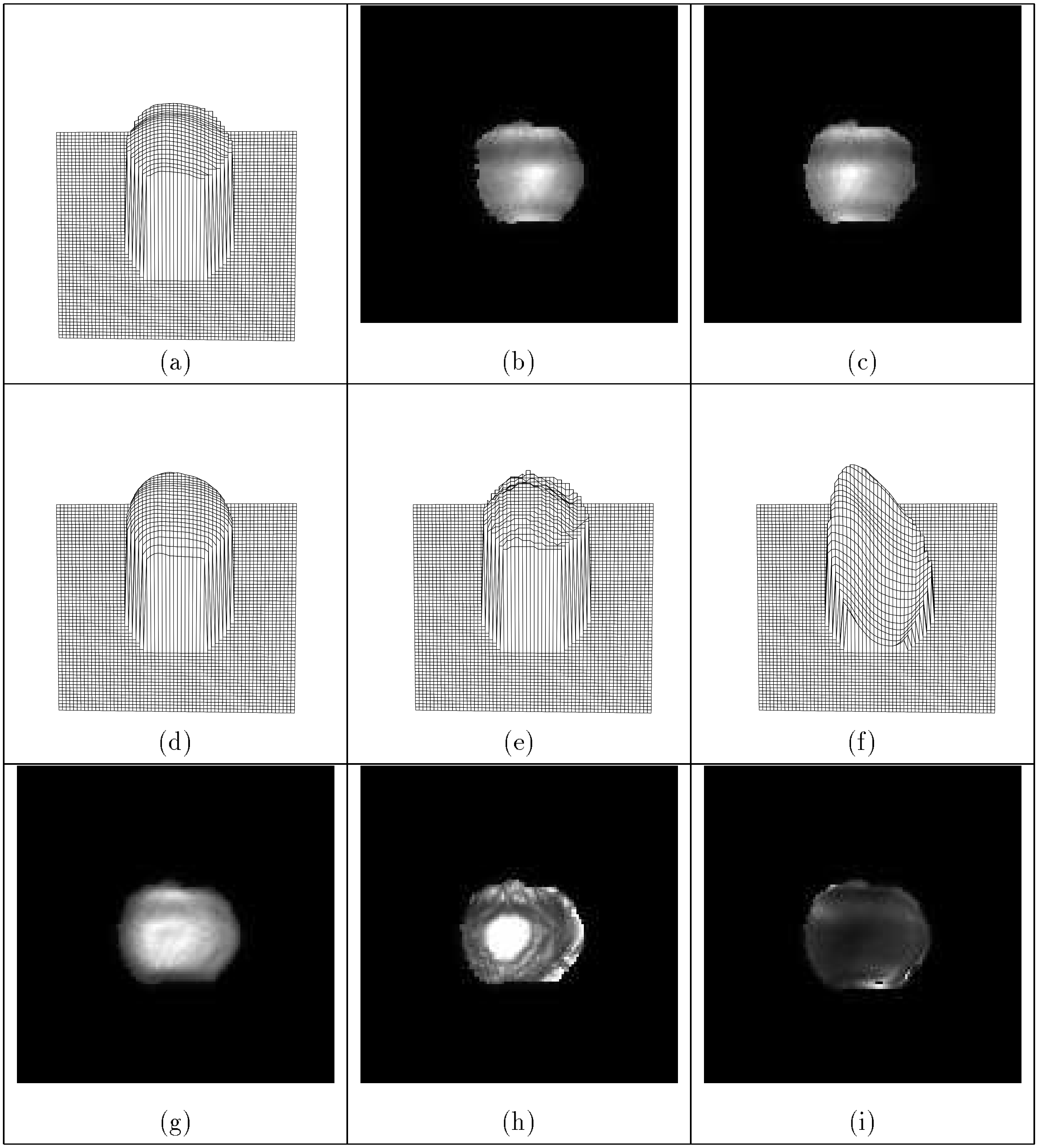


Figure 7: Results for the Tomato images. (a) A 3D plot of the range data. (b) Left stereo gray level image. (c) Right stereo gray level image. (d) A 3D plot of the estimated depth map by our method. (e) A 3D plot of the estimated depth map by stereo matching algorithm. (f) A 3d plot of the estimated depth map by Pentland's shape from shading algorithm. (g) A reconstructed gray level image using the estimated depth map in (d). (h) A reconstructed gray level image using the estimated depth map in (e). (i) A reconstructed gray level image using the estimated depth map in (f).

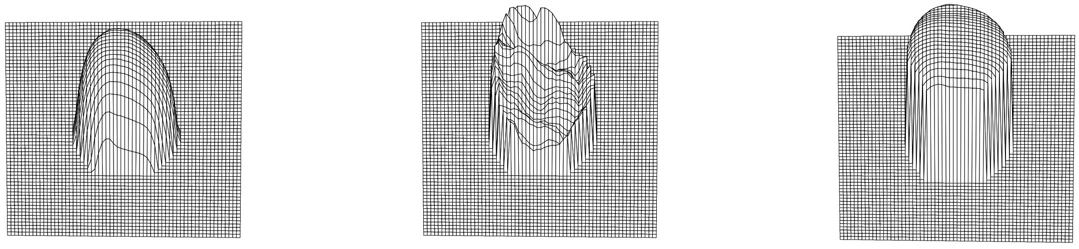

(a) (b) (c)

Figure 8: Tomato images. (a) The low frequency information from stereo. (b) The high frequency information from shading. (c) The combined low and high frequency information.

in Figure 10.(a), and the high frequency contents of shading are shown in Figure 10.(b). The combination of (a) and (b) by our method results in a better depth map, which is shown in 10.(c).

6.2 Results for real data

The proposed method was also tested on several real stereo pairs. The results are shown in Figures 11-15.

The results for the Renault images are shown in Figure 11. The stereo images are shown in Figure 11.(a)-(b), and the right stereo image is used for the shape from shading algorithm. The estimated depth map computed by the stereo matching algorithm is shown in Figure 11.(d), and the estimated depth map computed by the shape from shading algorithm is shown in Figure 11.(e). The obvious errors in details in the stereo results are noticeable. The object in this image does not have constant albedo, therefore Pentland's algorithm, which assumes constant albedo, encounters some problems. The results obtained by integrating stereo and shading using our method are shown in Figure 11.(c). This depth map is much better than the other two. The problems in surface details and problems due to variable albedo are almost eliminated. Figure 11.(f)–(h) shows the reconstructed gray level images using the estimated depth maps in (c)–(e) with the estimated light source direction $(-0.62, 0.50, 0.60)$. The reconstructed gray level images using the depth map obtained by integrating shape and

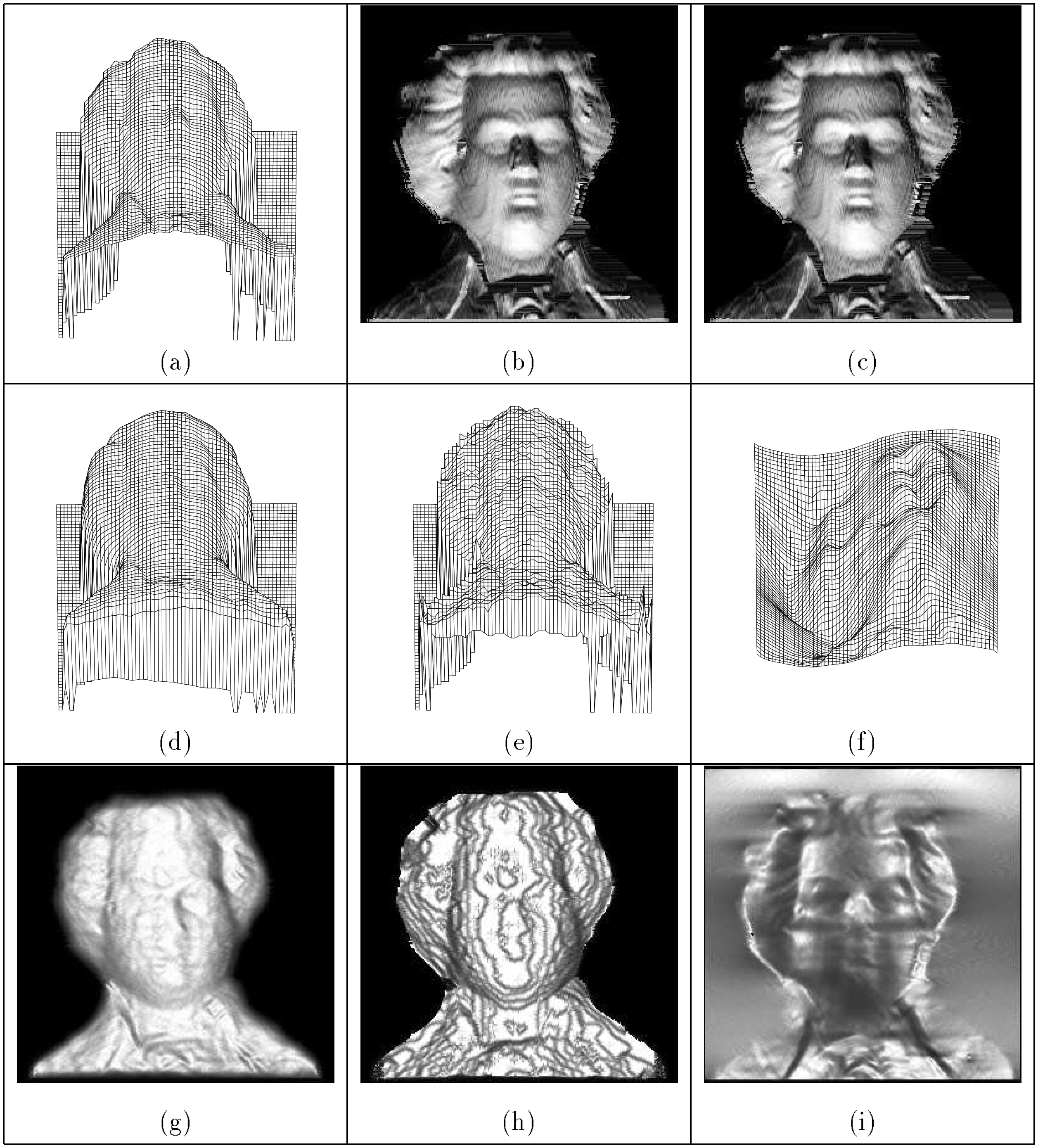


Figure 9: Results for the Mozart images. (a) A 3D plot of the range data. (b) Left stereo gray level image. (c) Right stereo gray level image. (d) A 3D plot of the estimated depth map by our method. (e) A 3D plot of the estimated depth map by stereo matching algorithm. (f) A 3d plot of the estimated depth map by Pentland's shape from shading algorithm. (g) A reconstructed gray level image using the estimated depth map in (d). (h) A reconstructed gray level image using the estimated depth map in (e). (i) A reconstructed gray level image using the estimated depth map in (f).

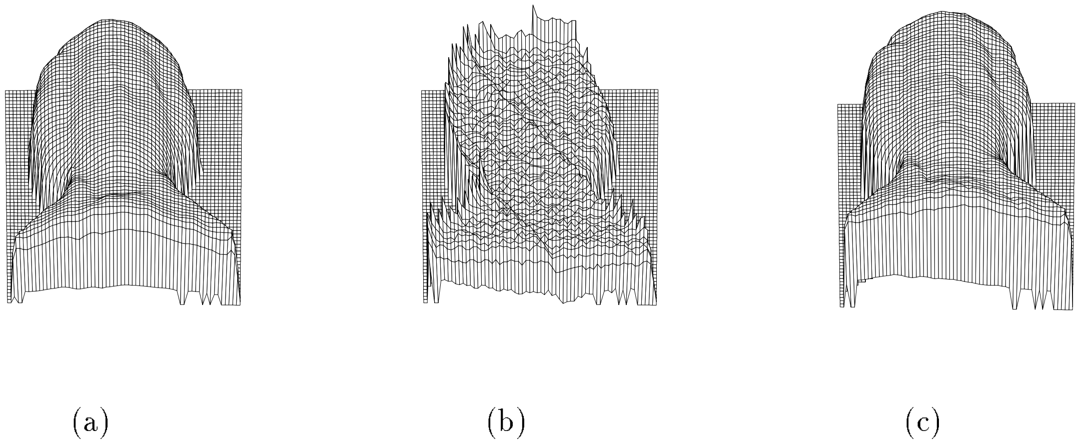


Figure 10: Mozart images. (a) The low frequency information from stereo. (b) The high frequency information from shading. (c) The combined low and high frequency information.

stereo are much closer to the original gray level images. The low frequency contents of stereo are shown in Figure 12.(a), and the high frequency contents of shading are shown in Figure 12.(b). The combination of (a) and (b) by our method results in a better depth map, which is shown in 12.(c).

Next, the results for Pentagon images are shown in Figure 13. The stereo images are shown in Figure 13.(a)-(b), and the right stereo image is used for the shape from shading algorithm. The estimated depth map computed by the stereo matching algorithm is shown in Figure 13.(d), and the estimated depth map computed by the shading algorithm is shown in Figure 13.(e). The stereo algorithm performs very poorly on this image. There are errors in depth everywhere. The shading depth map is more reasonable, and the detailed surface areas are reconstructed well. The results obtained by integrating stereo and shading using our method are shown in Figure 13.(c). The integrated depth map has some improvement over shading. Figure 13.(f)–(h) shows the reconstructed gray level images using the estimated depth maps in (c)–(e) with the estimated light source direction $(-0.76, -0.26, 0.59)$. The low frequency contents of stereo are shown in Figure 14.(a), and the high frequency contents of shading are shown in Figure 14.(b). The combination of (a) and (b) by our method results in a better depth map, which is shown in 14.(c).

Finally, the results for Sandwich images are shown in Figure 15. The stereo images are shown in Figure 15.(a)-(b), and the right stereo image is used for the shape from shading

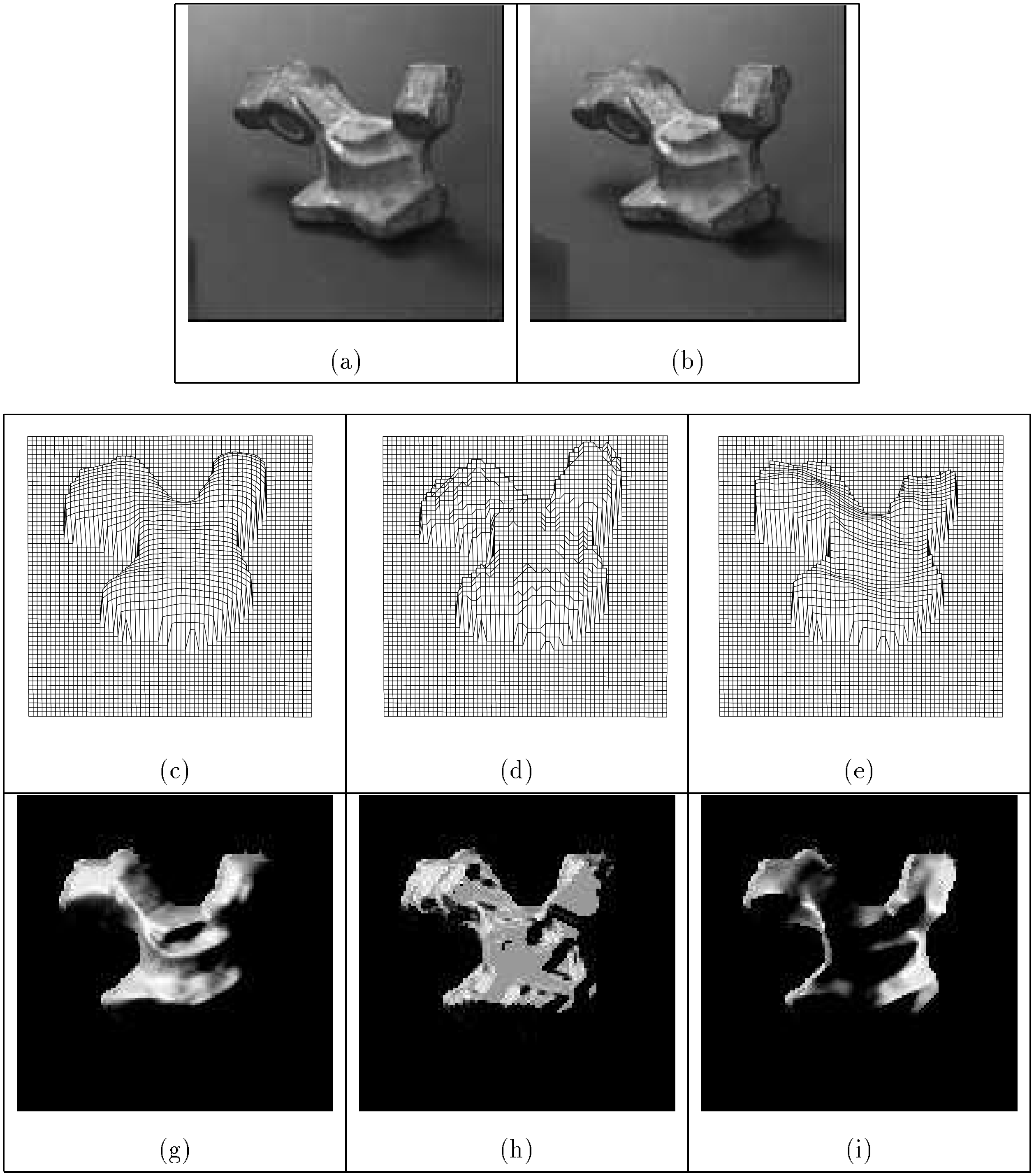


Figure 11: Results for the Renault images. (a) Left stereo image. (b) Right stereo image. (c) A 3D plot of the estimated depth map by our method. (d) A 3D plot of the estimated depth map by stereo matching algorithm. (e) A 3d plot of the estimated depth map by Pentland's shape from shading algorithm. (f) A reconstructed gray level image using the estimated depth map in (c). (g) A reconstructed gray level image using the estimated depth map in (d). (h) A reconstructed gray level image using the estimated depth map in (e).

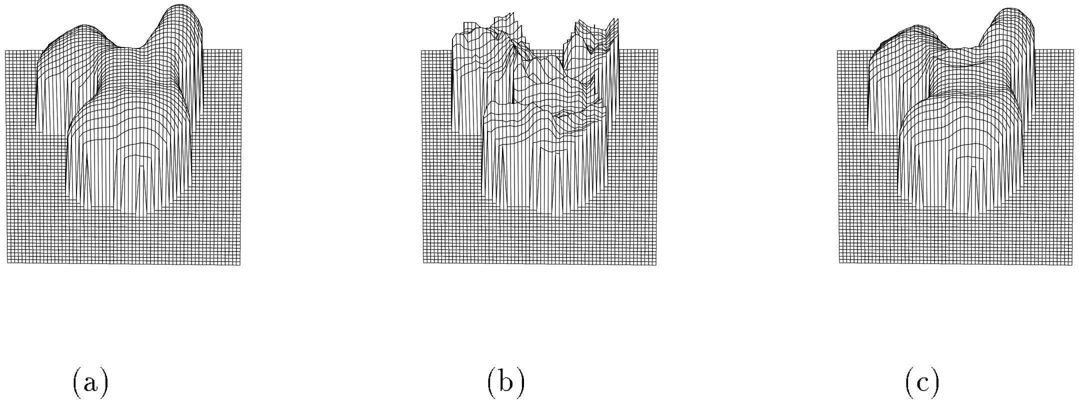
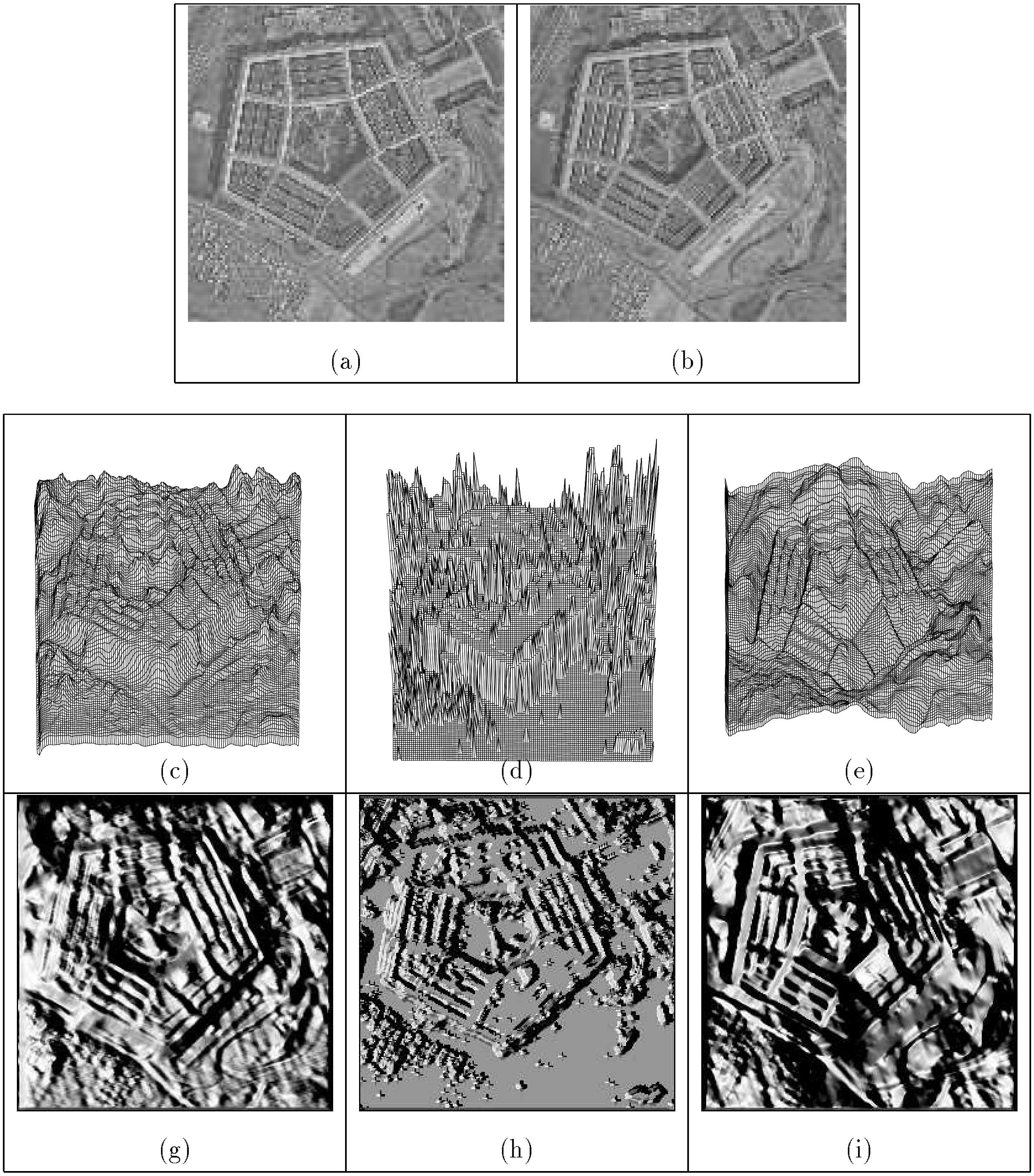


Figure 12: Renault images. (a) The low frequency information from stereo. (b) The high frequency information from shading. (c) The combined low and high frequency information.

algorithm. The estimated depth map computed by the stereo matching algorithm is shown in Figure 15.(d), and the estimated depth map computed by the shading algorithm is shown in Figure 15.(e). The results obtained by integrating stereo and shading using our method are shown in Figure 15.(c). The stereo depth map has many problems with surface details in the Sandwich surface; Instead of showing a flat planar surface it appears curved. The shading results are better than results for the stereo, but overall the sandwich surface does not appear planar due to changes in albedo. The integrated depth map is almost perfect, clearly showing one flat surface of the sandwich at a constant depth. Figure 15.(f)–(h) shows the reconstructed gray level images using the estimated depth maps in (c)–(e) with the estimated light source direction $(0.15, 0.78, 0.61)$. The problems in shading and stereo depth maps are highlighted in the reconstructed gray level images. However, the reconstructed gray level image obtained by the integrated depth map is reasonable. The low frequency contents of stereo are shown in Figure 16.(a), and the high frequency contents of shading are shown in Figure 16.(b). The combination of (a) and (b) by our method results in a better depth map, which is shown in 16.(c). Without the ground truth data for these real images, we cannot compute the error. However, we can clearly see the improvement from the 3-D plots and the reconstructed gray level images.



(a) (b)

(c) (d) (e)

(g) (h) (i)

Figure 13: Results for the Pentagon images. (a) Left stereo image. (b) Right stereo image. (c) A 3D plot of the estimated depth map by our method. (d) A 3D plot of the estimated depth map by stereo matching algorithm. (e) A 3d plot of the estimated depth map by Pentland's shape from shading algorithm. (f) A reconstructed gray level image using the estimated depth map in (c). (g) A reconstructed gray level image using the estimated depth map in (d). (h) A reconstructed gray level image using the estimated depth map in (e).

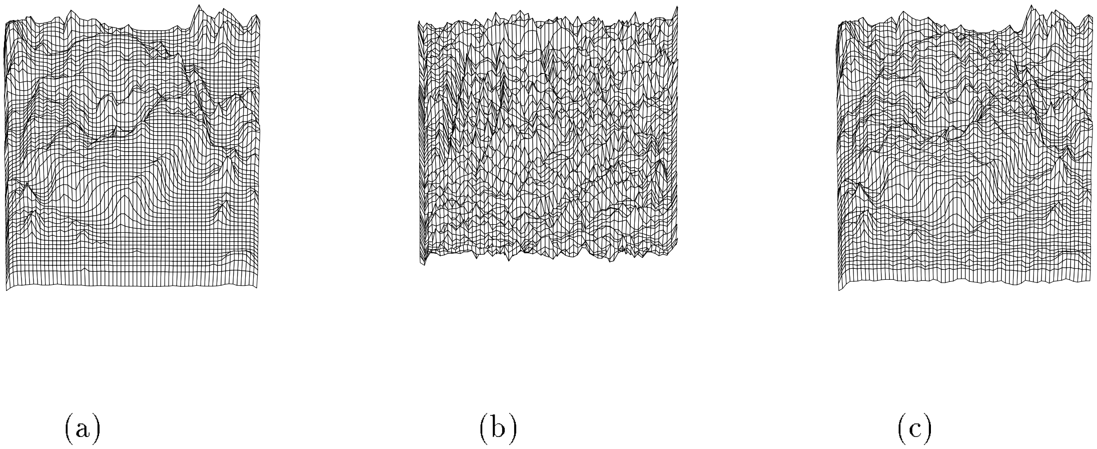


Figure 14: Pentagon images. (a) The low frequency information from stereo. (b) The high frequency information from shading. (c) The combined low and high frequency information.

7 Conclusions

In this paper we have addressed the problem of integration of *shape from X* modules. In particular, we have focussed on the combination of shape from shading and stereo. Our approach is very simple, and is motivated by the human vision system. Our criteria for integration shading and stereo is : *Keep the low frequency information from stereo, and add with the high frequency information from the shape from shading.*

Future work includes the integration of other shape from X modules, for example, motion, texture, and contour.

Acknowledgements: The authors are thankful to Professor Ikeuchi of Carnegie Mellon University for providing the *tomato* image, Dr. Stein of University of Southern California for providing the *Mozart* image, and Professor Ramesh Jain of University of Michigan for providing *Sandwich, Pentagon, and Renault* stereo pairs.

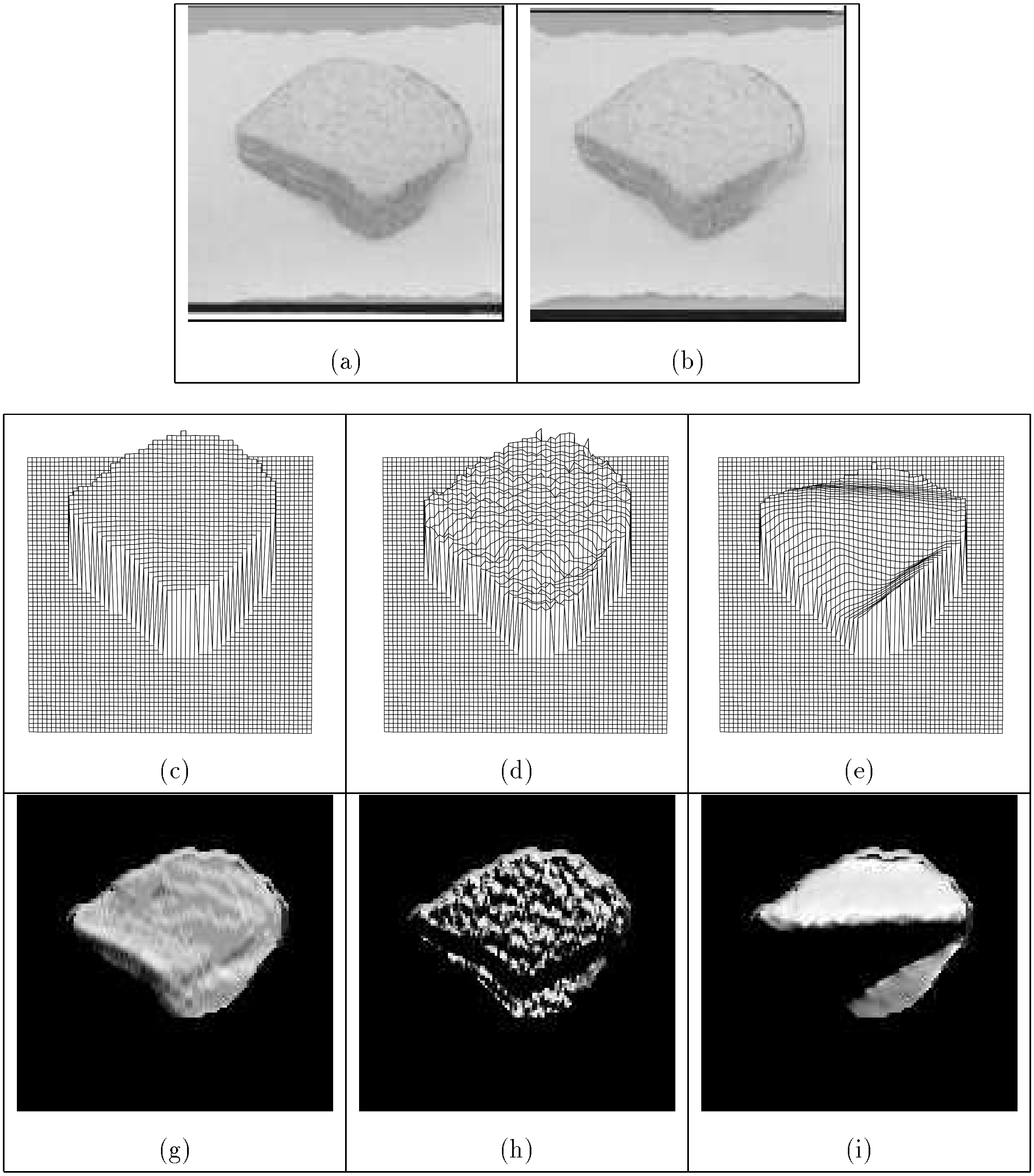


Figure 15: Results for the Sandwich images. (a) Left stereo image. (b) Right stereo image. (c) A 3D plot of the estimated depth map by our method. (d) A 3D plot of the estimated depth map by stereo matching algorithm. (e) A 3d plot of the estimated depth map by Pentland's shape from shading algorithm. (f) A reconstructed gray level image using the estimated depth map in (c). (g) A reconstructed gray level image using the estimated depth map in (d). (h) A reconstructed gray level image using the estimated depth map in (e).

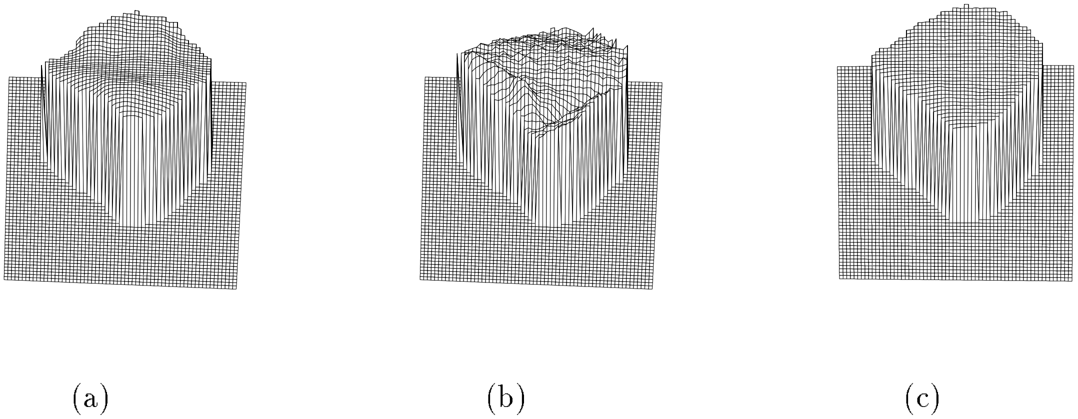


Figure 16: Sandwich images. (a) The low frequency information from stereo. (b) The high frequency information from shading. (c) The combined low and high frequency information.

References

[1] Aloimonos, J. and Shulman, David. Integration of Visual Modules. Academic press, 1989.

[2] Barnard, Stephen. A Stochastic Approach to Stereo Vision. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pp. 21-25, 1987.

[3] Blake, A., Zisserman, A. and Knowles, G. Surface Descriptions from Stereo and Shading. *Image and Vision Computing*, Vol. 3, No. 4, pp. 183-191, 1985.

[4] Frankot, Robert T. and Chellappa, Rama. A Method for Enforcing Integrability in Shape from Shading. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 4, pp. 439-451, 1988.

[5] Gonzalez, Rafael C. and Wintz, Paul. Digital Image Processing. Addison Wesley publishing company, 1987.

[6] Grimson, E. Binocular shading and visual surface reconstruction. *CVGIP*, pp. 18-44, 1984.

[7] Hall, Charles F. and Hall, Ernest L. A Nonlinear Model for the Spatial Characteristics of the Human Visual System. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 60, No. 7, 828-842, July 1972.

[8] G. Healey and T.O. Binford. Local shape from specularity. *Computer Vision, Graphics, Image Processing*, 42:62–86, 1988.

[9] Horn, B.K.P. Height and gradient from shading. *International Journal of Computer Vision*, 5:37–75, 1990.

[10] Ikeuchi, K. and Horn, B.K.P. Numerical shape from shading and occluding Boundaries. *Artifical Intelligence*, Vol. 17, No. 1-3, pp. 141-184, 1981.

[11] Leclerc, Yvan G. and Bobick, Aaron F. The Direct Computation of Height from Shading. *Proceedings of Computer Vision and Pattern Recognition*, pp. 552-558, 1991.

[12] Lee, C.H. and Rosenfeld, A. Improved methods of estimating shape from shading using the light source coordinate system. *AI*, 26:125–143, 1985.

[13] Marr, D. *Vision.* Freeman, San Francisco, CA, 1982.

[14] Ikeuchi, K., Nayar, S. K. and Kanade, T. Shape from interreflections. In *Third International Conference on Computer Vision*, pp. 1–11, 1990.

[15] Pentland, A. Shape information from shading: a theory about human perception. In *Second International Conference on Computer Vision (Tampa, FL, December 5–8, 1988)*, pp. 404–413, Washington, DC, 1988. Computer Society Press.

[16] Tsai, Ping-Sing and Shah, Mubarak. A fast linear shape from shading. *Proceedings of Computer Vision and Pattern Recognition*, pp. 734-736 , 1992.

[17] Waxman, A. and Duncan, J. Binocular image flows. Workshop on Motion, Kiawah Island, SC, 1987.