

# Integrating Stereo and Shape from Shading

Mostafa G.-H Mostafa\*, Sameh M. Yamany, and Aly A. Farag

Computer Vision and Image Processing Lab, University of Louisville, Louisville, KY 40292.  
{mostafa,yamany,farag}@cvip.louisville.edu, <http://www.cvip.uofl.edu>.

## Abstract

*This paper presents a new method for integrating different low level vision modules, stereo and shape from shading, in order to improve the 3D reconstruction of visible surfaces of objects from intensity images. The integration process is based on correcting the 3D visible surface obtained from shape from shading using the sparse depth measurements from the stereo module by fitting a surface into the difference between the two data sets. A feedforward neural network is used to fit a surface to the error difference. An extended Kalman filter algorithm is used for the network learning. It is found that the integration of sparse depth measurements has greatly enhanced the 3D visible surface obtained from shape from shading in terms of metric measurements.*

## 1 Introduction

The problem of three-dimensional (3D) object reconstruction from two-dimensional (2D) images has attracted much attention due to its theoretical challenge and its many practical applications. The 3D reconstruction of a sensed scene is crucial for machine vision and robotics systems in order for these systems to autonomously interact with their environment. Individual vision modules (stereo, shading, texture, etc.) cannot accurately reconstruct the 3D structure of the imaged scene due to the insufficient data constraints and the presence of sensor noise. In general, the performance of 3D vision systems is greatly enhanced when various sources of information about the 3D scene (e.g., stereo, range, shading, etc.) are incorporated[1, 2].

This paper presents a framework for integrating sparse depth data from stereo and dense depth maps from shape from shading in order to improve the 3D reconstruction of visible surfaces of 3D objects. The integration process is based on propagating the er-

ror difference between the two data sets by fitting a surface to that difference and using it to correct the visible surface obtained from shape from shading. A feedforward neural network is used to fit a surface to the sparse data. We also show a new technique for calculating shape from shading using perspective projection. It is found that the integration of stereo measurements has greatly enhanced the 3D visible surface obtained from shape from shading in terms of metric measurements.

This article is organized as follows. First, stereo reconstruction is described in Section 2, and the perspective shape from shading is given in Section 3. Our integration approach and its analysis are presented in Section 4. Section 5 presents our results and discussions. Finally, the paper is summarized and concluded in Section 6.

## 2 Stereo Reconstruction

Stereo reconstruction is a robust technique that infers the 3D structure and depth of a scene from two or more images taken from different viewpoints by cameras of known relative positions and orientations[3]. Because of the *correspondence problem*, i.e. finding the point matches in the images, stereo techniques known to produce an erroneous depth map in regions of the image with no or little texture.

Given a pair of stereo images, the relation between the 3-D coordinates of a point  $\mathbf{M} = [X, Y, Z]^T$  in the scene and its image coordinates  $\mathbf{m} = [x, y]^T$  can be obtained from the the pinhole camera model by

$$s\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{M}} \quad (1)$$

where  $s$  is an arbitrary scale,  $\mathbf{P}$  is a  $3 \times 4$  matrix, called the perspective projection matrix, and  $\tilde{\mathbf{m}}$  and  $\tilde{\mathbf{M}}$  are the homogeneous coordinates of the vectors  $\mathbf{m}$  and  $\mathbf{M}$ , respectively. In general,  $\mathbf{P} = \mathbf{A}[\mathbf{R}, \mathbf{t}]$  where  $\mathbf{A}$  is a matrix containing all the camera intrinsic parameters, and  $\mathbf{R}$  and  $\mathbf{t}$  are the rotation matrix and translation vector, respectively. The two perspective images of a scene are related by the *epipolar geometry*, which can be described by a  $3 \times 3$  singular matrix. It contains

---

\*This work is supported in part by DoD under contract: USNV N00014-97-11076. M. Mostafa is on leave from Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

all the information that is necessary to establishing correspondences between two images, from which the 3D structure of the sensed scene can be inferred[4].

### 3 Shape from shading using perspective projection and camera calibration

Shape From Shading (SFS) is a low-level vision module that produces a dense depth map, and it is defined as *the process of recovering the 3D visible surface of a 3D object from the shading in its intensity image by using information about the reflectance and illumination properties of the scene*[5].

The surface orientation at a point  $\mathbf{M}$  on a surface  $S$  is determined by the unit vector perpendicular to the plane tangent to  $S$  at  $\mathbf{M}$ . Most of the research done in SFS suppose that orthographic projection is used to map a point  $\mathbf{M}$  onto the image plane which is not adequate if the camera is very close to the object. We propose to use the perspective projection matrix to enhance the SFS algorithm.

Using perspective projection, a function  $s(x, y)$ , corresponding to the scalar  $s$ , maps a point  $\mathbf{M}$  in the space to a point  $\mathbf{m}$  in the image. The normal to the surface at  $\mathbf{M}$  is defined to be the cross product of the two gradient vectors  $\mathbf{p} = \frac{ds(x,y)}{dx}$ ,  $\mathbf{q} = \frac{ds(x,y)}{dy}$ . Thus  $\mathbf{p}$  and  $\mathbf{q}$  are vectors and not scalars as in the case of the orthographic projections. Their surface reflectance  $R(\cdot)$  becomes function of the scalar  $s$  defined in equation[1] as follows

$$R(s) = \frac{(\mathbf{p} \times \mathbf{q}) \cdot \mathbf{L}}{|\mathbf{p} \times \mathbf{q}| |\mathbf{L}|}, \quad (2)$$

where  $\mathbf{L}$  is the illumination direction. The new formulation of the SFS problem becomes finding the scalar  $s$  that solves the new brightness equation;  $g(s) = E(x, y) - R(s) = 0$ . This can be solved using Taylor's series expansion and applying the Jacoby iterative method where at the  $n^{th}$  iteration, for each point  $(x, y)$  in the image,  $s_{x,y}^n$  is as follows:

$$s_{x,y}^n = s_{x,y}^{n-1} + \frac{-g(s_{x,y}^{n-1})}{\frac{dg(s_{x,y}^{n-1})}{ds_{x,y}}}. \quad (3)$$

Finding  $\frac{d}{ds_{x,y}}g(s_{x,y}^{n-1})$  involves many steps using linear algebra and partial differentiation and the close form equation can be obtained as follows:

$$\begin{aligned} \frac{dg(s_{x,y}^{n-1})}{ds_{x,y}} &= -\frac{dN}{ds_{x,y}} \cdot \frac{\mathbf{L}}{|\mathbf{L}|} \\ \frac{dN}{ds_{x,y}} &= \frac{d\mathbf{v}}{ds_{x,y}} \frac{1}{\sqrt{\mathbf{v}^t \mathbf{v}}} - \frac{\mathbf{v}}{\sqrt{(\mathbf{v}^t \mathbf{v})^3}} \left( \mathbf{v}^t \frac{d\mathbf{v}}{ds_{x,y}} \right) \end{aligned}$$

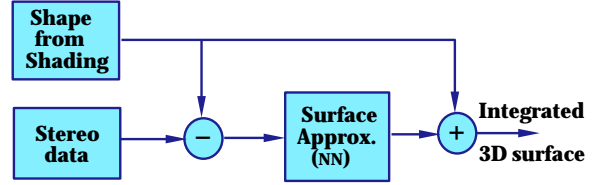


Figure 1: Functional block diagram for the integration process of shape from shading and stereo data.

$$\begin{aligned} \frac{d\mathbf{v}}{ds_{x,y}} &= \mathbf{B}^{-1} \mathbf{m} \times \mathbf{B}^{-1} (0, s_{x,y-1}, 0)^t \\ &+ \mathbf{B}^{-1} (s_{x-1,y}, 0, 0)^t \times \mathbf{B}^{-1} \mathbf{m}, \quad (4) \end{aligned}$$

where  $\mathbf{v} = \mathbf{p} \times \mathbf{q}$ . This representation will be accurate up to a scale factor.

### 4 The Integration Methodology

The main concern of this work is *the integration of the dense depth map obtained from the shape from shading with a sparse depth data from stereo for the reconstruction of 3D visible surfaces with accurate metric measurements*. This integration has two advantages. First, it helps in resolving, to some extent, the ambiguity of the 3D visible surface discontinuities produced by shape from shading due to highly textured regions. Second, it compensates for the missing metric information in the shape from shading by employing the data obtained from stereo, which work better at the highly textured regions. On the other hand, the sparse depth data from stereo does not contain all the depth information about the surface, so it cannot be used directly to represent the visible surfaces accurately. The integration process is based on propagating the error differences between the available depth data from stereo and the shape from shading throughout the remaining measurements where only shape from shading data are available. This can be done in three steps, as depicted in Fig. 1. First, the error difference in the depth measurements between the data sets is calculated. Second, we fit a surface to that error difference. Finally, the resultant surface is used to correct the shape from shading data. The surface fitting process, which is cast as a function approximation, is carried out in this paper using neural networks.

#### 4.1 Surface Approximation

Surface interpolation, has been one of the most intensely studied problems in low-level computer vision[6, 7, 8]. It plays a central role in the construction of a continuous  $2\frac{1}{2}$ -D *sketch* from sparse vi-

sual data[3]. The computational theories used in conjunction with surface interpolation include variational principles and regularization theory[6]. The common element of these computational theories is the minimization of a global energy function composed of many local energy components. This minimization has usually been implemented using iterative algorithms. A weak point in these methods is that it has adjustable parameters and does not perform well in case of sparse data. Another common problem is that they can find only locally optimal solutions, instead of finding the global minimum of the energy function.

Surface interpolation is considered a function approximation problem. Consider a nonlinear input-output mapping defined by the functional relationship  $\mathbf{Z} = \mathbf{f}(\mathbf{x})$ , where the vector  $\mathbf{x}$  is the input and the vector  $\mathbf{Z}$  is the output. The mapping valued function  $\mathbf{f}(\cdot)$  is assumed to be unknown. Given a limited set of input-output examples  $\{\mathbf{x}_i, \mathcal{Z}_i\}; i = 1, \dots, N$ , the requirement is to find the function  $\mathbf{F}(\cdot)$  that approximates  $\mathbf{f}(\cdot)$  over all inputs. That is,

$$\|\mathbf{F}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| < \epsilon \quad \text{for all } \mathbf{x}, \quad (5)$$

where  $\epsilon$  is a small threshold. This function approximation problem is a perfect candidate to supervise learning with  $\mathbf{x}_i$  playing the role of input vector and  $\mathcal{Z}_i$  serving the role of desired response.

In supervised learning of a multilayer neural network, the set of examples  $\{\mathbf{x}_i, \mathcal{Z}_i\}$  is used to train the neural network as a model of the unknown system. If  $\mathbf{y}_i$  is the output of the network produced in response to an input vector  $\mathbf{x}_i$ . The difference between the  $\mathcal{Z}_i$  (associated with  $\mathbf{x}_i$ ) and the network output  $\mathbf{y}_i$  provides the error signal  $\mathbf{e}_i = \mathcal{Z}_i - \mathbf{y}_i$ , which is equivalent to the difference in Eq.5. The total error energy in the network output is  $E(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n)$ , where  $C$  is the set of output neurons, is used to minimize the squared difference between the outputs of the unknown system and the neural network in a statistical sense. For a given training set,

$$E_{av} = \frac{1}{N} \sum_{n=1}^N E(n) = \frac{1}{N} \sum_{n=1}^N \left[ \frac{1}{2} \sum_{j \in C} e_j^2(n) \right], \quad (6)$$

is the average squared error energy which represents the *cost function* as a measure of learning performance. The objective of the learning process is to adjust the network free parameters (the synaptic weights) to minimize the cost function  $E_{av}$ . This minimization can be achieved by considering a training method in which the weights are updated on a pattern-by-pattern basis until one *epoch*, that is, one complete

presentation of the entire training set has been dealt with. The adjustments to the weights are made in accordance with the respective errors computed for each pattern presented to the network.

#### 4.1.1 The extended Kalman filter learning algorithms

Although the *back-propagation* (BP) learning algorithm[9, 10] is considered the most popular algorithm for training multilayer networks, it has two main problems. First, appropriate learning parameters  $\eta$ ,  $\alpha$ , and  $\lambda$  (for the activation function) need to be carefully chosen for each training set. The tuning of these free parameters is not trivial. Second, the rate of convergence tends to be relatively slow due to the stochastic nature of the algorithm, which in turn makes it computationally extensive.

The limitations of the BP algorithm may be overcome by viewing the supervised training of the network as an optimum filtering problem, the solution of which recursively utilizes information contained in the training data traced back to the first iteration of the learning process, a situation where the Kalman filter theory can be utilized[10]. Singal *et al.* suggested the idea of using the extended Kalman filter to train a multilayer perceptron[11]. A Kalman filter attempts to estimate the state of a system that can be modeled as a linear system driven by a white Gaussian noise and where the measurements available are linear combinations of the system states corrupted by additive white Gaussian noise[12]. The weights of the multilayer perceptron are the states the Kalman filter attempts to estimate and the output of the network is the measurement used by the Kalman filter.

Consider a multilayer perceptron with  $W$  synaptic weights and  $p$  output nodes. Let the vector  $\mathbf{w}(n)$  denotes the synaptic weights of the entire network at time  $n$ . The state-space equations of the network may be modeled as follows[11].

$$\mathbf{w}(n+1) = \mathbf{w}(n), \quad (7)$$

$$\mathbf{y}(n) = \mathbf{h}(\mathbf{w}(n), \mathbf{u}(n)) + \mathbf{v}(n). \quad (8)$$

where  $\mathbf{y}(n)$  is the output of the system at time  $n$ ,  $\mathbf{h}$  is the nonlinear function mapping the states to the observations,  $\mathbf{u}(n)$  is the input to the system, and  $\mathbf{v}(n)$  is a Gaussian noise. To apply the Kalman filter theory to the state-space model just described, Eq. 8 must be linearized in the form

$$\mathbf{d}(n+1) = \mathbf{H}(n)\mathbf{w}(n) + \nu(n), \quad (9)$$

where  $\mathbf{H}(n)$  is the  $b$ -by- $W$  measurement matrix of the linearized model, which consists of the partial derivatives of the  $p$  outputs of the whole network with respect to the  $W$  weights of the model. Using the linearized model just described, the Kalman filter equation (see [13] and [10] ) can be used for updating the weights. The decoupled nature of the extended Kalman filter (DEKF) algorithm[10] minimizes the cost function:

$$E(n) = \frac{1}{2} \sum_{j=1}^n \|e_j\|^2, \quad (10)$$

where  $e_j$  is the error vector defined by Eq. 6.

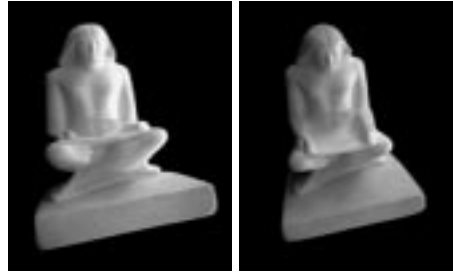
A performance analysis in a previous work[14] shows that the EKF algorithm outperforms the BP learning algorithm in terms of fast convergence and minimum root mean square (RMS) error. The best network topology is obtained by finding the topology that gives the minimum RMS error between the ground truth data and the results of the integration process.

## 5 Results and Discussions

We choose two different objects, a smooth object (vase) and a free form object (statue), to show the applicability of our approach. Figure. 2 shows the results for the statue of the Egyptian writer. It shows that the integration of few 3D data points, from the stereo reconstruction with the depth variations obtained from the perspective shading produces an enhanced 3D visible surface with metric information.

Figure. 3 shows the pair of stereo images of a vase, the visible surface obtained using shape from shading, and the visible surface obtained using the integration process. The figure shows that the integration has greatly improved the 3D reconstruction of the visible surface of the vase.

The results shown in the above two figures show the importance of integrating visual modules. Shape from shading module exploits the smoothness of a surface if the light source direction is well estimated, but performs badly when parts of the surface is highly textured. On the other hand, stereo module performs well when the camera is accurately calibrated and good correspondences are established. Correspondences are well established for textured surfaces. That is, shape from shading module works well where stereo module does not, and vice versa. Combining the results of both modules produces an enhanced 3D visible surface which can not be obtained from any of the two modules individually.



(a)



(b)



(c)

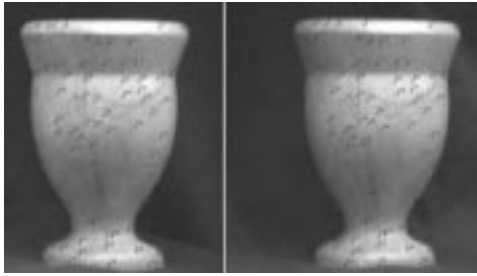


(d)

Figure 2: Results of the writer statue. (a) Intensity stereo images, (b)shape from shading, (c) Stereo reconstruction of matching points, and (d) Visible surface after the integration of the SFS and stereo data.

## 6 Conclusions

We have presented a methodology for integrating visual modules, shape from shading and stereo. The integration method is based on fitting a surface to the difference between the shape from shading and the stereo data. This approximated surface is used



(a)



(b)

(c)

Figure 3: Results of the vase. (a) Intensity stereo image pair with match points, (b) shape from shading, and (c) Visible surface after the integration of the SFS and stereo data.

to correct the shape from shading result. A feedforward multilayer neural network is used as a surface approximator. The network is trained using an extended Kalman filter algorithm. The surface approximation using the neural networks is found to work exceptionally well in learning and retrieving the smoothness of a surface, however it tends to smooth region of sharp discontinuities. The integration of a few stereo data points not only greatly improved the 3D visible surface reconstruction obtained from shape from shading, but also produces 3D visible surfaces representation with nearly accurate metric measurements.

The proposed integration method is found to perform exceptionally well for smooth surfaces, but for free form surfaces it found to have some difficulties in retrieving the exact surface at sharp discontinuities due to the existence of few stereo data points and the lack of information available from the shape from shading. In future work, we plan to improve the surface fitting technique such that it maintains the sharp discontinuities found in the few stereo data, and improve the stereo matching module to include

more matches at boundaries and discontinuities.

## References

- [1] M. A. Abidi and R. C. Gonzalez, eds., *Data Fusion in Robotics and Machine Intelligence*, Academic Press, 1992.
- [2] S. Pankanti and A. K. Jain, "Integrating vision modules: Stereo, shading, grouping, and line labeling," *IEEE Trans. on Patt. Anal. and Mach. Intell.* **17**, pp. 831–842, September 1995.
- [3] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman, San Francisco, California, 1982.
- [4] Z. Zhang, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Journal of Artificial Intelligence* **78**, pp. 87–119, 1995.
- [5] B. K. P. Horn, *Understanding image formation*, MIT Media Lab, Academic Press, 2nd ed., 1985.
- [6] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature* **317**(26), p. 314, 1985.
- [7] D. Terzopoulos, "The computation of visible-surface representation," *IEEE Trans. on Patt. Anal. and Mach. Intell.* **10**(4), pp. 417–438, 1988.
- [8] R. Szeliski, "Fast surface interpolation using hierarchical basis functions," *IEEE Trans. on Patt. Anal. and Mach. Intell.* **12**(6), pp. 513–528, 1990.
- [9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representation of back-propagation errors," *Nature* **323**, pp. 533–536, 1986.
- [10] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.
- [11] S. Singal and L. Wu, "Training feed-forward networks with the extended kalman algorithm," in *IEEE Int. Conf. Acoustic, Speech, and Signal Processing*, p. 1187, 1989.
- [12] P. S. Maybeck, *Stochastic Models, Estimation, and Control*, vol. 1, Academic Press, New York, 1979.
- [13] D. W. Ruck, S. K. Rogers, M. Kabrisky, P. S. Maybeck, and M. E. Oxley, "Comparative analysis of backpropagation and the extended kalman filter for training multilayer perceptron," *IEEE Trans. on Patt. Anal. and Mach. Intell.* **14**(6), pp. 686–691, 1992.
- [14] M. G. Mostafa, S. M. Yamany, and A. A. Farag, "Integrating shape from shading and range data using neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, p. 15, IEEE Computer Society, (Fort Collins, Colorado), June 23–25, 1999.