Shape from Appearance:
A Statistical Approach to Surface Shape Estimation*

Darrell R. Hougen and Narendra Ahuja

Beckman Institute and Coordinated Sciences Laboratory
University of Illinois, Urbana, Illinois 61801, USA

Abstract. This paper is concerned with surface shape estimation by a method in which an empirically determined associative model relating appearance to surface shape is used. Significantly, the estimated model is more accurate than the algorithm that generates the examples. The method presented here is a generalization of shape from shading methods that does not rely upon idealized models of the image formation process. As a relative of shape from shading, this method more accurately recovers small surface detail than is possible with methods such as stereo and motion. The present approach is a continuous analogue of pattern recognition and is closely related to methods of joint space learning used in robotics. Experiments on real scenes are used to illustrate the concepts involved.

1 Introduction

This paper describes a method of surface shape estimation that involves automatic generation of an *associative model* that relates surface shape to appearance. It is shown that through a scale change and the use of a smoothness requirement, the estimated model can be made to be more accurate than the algorithm that produced the examples. The performance increase is key to the utility of this method and sets it apart from the approach of Lehky and Sejnowski [7]

Associative modelling techniques are considered by the authors to be important because of the generality and precision made possible by such techniques. The shape estimation procedure described below is a generalization of physics based methods and embodies many of the advantages of such methods with few of the disadvantages. Physics based methods generally rely heavily on idealized models of the image formation process which do not capture the complexity of real scenes [2, 3, 4]. In addition, such models often contain hard to estimate parameters [3, 4]. However, shading information, in particular, is useful for recovering small surface detail.

* This research was supported by the Advanced Research Projects Agency and the National Science Foundation under grant IRI-89-02728 and by the Army Advance Construction Technology Center under grant DAAL 03-87-K-0006.

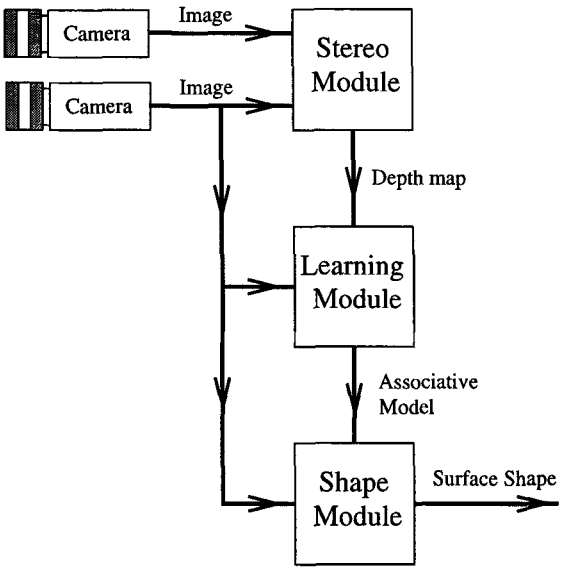

Fig. 1. A stereo module produces coarse shape estimates. A learning module produces an associative model of the relationship of shape to appearance. The shape module estimates the surface shape of novel objects.

In comparison, methods that rely upon the coincidence or correlation of features in two or more images, such as stereo and motion based methods, cannot be used to recover small surface detail due to the sparseness of discriminable image features [5] and the fact that the accuracy of such techniques drops with the square of depth [5]. In addition, such methods are only reliable in highly textured regions or in the presence of well defined image features. However, such methods are based on relatively weak assumptions and are therefore useful for recovering coarse or sparse depth estimates.

Recent papers by Leclerc and Bobick [6] and Hougen and Ahuja [3, 4] discuss integrated methods which combine the strengths of the above methods while avoiding many of the limitations. However, despite the increase in generality, such methods are still dependent upon highly restrictive idealized models of the image formation process.

In order to escape such restrictions, it should be noted that the relationship between local appearance and corresponding surface shape can always be captured in the form of a probability density function. The density estimation problem encountered here is the continuous analogue of the pattern recognition problem [1] and is closely related to function learning problems encountered in robotics [8]. Indeed, in the presence of nonlocal, contextual information, it may be possible to simplify the problem and treat the local relationship of shape to appearance as a functional relationship. This approach is explored in the following sections.

2 Algorithm Overview

The three major components of the local shape-from-appearance estimation procedure are illustrated in figure 1 including, (1) generation of coarse surface shape estimates for statistical modelling, (2) estimation of the associative model governing the statistical relationship of shape to appearance, and (3) surface shape estimation through application of the statistical model to the desired image.

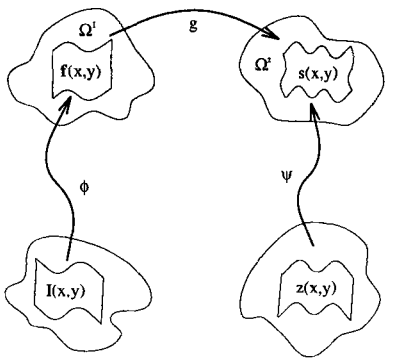

Fig. 2. Associative mapping. The image, I, is represented through ϕ by a set of features f and the surface z is represented through ψ by the shape function s. The learning module estimates an associative mapping g from f to s. The shape module finds the shape function, s, associated with a novel image and inverts ψ to produce the corresponding surface height function, z.

Generation of the coarse surface shape estimates involves the use of a standard vision algorithm such as a stereo or motion algorithm. Such algorithms are capable of producing coarse yet reasonable estimates under a wide range of conditions.

Estimation of the associative model governing the relationship of shape to appearance is accomplished by associating examples of estimated surfaces with corresponding examples from images as illustrated in figure 2. Each example of surface shape and appearance is represented by one or more mathematical features which are designed to capture important information about the examples. The relationship between corresponding examples is represented by a statistical model and serves as a substitute for the derivation of an idealized model.

Physics based methods are generally based upon restrictive assumptions. The present method requires only that the conditional density of shape given appearance be sufficiently informative. If the relationship is particularly simple, it may be possible to replace estimation of the probability density with estimation of a mapping function. In this case the model is referred to as an *associative map*. More detail appears in the following sections.

Surface shape estimation is accomplished through application of the previously estimated associative model to the desired image. The features calculated at a particular image location serve as input vectors to the associative model, allowing the statistically expected shape parameters to be computed at that location. A more recognizable surface description is obtained by inverting the feature calculation process as described in section 5.

3 Associative Modelling

Let $z : A \rightarrow \Re^-$ be a surface height function and $I : A \rightarrow [0, I_{\max}]$ be an image brightness function defined on $A \subset \Re^2$. The functions z and I may be viewed as members of the ensembles $\mathcal{Z}$ and $\mathcal{I}$ respectively where $\mathcal{Z}$ is the set of all viewable height functions and $\mathcal{I}$ is the set of all images of the height functions in $\mathcal{Z}$.

For a single surface height function, variable lighting conditions and surface marking patterns make possible a wide variety of possible images. Conversely, a single image may correspond to more than one surface height function. However,

some height functions are more likely than others to have produced a given image. The problem considered here is to find the surface *most likely* to have produced a given image.

3.1 Choosing Features

Important criteria for choosing a surface representation are simplicity, symmetry, completeness, and learnability. An operator, ψ, that at least partially satisfies these criteria is the Laplacian of a Gaussian smoothing filter. It is simple, symmetric, and complete; z can be reconstructed from $s(x,y;\sigma) = \nabla^2 G(\sigma) * z$. It is also more apparent locally, and therefore more learnable, than, for example, surface height or slope.

The most important criterion for choosing an image representation is that it be locally informative. The image irradiance at a point is uninformative but higher order functions of the local image irradiance are statistically related to surface shape and therefore useful. In general, the image, I, is represented by an image feature vector, $\boldsymbol{f}(\sigma)$, where $f_j(\sigma) = \phi_j(\sigma)I$, $j = 1,\ldots,k$. Here, polynomial coefficients are used.

3.2 Conditional Density

At a point, (x,y), the conditional density of shape given appearance in the local neighborhood of (x,y) is written $p_{\Lambda(\sigma)}(s(x,y;\sigma)|\boldsymbol{f}(x,y;\sigma))$ and is scale dependent. If the image feature vector is given on $A \subset \Re^2$ and the surface feature vector at a particular point depends only on the image feature vector at that point then the maximum log-likelihood solution for surface shape given the information about appearance is given by

$$z(x,y) = \max_{z(x,y)} \int\int_A \log(p_{\Lambda(\sigma)}(s(x,y;\sigma)|\boldsymbol{f}(x,y;\sigma)))dxdy$$

where the maximization is over all functions $z : A \rightarrow \Re^-$.

The above formulation can be simplified by using the marginal probability density in which the dependency upon σ is removed. This is a reasonable simplification given the fact that most objects that appear in the world are observed at many ranges and therefore at many scales making the conditional density nearly independent of scale. More formally, it is assumed that, $p_{\Lambda(\sigma)} \approx p_\Lambda$.

3.3 Scaling to Increase Performance

The fact that a surface curve or bend or marking has previously been seen up close, allowing its shape to be accurately determined, means that the shape can be accurately determined later when the surface is far away. The statistical association between surface shape and appearance determined at a coarse scale is stored in the form of a conditional probability density and used later to estimate

the surface shape at a finer scale. If the probability density is scale independent, it is not necessary to know the change in scale to reconstruct the surface.

As a preliminary, it should be noted that scaling to increase power can only work if the new shape estimation procedure has higher performance than the example generator. As an example, the resolution and accuracy are higher for shape from shading methods than for stereo, motion, or focus methods [5]. The present method is a relative of shape from shading and has similar performance characteristics, making possible an increase in performance.

Although the precise increase in performance is the subject of ongoing research, the following considerations are relevant. The probability density, p_A, should be estimated using examples given at a scale σ_e that is chosen to obtain the maximum performance from the example producing algorithm. If the resultant surface is estimated at scale a σ_r that is chosen optimally or suboptimally with $\sigma_r \geq \sigma_{\text{opt}}$. Then, if $\sigma_r < \sigma_e$ the resolution and hence the performance is increased by a factor related to σ_e/σ_r.

4 Associative Mapping

In many instances considered in computer vision, the probability density may be simple enough to be well approximated by a sum of normal variates. The correspondence of a smooth surface to a smoothly varying image or of a long narrow specularity to a surface with a convex or concave bend are examples of unimodal or bimodal distributions. If the mode is assumed known, the maximization of the log-likelihood reduces to the solution of a least squares problem. This is not unreasonable in cases in which a single choice is required for an entire region as is the case that the surface curvature has constant sign in a region of interest.

Let $g : \Omega^I \rightarrow \Omega^z$ be a function from the image feature space, Ω^I, into the surface feature space, Ω^z, such that $s(x,y;\sigma) = g(\boldsymbol{f}(x,y;\sigma)) + \epsilon$ where ϵ is zero mean Gaussian white noise. If the conditional density, p_A, is Gaussian white noise with mean, $s(x,y;\sigma)$, then g is guaranteed to exist and is given by $g(\boldsymbol{f}(\sigma)) = \langle s(\sigma)|\boldsymbol{f}(\sigma)\rangle$ where $\langle\cdot\rangle$ denotes expectation. Thus, estimation of g is a regression problem. Note that the mapping function, g, is assumed to be independent of the the scale factor σ. If the mapping is scale dependent, then $g(\boldsymbol{f}(\sigma))$ may be written $g(\boldsymbol{f}(\sigma);\sigma)$.

Let $(\boldsymbol{f}_i(\sigma_e), s_i(\sigma_e))$, $i = 1,\ldots,N$, be pairs of image feature vectors and corresponding surface shape estimates generated by the stereo program. In order to obtain the least squares estimate of g a criterion function Q is defined by

$$Q(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(s_i(\sigma_e) - g(\boldsymbol{f}_i(\sigma_e);\boldsymbol{\theta}))^2$$

where g is parameterized by $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)$.

If g is modelled by a linear sum of basis functions, $(B_1,\ldots,B_m)$, with coefficients $(\theta_1,\ldots,\theta_m)$. Then $\hat{g}$ may be written, $\hat{g}(\boldsymbol{f}_i(\sigma_e)) = \sum_{j=1}^{m}\theta_j B_j(\boldsymbol{f}_i(\sigma_e))$. In this case, the regression reduces to an ordinary linear least squares problem.

Figure 2 illustrates the relationships between $I, \phi(\sigma), \boldsymbol{f}(\sigma), z, \psi(\sigma), s(\sigma)$ and g. The feature operators, $\phi(\sigma)$ and $\psi(\sigma)$ transform the functions $I(x,y)$ and $z(x,y)$ into $\boldsymbol{f}(x,y;\sigma)$ and $s(x,y;\sigma)$ respectively. The function g maps each point of $\boldsymbol{f}(x,y;\sigma)$ to a point in the surface feature space that differs from $s(x,y;\sigma)$ by an amount ϵ.

5 Surface Estimation

In the preceding analysis, g was treated as a random variable with mean $s(\sigma_e)$. For purposes of surface estimation, $s(\sigma_r)$ is identified as a random variable with mean g. If $s(\sigma_r)|\boldsymbol{f}(\sigma_r)$ has a normal distribution with mean $g(\boldsymbol{f}(\sigma_r))$ and variance ρ^2 then $\log(p_A(s(\sigma_r)|\boldsymbol{f}(\sigma_r))) = (s(\sigma_r) - g(\boldsymbol{f}(\sigma_r)))^2/(2\rho^2) - \kappa$ where $\kappa = \frac{1}{2}\log(2\pi\rho^2)$ is a constant. Therefore, the maximum likelihood estimate of the surface is found by maximizing the criterion function

$$D(z) = \int\int_A (s(x,y;\sigma_r) - g(\boldsymbol{f}(x,y;\sigma_r)))^2 dxdy$$

over all surfaces $z : A \rightarrow \Re^-$.

Let $Z, H(\sigma_r), S(\sigma_r)$ and $\Psi(\sigma_r)$ be the Fourier transforms of $z, g(\boldsymbol{f}(\sigma_r)), s(\sigma_r)$ and $\psi(\sigma_r)$ respectively. Then, by Parseval's theorem,

$$D(z) = \int\int |S(\omega_1,\omega_2;\sigma_r) - H(\omega_1,\omega_2;\sigma_r)|^2 d\omega_1 d\omega$$

The minimum integrated squared error is achieved by the function that minimizes the error at each point of the domain. Since $S(\sigma_r) = \Psi(\sigma_r)Z$, that minimum is achieved by setting, $\hat{Z} = H(\sigma_r)/\Psi(\sigma_r)$. The solution surface, $\hat{z}$, is the inverse transform of $\hat{Z}$.

Although $\hat{z}$ is the maximum likelihood estimate in the absence of noise, a better estimate in the presence of noise is found by Wiener filtering. The resulting optimal estimate is found by setting

$$\hat{Z} = \frac{H(\sigma_r)\Psi(\sigma_r)}{\Psi^2(\sigma_r) + K^2}$$

where $K^2(\omega_1,\omega_2) = \langle \eta^2(\omega_1,\omega_2)/Z^2(\omega_1,\omega_2) \rangle$ is the variance of the noise divided by the expected power spectrum of the surface. Under the assumption of white noise, the noise term reduces to a constant. If the surface is assumed to be fractal Brownian, the final value of K^2 is given by $K^2 = \eta^2(\omega_1^2 + \omega_2^2)^2$.

6 Experimental Results

The algorithm described in section 2 can be thought of as operating in two major modes, the model estimation mode and the surface estimation mode. In the model estimation mode, the input images are used by the stereo module to produce surface shape estimates which serve as examples to be used in estimating

the associative model. In the surface estimation mode, the model is used to estimate the shape of a previously unseen surface. The experiments explained in this section are designed to illustrate both major operational modes.

6.1 Model Estimation Results

Figure 3 shows one of four images of an oriented ridge surface with its stereo depth map and corresponding level curves. The image is the left image of a stereo pair of images taken at a depth of about 10cm with a baseline of about 1cm. Once the surface estimates have been computed by the stereo program, regions from each image and corresponding surface are selected to act as input data for the model estimation procedure.

The original data regions are converted to data points, $(\boldsymbol{f}_i(\sigma_e), s_i(\sigma_e)), i = 1, \ldots N$ through the action of the feature operators, $\boldsymbol{\phi}(\sigma_e)$ and $\psi(\sigma_e)$. For the experiments reported here, the image feature operators were defined to be local, second degree polynomial fits and the output features were the polynomial coefficients. Figure 4 shows a plot of the surface data projected onto a two-dimensional subspace of the image feature space along with two projections of the mapping function, g, which is represented by a low degree polynomial. The size of each dot shows its magnitude. The clear trend in the data suggests that a low order model should account for a large percentage of the variance.

6.2 Surface Estimation Results

This test is designed to show that the system can recover the shape of a previously unseen surface that is very different from the surfaces used in the model estimation phase. The shape estimation procedure is conducted using the model estimated from the ridge surfaces of the previous section.

Figures 5 and 6 illustrate the shape recovery process applied to the image of a clay face. The first step of the surface recovery procedure is calculation of the image feature vector at every image location. The middle image in figure 5 shows one of six feature arrays produced using the local polynomial fit method described above using a scale factor σ_r. Each feature can be thought of as encoding a particular type of information about the local image structure. Once the feature arrays have been computed, the mapping function g is evaluated at each point producing a surface map, $s(\sigma_r)$, shown at the right of figure 5.

The final step of the surface estimation procedure is deconvolution of the surface map using the kernel, $\psi(\sigma_r)$, to obtain a surface height map, z. The resultant depth map, corresponding level curves, and reconstructed image are shown in figure 6. Note that the current method does not depend upon stereo during the surface estimation procedure and therefore produces a smooth surface. Another example is shown in figure 7 which shows a picture of a human subject followed by the corresponding depth map, level curves, and shaded depth map.

7 Conclusions

The shape from appearance method is a new method for estimating surface shape based on a learned associative model. The model is generated by associating examples of local surface shape with corresponding image features. The model may be a probability density or an associative map. The associative map is easier to use but can only be used in the presence of sufficient contextual information.

The experiments described in this paper involve the use of a stereo module as a source of local shape examples. It has been shown that through a scale change, the associative model can be made more accurate than the stereo algorithm. As a consequence, it is possible to recover the shape of an unknown surface more accurately than is possible with the stereo algorithm.

In general, the method reported here is more accurate than stereo, motion, or focus based methods. Shading based methods are also quite accurate, but they are based on strong and often unrealistic assumptions about reflectance, lighting, shadowing and other scene characteristics. The method described here is based on much weaker assumptions. There is still much theoretical development to be done. However, its generality promises to make it useful under a wider range of conditions than existing methods.

References

1. R. O. Duda and P. E Hart, *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons, 1973.
2. D. Forsyth and A. Zisserman, "Shape from shading in the light of mutual illumination," *Image Vision Comput.*, vol. 8, no. 1, pp 42–49, 1990.
3. D. R. Hougen and N. Ahuja, "Estimation of the Light Source Distribution and its Use in Shape Recovery from Stereo and Shading," in *Fourth Intl Conf. Comput. Vis.*, pp. 148–155, May 1993.
4. D. R. Hougen and N. Ahuja, "Adaptive Polynomial Modelling of the Reflectance for Shape Estimation from Stereo and Shading," in *Proc. IEEE Conf. Comput. Vis. Patt. Recog.*,, pp. 991–994, June 1994.
5. D. R. Hougen and N. Ahuja, "Resolution and Accuracy of Stereo, Motion, and Shading Methods," submitted to *Comput. Vis. Patt. Recog.*,, 1996.
6. Y. G. Leclerc and A. F. Bobick, "The direct computation of height from shading," *Proc. Comput. Vis. Patt. Recog.*, pp. 552–558, June 1991.
7. S. R. Lehky and T. J. Sejnowski, "Neural network model of visual cortex for determining surface curvature from images of shaded surfaces," *Proc. R. Soc. Lond. B,* vol. 240, pp. 251–278, 1990.
8. S. Omohundro, "Geometric learning algorithms," *Physica D,* vol. 42 pp. 307-321, 1990, and in *Emergent Computation,* ed. Stephanie Forrest, MIT Press 1991.

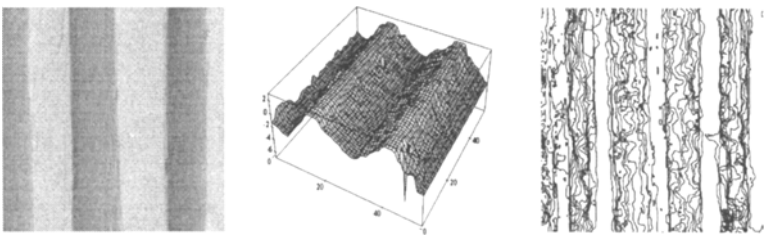

Fig. 3. One of four oriented ridge surface images with stereo depth map and corresponding level curves. The stereo program extracts a coarse estimate of the surface for use as input to the model building program.

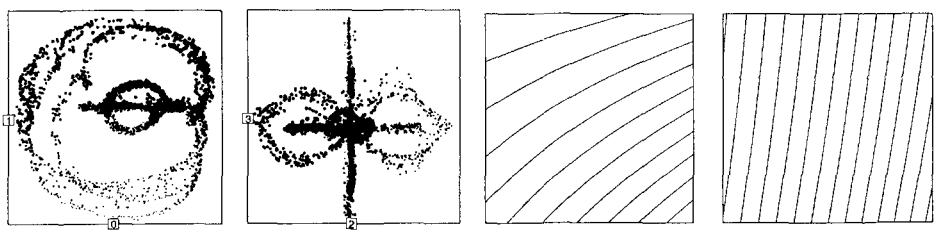

Fig. 4. Data collected from ridge images and resultant model projected onto a two-dimensional subspace of the image feature space. The size of each dot corresponds to value of the LOG surface feature. The model is a low order polynomial.

Fig. 5. Image of clay cherub figure, one of six feature maps, and shaded surface map, $s(x, y) = g(f(x, y))$. Each feature is one coefficient of second degree polynomial fit to the local gray level surface.

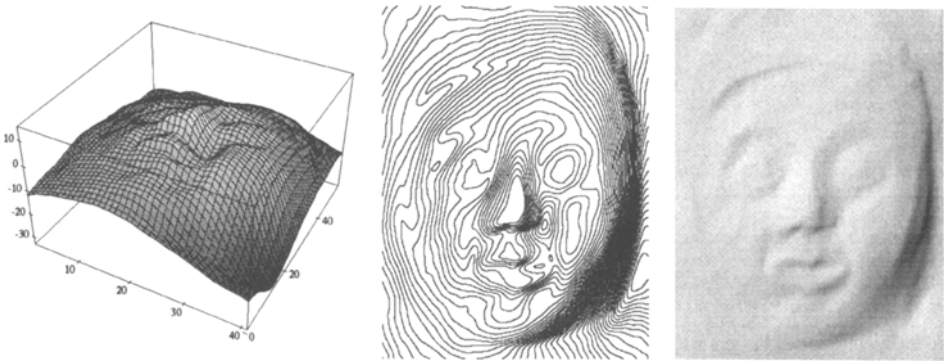

Fig. 6. Depth map corresponding to cherub image, level curves, and an image produced by shading the depth map from $(-1, 1, 1)/\sqrt{3}$.

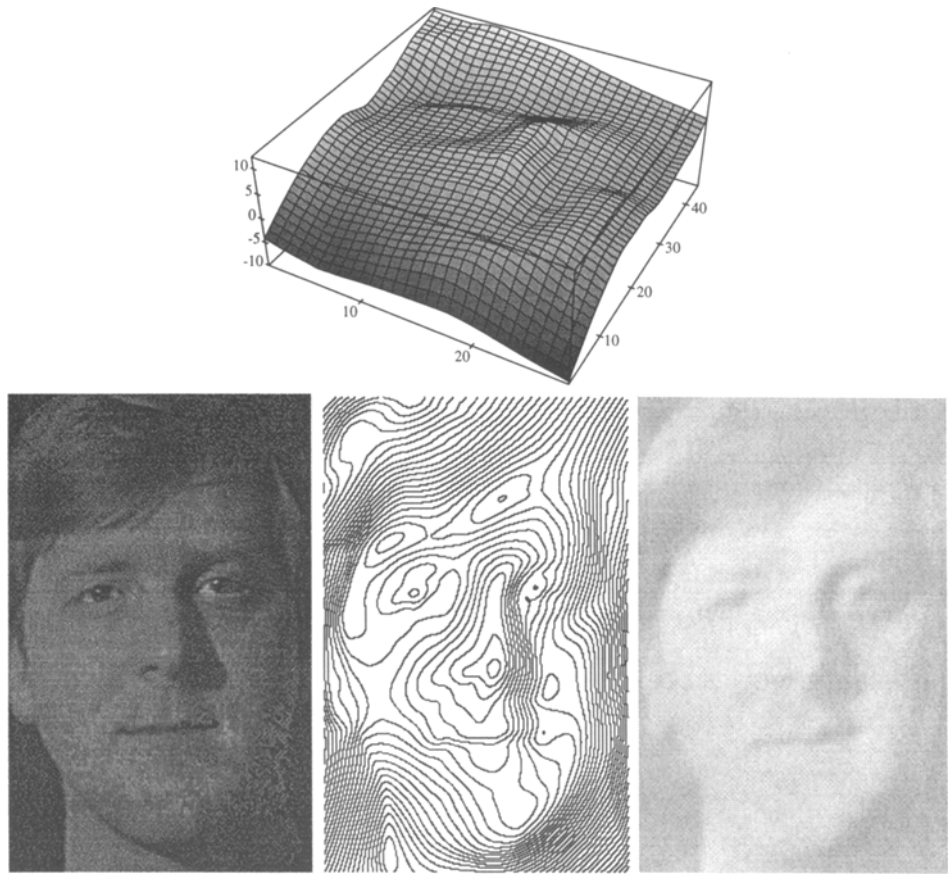

Fig. 7. Image of one author, depth map, level curves, and an image produced by shading the depth map from $(-1, 1, 1)/\sqrt{3}$.