Some Experiments in Applying Inductive Inference Principles to Surface Reconstruction

Edwin P.D. Pednault
AT&T Bell Laboratories
Crawfords Corner Road
Holmdel, New Jersey 07733

Abstract

This paper presents an application of the Minimum Description-Length (MDL) principle of inductive learning to the surface reconstruction problem of computer vision. The application demonstrates that the MDL principle can be applied to practical problems while preserving its convergence properties. It also illustrates how at least one aspect of computer vision (i.e., surface reconstruction) can be treated as an inductive inference problem using this principle. This has the advantage that the convergence properties of the MDL principle enable an exact surface reconstruction to be obtained asymptotically as the number of available data points increases. Moreover, convergence to the true surface occurs independent of the amount of error in the measurements.

1. Introduction

In its simplest form, the Minimum Description-Length (MDL) principle states that induction can be accomplished by finding the shortest description for a set of observations in a suitable language, where the description includes a theory that accounts for the observations. It has been proved mathematically that choosing the shortest description enables one to eventually converge on an "appropriate" theory, given a sufficient number of observations, where appropriateness is judged in terms of the ability to predict observations [Pednault 1988; Barron 1985; Barron and Cover 1983; Rissanen 1978]. This convergence property gives the MDL principle tremendous potential for application to vision and other induction problems, since it guarantees robust inferences for sufficient amounts of data, even in the presence of noise.

The MDL principle, though, is computationally intractable in its most general form. Its convergence properties have been studied with respect to languages that arc Turing-equivalent and thus allow any computable function to be represented. To actually apply the principle in this general context, one would have to solve the halting problem of Turing machines, which is clearly impossible.

However, as I have argued elsewhere [Pednault 1984], the halting problem does not doom the MDL principle, since one can limit the range of theories to a tractable subset, and/or employ approximation techniques that attempt to find theories that are as close to the optimum as is computationally feasible. This pragmatic approach enables one to construct efficient algorithms by tailoring the optimization to specific properties of interest. This leads to the following methodology for applying the MDL principle to computer vision and other induction problems:

(1) Determine the kinds of structures that needs to be detected.
(2) Develop a language well-suited for expressing these structures.
(3) Develop algorithms that find short descriptions in the language.
(4) Run tests to find inappropriate behavior.
(5) Determine whether the problems lie with the language, the algorithms, or both.
(6) Modify the language and/or algorithms accordingly and iterate.

The question, though, is whether the convergence properties of the MDL principle can be preserved when using this methodology.

This paper demonstrates experimentally that convergence can indeed be preserved in practical applications by applying the MDL principle to the surface reconstruction problem in computer vision. Surface reconstruction seeks to recover the mathematical functions that describe a surface given a set of points on the surface le.g., Terzopoulos 1988; Crimson 1983; Barrow and Tenenbaum 1979]. The problem is compounded when dealing with real data, since the points usually will not lie on the actual surface, but instead will be randomly displaced away from it due to measurement errors.

Since the goal of surface reconstruction is to infer general properties (i.e., surfaces) from sets of observations (i.e., data points), it may be treated as a problem of induction. The MDL principle is especially well-suited to this problem, since it guarantees convergence of the inferences for any amount of noise in the measurements. As the number of available data points increases, the reconstructions obtained converge asymptotically to the true surface. This property is demonstrated in the examples presented in Section 4.

Another important characteristic is that if the number of available points is insufficient for exact convergence, an approximation to the true surface is obtained in which the degree of approximation is adjusted automatically according to the number of data points and to the amount of error in the measurements. The effect is to choose coarser approximations when the measurement errors are large and/or the number of points is small, and finer approximations when the errors are small and/or the number of points large. This occurs automatically without the manual adjustment of parameters, such as the stiffness parameters used in thinplate interpolation (see [Tcrzopoulos 1988] for a discussion of such parameters). This property is likewise demonstrated in the examples presented.

2. The Minimum Description-Length Principle

Let us first consider how the MDL principle applies in general to surface reconstruction and then to the specific case

considered in the examples. According to the principle, the theory that best accounts for a collection of observations is the one that yields the shortest description [Pednault 1988; Barron 1985; Barron and Cover 1983; Rissanen 1978]. A description consists of a machine-readable representation of the theory plus an encoding of the observations with respect to the theory. When applied to surface reconstruction, the "theory" is the function that defines the surface and the "observations" are the data points. These points are encoded in terms of the difference between the points and the reconstructed surface. The optimal reconstruction is then the function *S* that minimizes the sum

$$l(S) + l(z_1,\ldots,z_n \mid S)$$

where *l(S)* is the length in bits of a machine-readable description of 5, $z_1,\ldots,z_n$ are the data points, and $l(z_1 \ldots z_n \mid S)$ is the number of bits needed to encode the difference between these points and 5. The combination of a function and an encoding of the difference constitutes an exact representation of the data points. The optimal surface is the one that minimizes the length of this representation.

In general, *S* will be a pieccwise-continuous function that can be decomposed into a collection of regions and continuous functions, with one function per region. *l(S)* will thus incorporate the number of bits needed to specify both the region boundaries and the functions within each region. If these functions form a collection of parametric families, the function associated with a region can be described by first specifying the family it belongs to and then specifying its parameters. The number of bits needed to supply this information is included in *l(S).* Note that *l(S)* will thus increase as the number of regions and function parameters increase. We can therefore view *l(S)* as measuring the complexity of the reconstructed surface.

To encode the data points, the difference between their values and *S* is analyzed statistically by viewing the difference as a random process. This random process, together with the function 5, induces a probability distribution *p* on the data points. If we employ the Shannon coding technique developed in information theory [Gallager 1968], we can use *p* to encode the points using no more than

$$l(z_1,\ldots,z_n \mid S) = -\log_2 p(z_1,\ldots,z_n)$$

bits, $l(z_1\ldots z_n \mid S)$ can thus be thought of as a degree-of-fit term. If the *S* reproduces the data points exactly, $p(z_1\ldots,z_n)$ will equal one and $l(z_u\ldots_t z_n \mid S)$ will be zero. As the fit degrades, $p(z_u\ldots_y z_n)$ will decrease and $l(z_{Xt}\ldots,z_n \mid S)$ will increase.

Since the degree of fit can vary from region to region, we are not justified in assuming that the probability distribution *p* remains constant. To allow it to vary, the distribution must be specified along with the interpolating function within each region. This can be accomplished by treating *p* as a member of a parametric family of distributions and then encoding its parameters. The number of bits needed to specify this information is included in *l(S).*

The optimum surface is the one that achieves the minimal total coding length over all partitionings of the points into regions and all possible assignments of surfaces and noise models to each region.

3. Polynomial Surface Reconstruction

For the experiments reported in this paper, piecewise-polynomial functions were used in conjunction with a Gaussian noise model to describe surfaces and the random displacement of the data points. The algorithms developed actually allow any linear combination of basis functions. Polynomials were used in the experiments because of their familiarity in order to facilitate the presentation of the results. A Gaussian noise model was selected, since optimal polynomial coefficients can then be determined using a computationally-efficient least-squares algorithm. Nonlinear families of interpolating functions can also be used within the framework, as can other noise models, but different techniques must then be employed to find optimal function parameters (the least-squares algorithm presumes a Gaussian noise model and linear combination of basis functions).

A polynomial-time algorithm has been developed to find optimal reconstructions for one dimensional surfaces (i.e., curves). Finding an optimal reconstruction is computationally infeasible for a two dimensional surface, since an exponential number of regions must be examined in order to find the optimum. For a multidimensional surface, one has no choice but to use approximation techniques to find reconstructions that are as close to the optimum as is computationally feasible. One approach I am investigating is to use a series of optimal ID reconstructions to guide the reconstruction of 2D surfaces. Early experiments with this approximation technique have proved quite promising.

With ID surface reconstruction, regions become intervals on a line. Polynomials are used to describe the surface (curve) within each region. To obtain an efficient algorithm, each interval is encoded independently and the resulting codes are then concatenated to form the encoding of the overall surface. This allows the total coding length to be minimized using dynamic programming techniques, which produces a polynomial-time algorithm.

The data points are assumed to take on integer values from 0 to 255. This corresponds to range data quantized to 8 bits accuracy. Because of the discrete nature of the data, the noise model employed is a bounded discrete Gaussian distribution having the probability mass function

$$p(z) = \beta(\delta,\gamma)\, e^{-\frac{(z-\delta)^2}{2\gamma^2}}$$

where z represents the value of a data point, and

$$\beta(\delta,\gamma) = \left[\sum_{k=0}^{255} e^{-\frac{(k-\delta)^2}{2\gamma^2}}\right]^{-1}$$

The parameter 5 represents the true height of the surface at a point, while y^2 determines the distribution of errors when measuring 6. Note that 5 and y^2 only loosely correspond to the mean and variance of the distribution. They arc not numerically equal to these statistics because the distribution is bounded and discrete.

The underlying surface is assumed to be sampled at uniformly spaced increments so that z, represents the value of the i'th sample. The polynomials are then expressed as functions of i. The measurement errors for the samples within an interpolation interval are assumed to be statistically independent. The joint distribution is then the product of the marginal distributions for the points in the interval. The parameter y^2 is a assumed to be constant over an interval, while 5 is given by the interpolating polynomial. Thus, if $a_0+\ldots\ldots a_m i^m$ is the polynomial associated with the interval containing points i_1 through j_i, the joint distribution for these points is given by

$$p(z_{i_1},\ldots,z_{j_1}) = \prod_{k=i_1}^{j_1} \beta(a_0 + \cdots + a_m k^m, \gamma)\, e^{-\frac{(z_k - a_0 - \cdots - a_m k^m)^2}{2\gamma^2}}$$

The number of bits needed to encode points i_1 through j_1 relative to the polynomial is given by the negative log of this expression:

$$l(z_{i_1},\ldots,z_{j_1} \mid S_{i_1 j_1}) = -\log_2 p(z_{i_1},\ldots,z_{j_1})$$

$S_{i_1 j_1}$ represents an encoding of the polynomial function, the number of points in the interval, and the noise parameter γ^2. An upper bound on $l(z_{i_1},\ldots,z_{j_1} \mid S_{i_1 j_1})$ can be obtained by replacing the $\beta(\cdots,\gamma)$ term by $[\min_\delta \beta(\delta,\gamma)]$. Minimizing this upper bound with respect to the polynomial coefficients reduces to a least-squares fit problem and produces near-optimal coefficients. Minimization with respect to γ^2 can be accomplished via a Newton-Raphson method.

The total coding length is the sum of the coding lengths for each of the intervals. Thus, if $i_1,\ldots,i_t$ are the starting points of the intervals and $j_1,\ldots,j_t$ are the ending points, where $i_1 = 1$, $j_t = n$, and $i_{k+1} = 1 + j_k$ for $1 \le k < n$, then the total coding length is given by

$$l(S) + l(z_1,\ldots,z_n \mid S) = \sum_{k=1}^{t} l(S_{i_k j_k}) + l(z_{i_k},\ldots,z_{j_k} \mid S_{i_k j_k})$$

The number of intervals and their endpoints are determined by minimizing this sum using dynamic programming techniques.

The quantity $l(S_{i_k j_k})$ is equal to the number of bits needed to encode the width of the interval (i.e., $j_k - i_k + 1$), the order of the polynomial, the polynomial coefficients, and the noise parameter γ^2. The interval width and the order of the polynomial are integers and are encoded using the Rissanen-Cover-Elias universal prior for integers [Rissanen 1983; Leung-Yan-Cheong and Cover 1978; Elias 1975]. If m is the order of the polynomial, the number of bits needed to encode m and the interval width is given by $\log_2^*(m+1) + c$ and $\log_2^*(j_k - i_k + 1) + c$, respectively, where $c \approx 2.865$ and where $\log_2^*(x) = \log_2(x) + \log_2\log_2(x) + \cdots$ up to but not including the first negative term.

Whereas the interval width and the order of the polynomial are integers, γ^2 and the polynomial coefficients are real-valued and must therefore be quantized. Rissanen [1983] has proposed a quantization scheme that would minimize the worst-case coding length assuming uniform quantization steps. However, this scheme was found to produce somewhat inefficient codes in the experiments reported here, mainly because the data points are already quantized to 8 bits. A nonuniform quantization scheme was therefore developed instead to achieve more compact encodings.

4. Experimental Results

Figures 1 and 2 illustrate the performance of the reconstruction algorithm using synthetic data. In these figures, Gaussian white noise of various standard deviations is added to two different piecewise polynomial curves. The noisy curves and their resulting reconstructions are plotted side by side for comparison. The underlying curve in each figure consists of three polynomial regions of orders 0, 1, and 2, respectively. The curve in Figure 1 is continuous, while in Figure 2 it is discontinuous. In Figures 1a and 2a, the curves are sampled at 48 points and the standard deviation of the noise is increased from 0 to 32 in a logarithmic fashion. In Figures 1b and 2b, the standard deviation of the noise is held constant at 32 and the number of data points is increased from 48 to 768 by factors of two in Figure 1b, and from 48 to 384 in Figure 2b.

Figure 3 presents the results of applying the algorithm to the rows of a 128 by 128 range image of a mechanical part. Figure 3a shows the original image, while Figure 3b shows the reconstruction. In Figure 3c, a row and a column from the original image are plotted along side their reconstructions. The positions of the row and column are indicated by horizontal and vertical marks along the borders of Figures 3a and 3b.

In all of these examples, the number of intervals, their locations, and the order of the polynomials were determined entirely by minimizing the total encoding length of the data points. The initial implementation allowed up to 15th order polynomials. Problems with numerical instability, however, required that the polynomials be limited to at most 5th order in the examples presented. Further work is being done to resolve these stability problems.

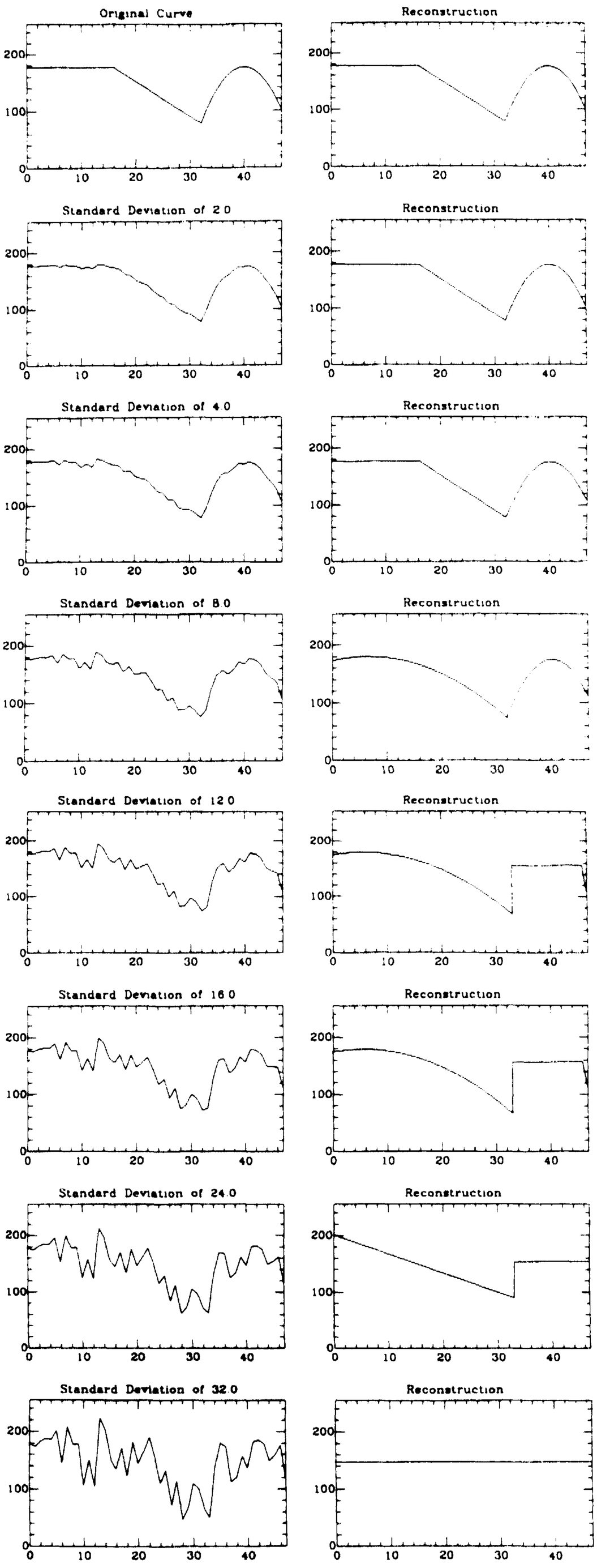
5. Discussion

The experimental results demonstrate that the convergence properties of the MDL principle can be preserved in practical applications. Figures 1b and 2b clearly show that an exact surface reconstruction is obtained asymptotically when the true surface is a member of the subset considered (i.e., polynomials) and one is minimizing the description length over this subset. In both cases, the structures of the underlying curves are recovered at the highest data samplings. The use of a high noise level illustrates that convergence occurs independent of the amount of noise. The noise level does affect the rate of convergence (i.e., the higher the noise, the slower the convergence), but not the eventual convergence. Note that convergence is much faster for the discontinuous curve.

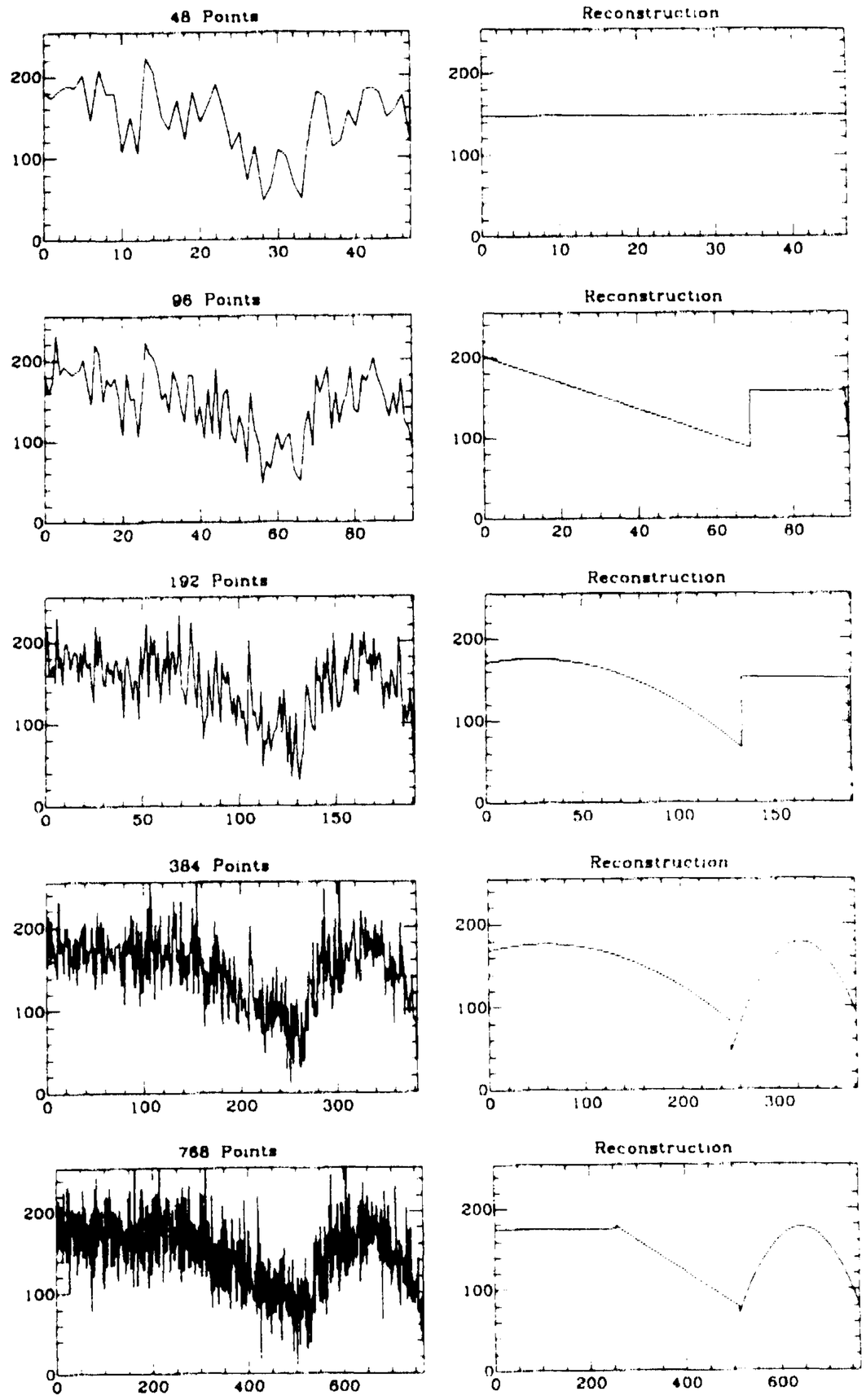
Figures 1 and 2 together demonstrate that the degree to which the reconstruction approximates the true surface adjusts automatically to the number of data points and to the amount of error in the measurements when the number of points is insufficient for exact reconstruction. Coarser approximations are made when the measurement errors are large and/or the number of points is small, and finer approximations are made when the errors are small and/or the number of points large. The result is a graceful decrease in the accuracy of the reconstruction as measurement error increases or the number of data points decreases. This compensation occurs automatically without the manual adjustment of parameters. Note that the decrease in accuracy is more gradual for the discontinuous curve.

The mechanical part shown in Figure 3 does not have polynomial surfaces, yet the reconstruction produced by the algorithm is quite accurate. Thus, even when the true surface lies outside of the subset considered, a good approximation to the surface can still be obtained. The principal requirement is for the subset to have sufficient latitude for an adequate approximation. In addition, the noise model must adequately reflect the statistical properties of the measurement errors, which does happen to be the case for the mechanical part.

The surfaces in Figure 3 are actually two dimensional, not one dimensional as is assumed by the algorithm. However, finding a minimal encoding of a 2D surface is computationally infeasible, since an exponential number of

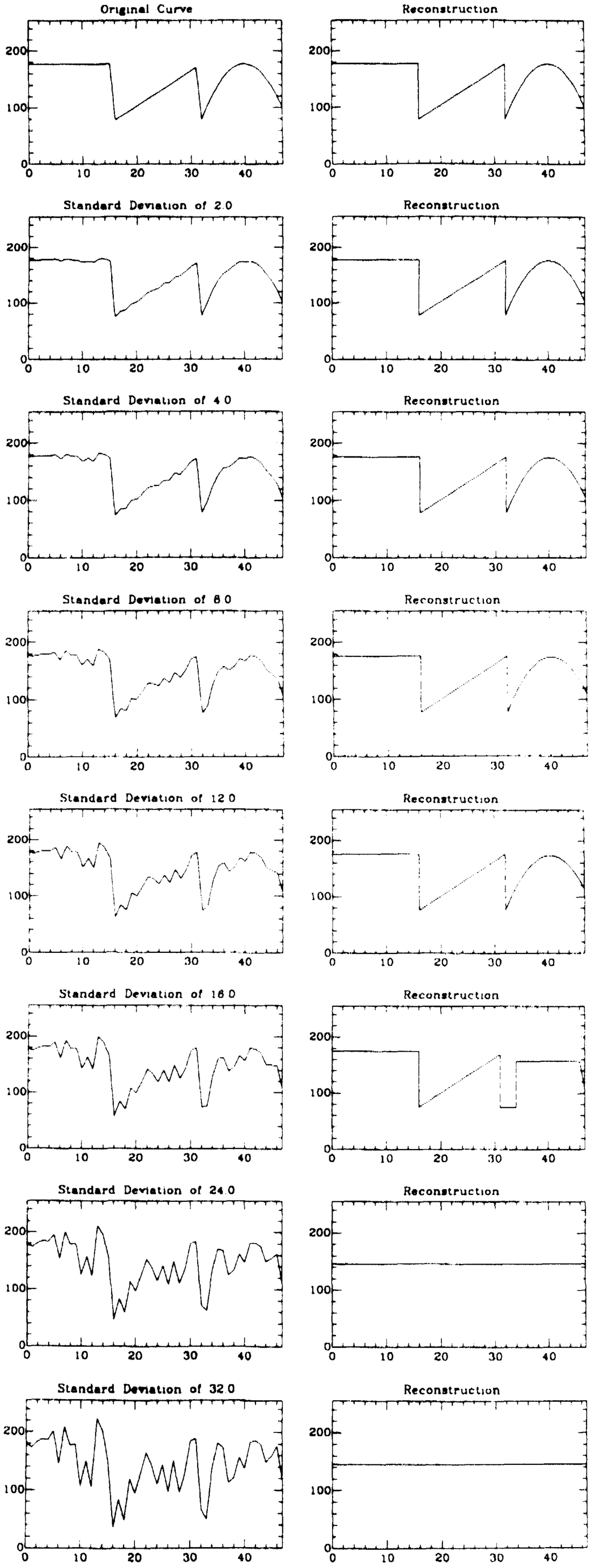


(A)



(B)

Figure 1: Result of applying the algorithm to a continuous piecewise polynomial curve with additive Gaussian noise. (A) Reconstructions obtained when noise of increasing standard deviations is added to 48 sample points on the curve. (B) Reconstructions obtained when the standard deviation is held constant at 32 and the number of sample points is increased from 48 to 768.



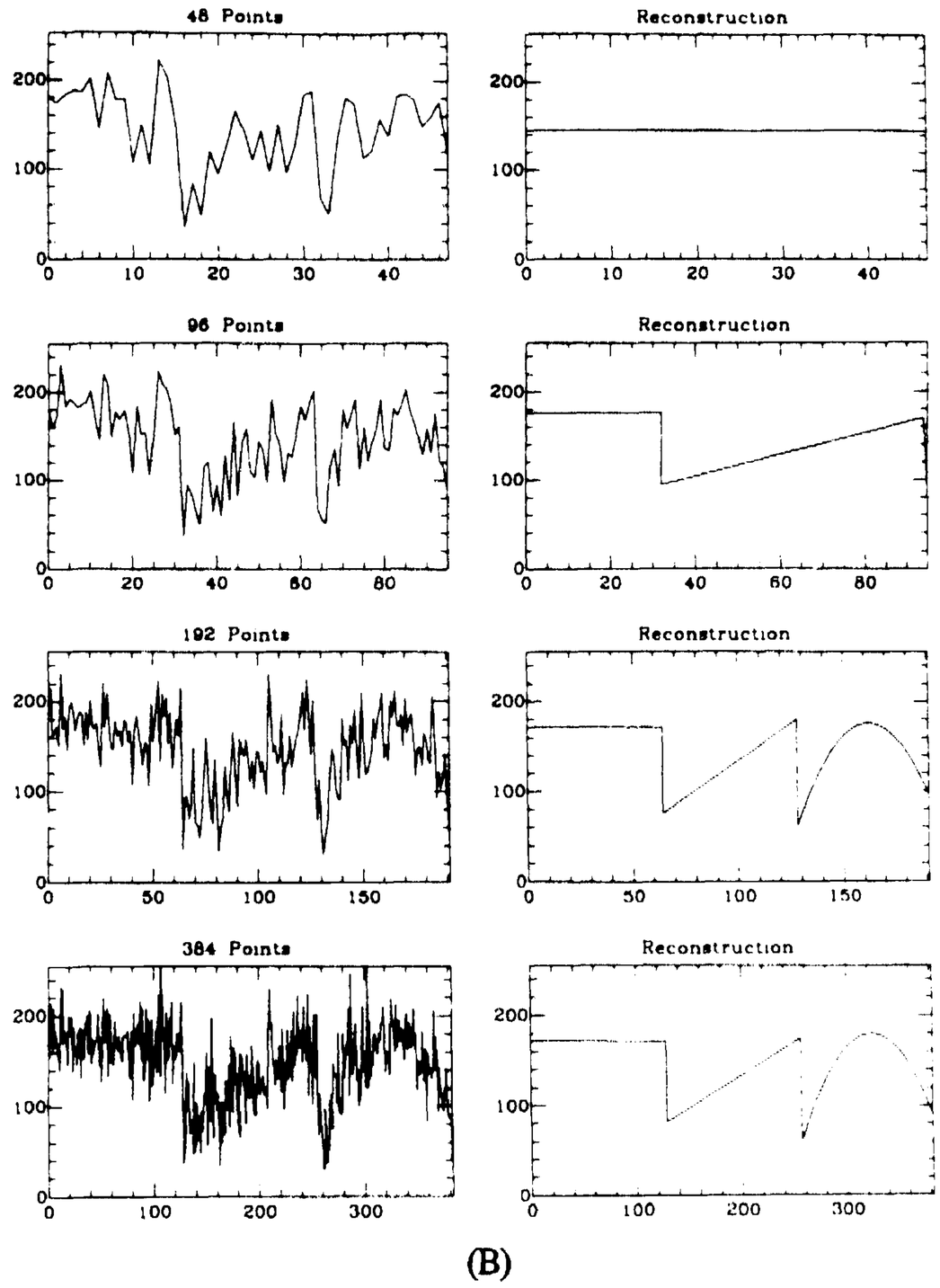
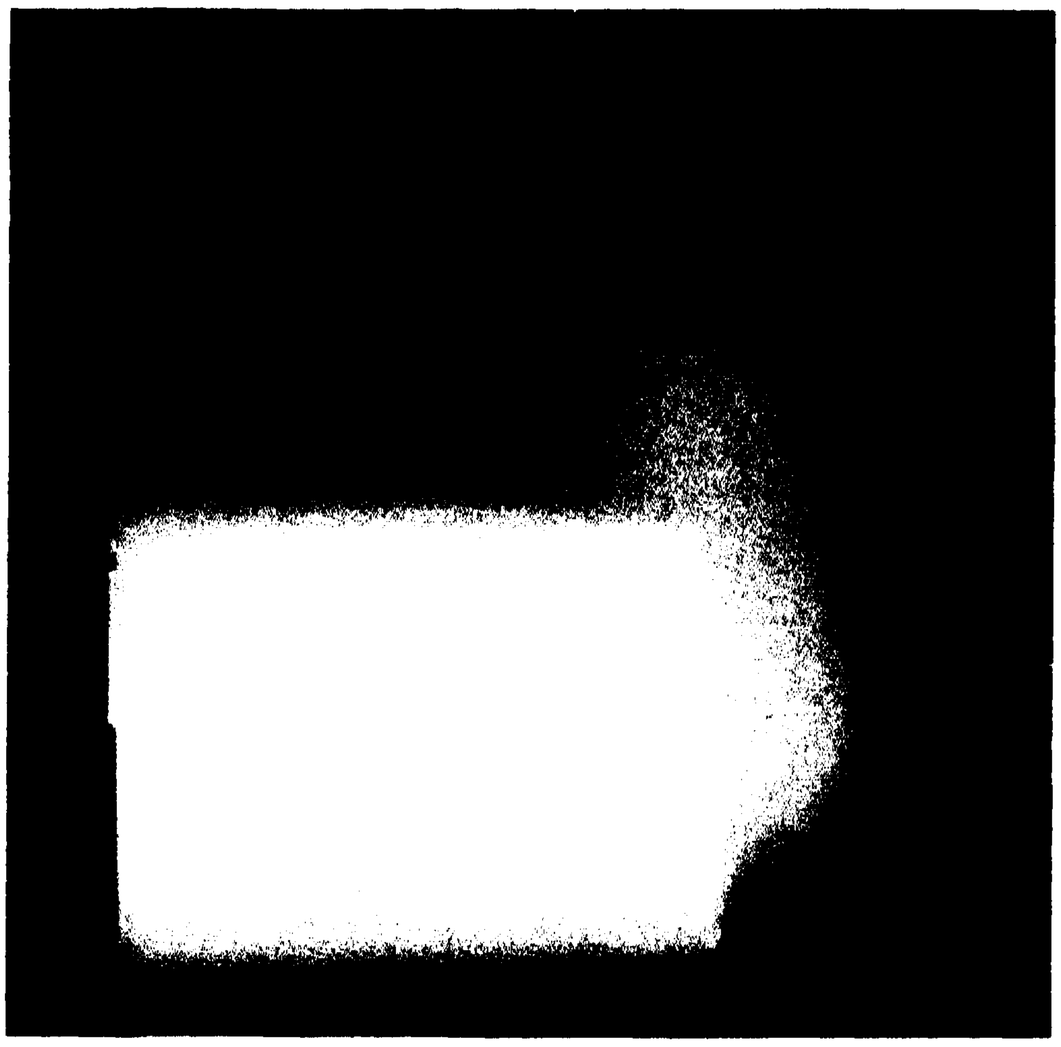
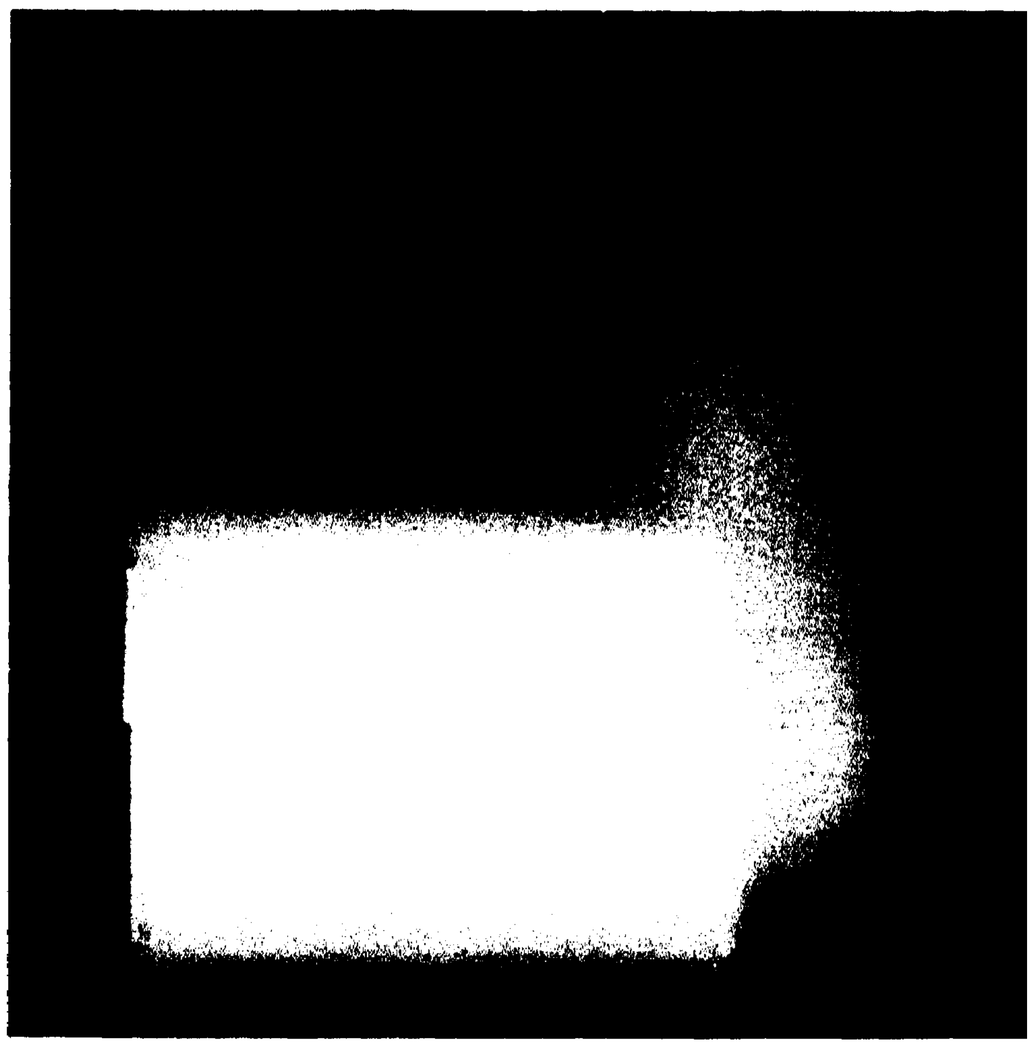


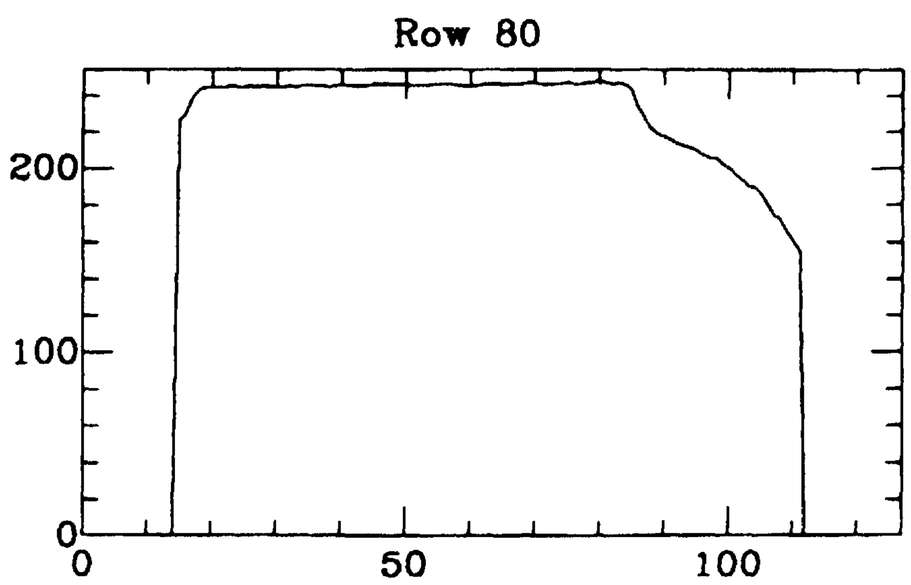
Figure 2: Result of applying the algorithm to a discontinuous piecewise polynomial curve with additive Gaussian noise. (A) Reconstructions obtained when noise of increasing standard deviations is added to 48 sample points on the curve. (B) Reconstructions obtained when the standard deviation is held constant at 32 and the number of sample points is increased from 48 to 384.

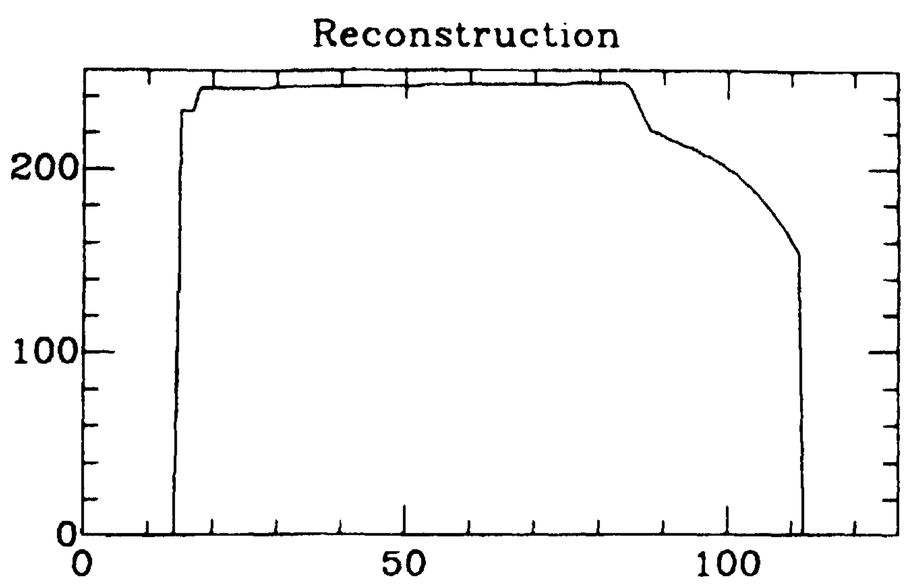


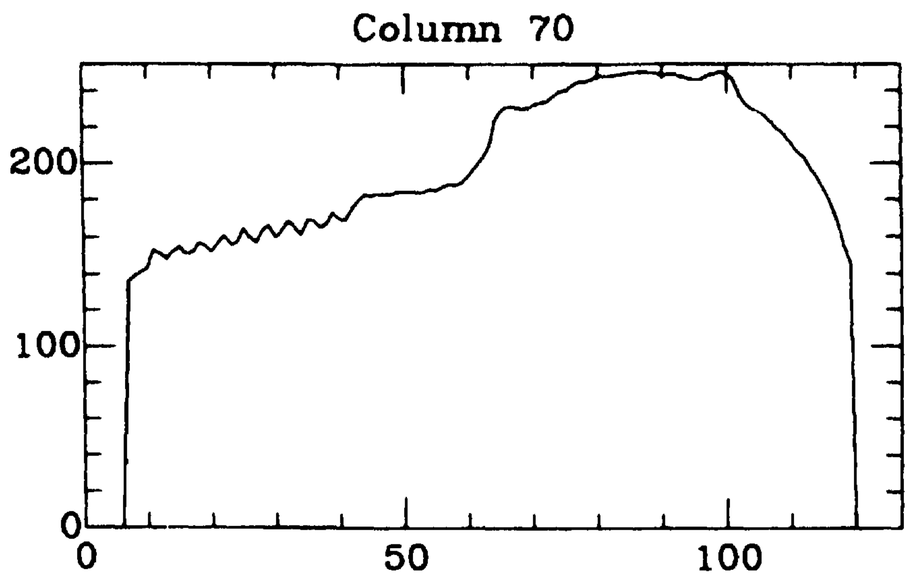
(A)

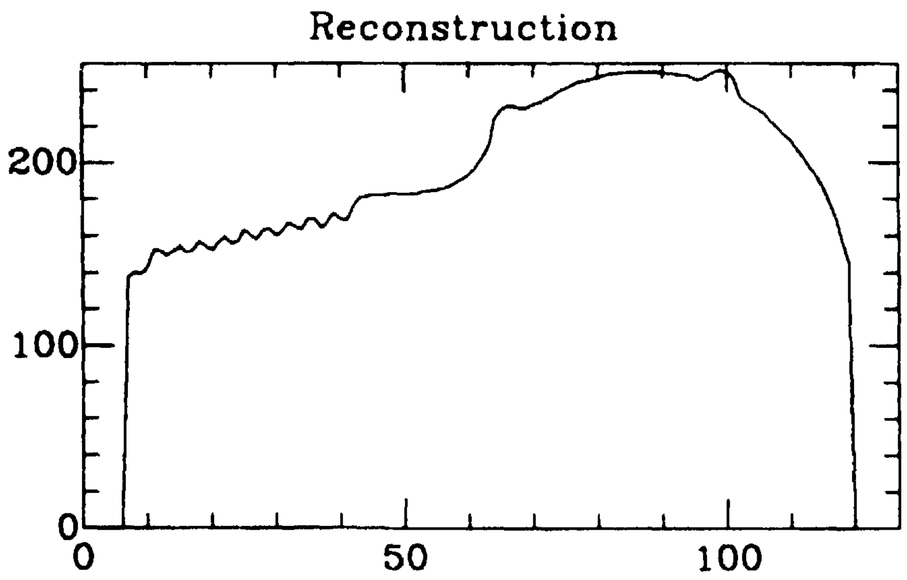


(B)









(C)

Figure 3: Result of applying the algorithm to the rows of a 128 by 128 range image of a mechanical part. (A) The original image. (B) The row reconstruction. (C) A row and column selected from the original image together with their reconstructions. The position of the row and column are indicated by marks along the borders of the images.

regions would have be examined to find the optimum. I am investigating an approach to 2D reconstruction that utilizes a series of ID reconstructions determined through exact minimal encodings. The intuition is that high-quality ID reconstructions based on polynomials of sufficient degree will provide tight constraints on the 2D surface. The quality of the reconstructions obtained on the mechanical part strongly supports the viability of this approach. It also provides evidence that one need not necessarily find the absolute shortest description in order to obtain a good approximation to the true surface—a description that is short enough can do just as well.

6. Conclusions

The strength of the MDL principle lies in its convergence properties [Pednault 1988; Barron 1985; Barron and Cover 1983; Rissanen 1978]. The benefits of these properties provide strong motivation for discovering ways of overcoming the computational barriers posed by the principle in its most general form. The results presented here demonstrate that these barriers can be overcome in specific applications by limiting the range of admissible theories to a tractable subset and/or by employing approximation techniques that attempt to find theories that are as close to the optimum as is computationally feasible.

The approach that I am exploring in applying the MDL principle to surface reconstruction illustrates one way of dealing with these computational issues. Different means are also being explored by other researchers in computer vision. Segen [1980, 1985, 1988] has applied the principle to several problems in vision and perception using various approximation techniques. Smith and Wolf [1984] have considered an approach closely related to the one presented here in applying the MDL principle to curve matching for image correspondence. Leclerc [1988] is investigating a distributed 2D approximation to the MDL principle for the purpose of image partitioning. Pentland [1988] has developed an algorithm motivated by the MDL principle for the reconstruction of superquadrics from binary images. Though different approaches to constructing approximation algorithms are taken in each instance, they all fall within the basic paradigm outlined above and in Section 1.

This previous work focused primarily on the practical application of the MDL principle without considering its convergence properties. The results presented in this paper demonstrate that convergence is one of the key aspects of the MDL principle. Since convergence is not guaranteed when approximation techniques are employed, it should be included as one of the tests for inappropriate behavior when following the methodology described in Section 1. The variety of applications previously considered, together with the convergence properties examined here, demonstrate the flexibility and viability of the MDL principle. Undoubtedly, the future will see many more applications of this principle.

Acknowledgements

I wish to thank Steve Zucker for introducing me to computer vision years ago, Tom Cover for having introduced me to Kolmogorov complexity and the minimal description-length principle, and Steve, Tom, Peter Cheeseman, Mike Georgeff, Yvan Leclerc, John Mohammed, Jakub Segen, and many others for the stimulating and challenging (if not heated) discussions on these topics over the years. John Gabbe and Larry O'Gorman provided helpful comments on earlier drafts. Again, I thank my faithful editors Corinne and Maria Babcock. Gerard Medioni at the University of Southern California kindly provided the range image that appears in Figure 3.

References

[Barron and Cover 1983] A. R. Barron and T. M. Cover, "convergence of logically simple estimates of unknown probability densities," presented at the *International Symp. on Info. Theory,* St. Jovite, Quebec, Canada, 1983.

[Barron 1985] A. R. Barron, "Logically smooth density estimation," Tech. Report 56, Dept. of Statistics, Stanford University, Stanford, California, 1985.

[Barrow and Tenenbaum 1979] H. G. Barrow and J. M. Tenenbaum, "Reconstructing smooth surfaces from partial, noisy information," *Proc. DARPA Image Understanding Workshop,* University of Southern California, pp 76-86, 1979.

[Elias 1975] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans, on Info. Theory,* Vol. 21, No. 2, pp 194-203, 1975.

[Gallager 1968] R. G. Gallager, *Information Theory and Reliable Communication,* John Wiley and Sons, New York, 1968.

[Grimson 1983] W. E. L. Grimson, "An implementation of a computational theory of visual surface interpolation," *Comp. Vision, Graphics, and Image Proc,* Vol 22, pp 39-69, 1983.

[Leclerc 1988] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," *Proc. DARPA Image Understanding Workshop,* pp 365-382, April 1988.

[Leung-Yan-Cheong and Cover 1978] S. K. Leung-Yan-Cheong and T. M. Cover, "Some equivalences between Shannon entropy and Kolmogorov complexity," *IEEE Trans, on Info. Theory,* Vol. 24, No. 3, pp 331-338, 1978.

[Pednault 1984] E. P. D. Pednault, "Vision, induction, and minimum-length descriptions," talk presented at the AI Center, SRI International, Menlo Park, California, March 1984.

[Pednault 1988] E. P. D. Pednault, "Inferring probabilistic theories from data," *Proc. AAA1-88,* St. Paul, Minn., pp 624-628.

[Pentland 1988] A. Pentland, "Automatic extraction of deformable part models," Tech. Report 104, Vision Sciences, MIT Media Lab, Cambridge, Mass., 1988.

[Rissanen 1978] J. Rissanen, "Modeling by shortest data description," *Automatica,* Vol. 14, pp 465-471, 1978.

[Rissanen 1983] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics,* Vol. 11, pp 416-431, 1983.

[Segen 1980] J. Segen, *Pattern Directed Signal Analysis,* Ph.D. Thesis, Dept. of Electrical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, May 1979.

[Segen 1985] J. Segen, "Learning structural descriptions of shape," *Proc. Comp. Vision and Patt. Rec,* pp 69-99, 1985.

[Segen 1988] J. Segen, "From features to symbols: learning relational models of shape," *Proc. COST 13 Workshop,* Bonas, France, August 1988.

[Smith and Wolf 1984] G. B. Smith and H. C. Wolf, "Image-to-image correspondence: linear structure matching", Tech Note 331, AI Center, SRI International, Menlo Park, California, July 1984.

[Terzopoulos 1988] D. Terzopoulos, "The computation of visible-surface representations," *IEEE Trans, on Pattern Anal, and Machine IntelL,* Vol. 10, No. 4, July 1988.