Closed-Form Solutions for Physically-Based Shape Modeling and Recognition

Stan Sclaroff
Alex Pentland

Vision and Modeling Group, The Media Lab
Massachusetts Institute of Technology

Abstract
We present an efficient, physically-based solution for recovering a 3-D solid model from collections of 3-D surface measurements. Given a sufficient number of independent measurements, the solution is overconstrained and unique except for rotational symmetries. We then present a physically-based object recognition method that allows simple, closed-form comparisons of recovered 3-D solid models. The performance of these methods is evaluated using both synthetic and real laser range finder data.

1 Introduction
Vision research has a long tradition of going from collections of low-level measurements to higher-level "part" descriptions such as generalized cylinders [1; 2], deformed superquadrics [3; 4; 5; 6], or geons [7], and then attempting to perform object recognition. The general idea is to use part-level modeling primitives to bridge the gap between image features and the symbolic, parts-and-connections descriptions useful for recognition and reasoning.

Recently, several researchers have successfully addressed the first of these problems — that of recovering part descriptions — using deformable models. There have been two classes of deformable models: those based on parametric solid modeling primitives, such as our own work using superquadrics [3], and those based on mesh-like surface models, such as employed by Terzopoulos, Witkin, and Kass [8]. With parametric models, fitting has been performed using the modeling primitive's "inside-outside" function [4; 5; 9; 10], while with mesh surface models, an energy function has been employed [6; 8].

Shape description by use of orthogonal parameters has the advantage that it produces a unique, compact description that is well-suited for recognition and database search, but has the disadvantage that it does not have enough degrees of freedom to account for fine surface details. In contrast, deformable meshes are very good for describing shape details, but produce descriptions that are neither unique nor compact, and consequently cannot be used for recognition or database search without additional processing. Both approaches share the disadvantage that they are relatively slow, requiring dozens of iterations in the case of parametric formulation, and up to hundreds of iterations in the case of physically-based mesh formulation.

We have addressed these problems by adopting an approach based on the finite element method (FEM) and solid modeling. This approach provides both the expressiveness and convenience of the physically-based mesh formulation, and in addition can provide greater accuracy in physical simulation. Thus it is possible to use the models to recover from range data to accurately simulate particular materials and situations for purposes of prediction, visualization, planning, and so forth [10].

More importantly, we have developed a formulation whose degrees of freedom are orthogonal, and thus decoupled, by posing the dynamic equations in terms of the FEM equations' eigenvectors. These eigenvectors are known as the object's free vibration modes, and together form a frequency-ordered orthonormal basis set.

By decoupling the degrees of freedom we achieve substantial advantages — the most important of these being that the fitting problem has a simple, efficient, closed-form solution. In addition, the model's intrinsic complexity can be adjusted to match the number of degrees of freedom in the data measurements, so that the solution can always be made overconstrained. When overconstrained, the solution is unique, except for rotational symmetries and degenerate conditions. Thus the solution is well-suited for recognition and database tasks.

2 Background: The Representation
Our representation describes objects using the force-and-process metaphor of modeling clay: shape is thought of as the result of pushing, pinching, and pulling on a lump of elastic material [3; 10]. The mathematical formulation is based on the finite element method (FEM), the standard engineering technique for simulating the dynamic behavior of an object.

In the FEM, energy functionals are formulated in terms of nodal displacements U, and iterated to solve for the nodal displacements as a function of impinging loads R:

\[ \mathbf{MU} + \mathbf{CU} + \mathbf{KU} = \mathbf{R} \] (1)

This equation is known as the FEM governing equation. The solution of the equilibrium equation for the nodal displacements U is the most common objective of finite element analyses.

2.1 Modal Analysis
To obtain an equilibrium solution U, one integrates Equation 1 using an iterative numerical procedure at a cost proportional to the stiffness matrices' bandwidth. To reduce this cost we can transform the problem from the original nodal coordinate system to a new coordinate system whose basis vectors are the columns of an n x n matrix P. In this...
new coordinate system the nodal displacements $U$ become 
*generalized displacements* $\bar{U}$:

$$U = \Phi \bar{U}$$ (3)

Substituting Equation 3 into Equation 1 and premultiplying by $\Phi^T$ transforms the governing equation into the co-
ordinate system defined by the basis $\Phi$:

$$\bar{\Phi} \bar{M} \bar{\Phi} \bar{U} + \bar{\Phi} \bar{C} \bar{U} + \bar{\Phi} \bar{K} \bar{U} = \bar{\Phi} \bar{R}$$ (4)

where

$$\bar{M} = \Phi^T M \Phi, \quad \bar{C} = \Phi^T C \Phi, \quad \bar{K} = \Phi^T K \Phi, \quad \bar{R} = \Phi^T R$$ (5)

With this transformation of basis, a new system of stiffness, mass, and damping matrices can be obtained which has a
smaller bandwidth than the original system.

The optimal basis $\Phi$ has columns that are the eigenvectors of $M^{-1}K$ [11]. These eigenvectors are also known as
the system’s *free vibration modes*. Using this transformation matrix we have

$$\Phi^T K \Phi = \Omega^2, \quad \Phi^T M \Phi = I$$ (6)

where the diagonal elements of $\Omega^2$ are the eigenvalues of $M^{-1}K$ and remaining elements are zero. When the damping
matrix $C$ is restricted to *Rayleigh damping*, then it is also diagonalized by this transformation.

The lowest frequency modes are always the rigid-body
modes of translation and rotation. The next-lowest frequency
modes are smooth, whole-body deformations that leave the center of mass and rotation fixed. Compact bod-
ies — solid objects like cylinders, boxes, or heads, whose
dimensions are within the same order of magnitude — nor-
mally have low-order modes which are intuitive to humans: bending, pinching, tapering, scaling, twisting, and shearing.
Bodies with very dissimilar dimensions, or which have holes, etc., can have very complex low-frequency modes.

3 Recovering 3-D Part Models From
3-D Sensor Measurements

Let us assume that we are given $n$ three-dimensional sensor
measurements $X^w$ (in the global coordinate system) that
originate from the surface of a single object

$$X^w = [x_1^w, y_1^w, z_1^w, \ldots, x_n^w, y_n^w, z_n^w]^T$$ (7)

Following Terzopoulos et al., we then attach virtual springs between these sensor measurement points and particular
nodes on our deformable model. This defines an equili-
rium equation whose solution $\bar{U}$ is the desired fit to the
sensor data. Consequently, for $m$ nodes with correspond-
ing sensor measurements, we can calculate the virtual loads $\bar{R}$ exerted on the undeformed object while fitting it to the sensor measurements. For node $k$ these loads are simply

$$[r_{2k-1}, r_{2k}, r_{2k+1}, r_{2k+2}]^T = [x_k^w, y_k^w, z_k^w]^T - [x_k, y_k, z_k]^T$$ (8)

where $x_1, y_1, z_1, \ldots, x_n, y_n, z_n$ are the nodal coordinates of the undeformed object in the object’s coordinate frame.

When sensor measurements do not correspond exactly with
existing nodes, the loads can be distributed to surrounding
nodes using the FEM interpolation functions.

To fit a deformable solid to the measured data we solve
the equilibrium governing equation $K\bar{U} = \bar{R}$. The diffi-
culty in calculating this solution is the large dimensionality
of $K$, so that iterative solution techniques are normally em-
ployed. However, a closed-form solution is available sim-
ply by converting this equation to the modal coordinate
system. This is accomplished by substituting $U = \Phi \bar{U}$ and premultiplying by $\Phi^T$, so that the equilibrium equa-
tion becomes

$$\Phi^T K \Phi \bar{U} = \Phi^T \bar{R}$$ (9)

The solution to the fitting problem, therefore, is obtained
by inverting the diagonal matrix $K$. Note, however, that as
this formulation is posed in the object’s coordinate
system the rigid body modes have zero eigenvalues, and
must therefore be solved for separately by setting $\bar{u}_i = \bar{r}_i$, $1 \leq i \leq 6$. The complete solution may be written in the
original nodal coordinate system, as follows

$$U = \Phi(K + I)^{-1}\Phi^T \bar{R}$$ (10)

where $I$ is a matrix whose first six diagonal elements are
ones, and remaining elements are zero.

3.1 Reducing the Degrees of Freedom

The major difficulty in calculating this solution occurs
when there are fewer degrees of freedom in sensor measure-
ments than in the nodal positions — as is normally the case in
computer vision applications. Previous researchers have
suggested adopting heuristics such as smoothness and sym-
metry to obtain a well-behaved solution; however in many
cases the observed objects are neither smooth nor symmetric,
and so an alternative method is desirable.

We believe that a better, and certainly simpler, method
is to discard some of the high-frequency modes, so that
the number of degrees of freedom in $\bar{U}$ is equal to or
less than the number of degrees of freedom in the sen-
or measurements. Discarding some of the high-frequency
modes allows Equation 10 to provide a generically over-
constrained estimate of object shape. Note that discard-
ing high-frequency modes is not equivalent to a smooth-
ness assumption, as sharp corners, creases, etc., can still
be obtained. What we cannot do with a reduced-basis
modal representation is place many creases or spikes close
together.

3.2 Determining Spring Attachment

Attaching a virtual spring between a data measurement and a deformable object implicitly specifies a correspon-
dence between some of the model’s nodes and the sensor
measurements. In most situations this correspondence is
not given, and so must be determined in parallel with the
equilibrium solution. In our experience this attachment
problem is the most problematic aspect of the physically-
based modeling paradigm. This is not surprising, however,
as the attachment problem is similar to the correspondence
problems found in many other vision applications.

Our approach to the spring attachment problem is sim-
ilar to that adopted by other researchers [4; 8]. Given a
segmentation of the data into objects or “blobs,” the
first step is to define an ellipsoidal coordinate system by
examination of the data’s moments of inertia. These esti-
mates of position, orientation, and size define an elliptical
coordinate system. Data points are then projected onto

---

1 Closed-form in the sense that the solution is non-
iterative, requiring only multiplying a data term by a pre-
computed matrix of eigenvectors.
Figure 1: Front (a) and side (b) views of fitting a rectangular box, cylinder, banana, etc., using sparse 3-D point data with 5% noise. The original models are shown in white, and the recovered models are shown in darker grey.

3.3 A Synthetic Example

Figure 1 shows an example using very sparse synthetic range information. White objects are the original shapes that the range data was drawn from, and the grey objects are the recovered 3-D models. For each white object in Figure 1(a) the position of visible corners, edges, and face centers was measured, resulting in between 11 and 18 data points. These data points were then corrupted by adding 5 mm of uniformly distributed noise to each data point's x, y, and z coordinates; the maximum dimension of each object is approximately 100 mm. Note that only the front-facing 3-D points, e.g., those visible in Figure 1(a) at the left, were used in this example. Total execution time on a standard Sun 4/330 is approximately 0.1 seconds.

Despite the rather large amount of noise and a complete lack of information about the back side of the object, it can be seen that Equation 10 does a good job of recovering the entire 3-D model. This is especially apparent in the side view, shown in Figure 1(b), where we can see that even the back side of the recovered models (grey) are very similar to the originals (white). This accuracy despite lack of 360° data reflects the fact that Equation 10 provides the shape estimate with the least internal strain energy, so that symmetric and mirror symmetric solutions are preferred.

3.4 Fitting Range Data

A second example uses 360° laser rangefinder data of a human head, as shown in the left-hand image of Figure 2. There are about 2500 data points. Equation 10 was then used to estimate the shape, using only the low-frequency 30 modes. The low-order recovered model is shown in the middle column; because of the large number of data points execution time on a Sun 4/330 was approximately 3 seconds. It can be seen that the low-order modes provide a sort of qualitative description of the overall head shape.

A full-dimensionality recovered model is shown in the right-hand image of 2. In the ThingWorld system [12], rather than describing high-frequency surface details using a finite element model with as many degrees of freedom as there are data points, we normally augment a low-order finite element model with a spline description of the surface details. This spline description can, of course, be similar to that used by Terzopoulos et al. This provides us with a two-layered representation (low-order finite element model + surface detail spline description = final model) that we find to be both more efficient to recover and more useful in recognition, simulation, and visualization tasks than a fully-detailed finite element model.

4 Object Recognition

Perhaps the major drawback of physically-based models has been that they are not very useful for recognition, comparison, or other database tasks. This is primarily because they normally have more degrees of freedom than there are sensor measurements, so that the recovery process is underconstrained. Therefore, although heuristics such as smoothness or symmetry can be used to obtain a solution, they do not produce a stable, unique solution.

The major problem is that when the model has more degrees of freedom than the data, the model's nodes can slip about on the surface. The result is that there are an infinite number of valid combinations of nodal positions
for any particular surface. This difficulty is common to all spline and piecewise polynomial representations, and is known as the knot problem.

The obvious solution to the problem of non-uniqueness is to discard enough of the high-frequency modes to ensure an overconstrained estimate of shape, as was done for the shape recovery problem above. Use of a reduced-basis modal representation results in a unique description of shape because the modes (eigenvectors) form an orthonormal basis set. Therefore, there is only one way to represent an object in its canonical position.

Further, because the modal representation is frequency-ordered, it has stability properties that are similar to those of a Fourier decomposition. Just as with the Fourier decomposition, an exact subsampling of the data points points does not change the low-frequency modes. Similarly, irregularities in local sampling and measurement noise tend to primarily affect the high-frequency modes, leaving the low-frequency modes relatively unchanged.

4.1 Comparing Models

Thus by employing a reduced-basis modal representation we can obtain overconstrained shape estimates that are also unique except for rotational symmetries. To compare two objects with known mode values $\tilde{U}_1$ and $\tilde{U}_2$, one simply compares the two vectors of mode values:

$$\epsilon = \frac{\tilde{U}_1 \cdot \tilde{U}_2}{\|\tilde{U}_1\| \|\tilde{U}_2\|}$$  \hspace{1cm} (11)

Vector norms other than the dot product can also be employed; in our experience all give roughly the same recognition accuracy.

The first six entries of $\tilde{U}$ are the rigid-body modes (translation and rotation), which are normally irrelevant in object recognition. Similarly, the seventh mode (overall volume) is sometimes irrelevant in object recognition, as many machine vision techniques recover shape only up to an overall scale factor. Thus, rather than computing the dot product with all of the modes, we typically use a modified version of Equation 11 where only modes numbered eight and higher are included; this allows for translation, rotation, and scale-invariant matching.

2 The primary limitation of this uniqueness property stems from the finite element method's linearization of rotation, so that object symmetries can lead to multiple descriptions, and errors in measuring object orientation can cause commensurate errors in shape description.

The ability to compare the shapes of even complex objects by a simple dot product makes the modal representation well suited to recognition, comparison, and other database tasks. We will now evaluate the reliability of the combined shape recovery/recognition process using both synthetic and laser rangefinder data.

4.2 Head Recognition from Range

We conducted an experiment to recover and recognize face models from range data generated by a laser range finder. In this experiment we obtained laser rangefinder data of eight people's heads from a five different viewing directions: the right side (-90°), halfway between right and front (-45°), front (0°), halfway between front and left (45°), and the left side (90°). We have found that people's heads are only approximately symmetric, so that the ±45° and ±90° degree views of each head have quite different detailed shape. In each case the range data was from the forward-facing, visible surface only.

Data from a 360° scan around each head was then used to form the stored model of each head that was later used for recognition. Full-detail versions of these eight reference models are shown in Figure 3. Only the low order 30 deformation modes were used for shape extraction and recognition. Because the low order modes provide a coarse, qualitative summary of the object shape (see the middle column of Figure 2) they can be expected to be the most stable with respect to noise and viewpoint change. Total execution time on a standard Sun 4/330 averaged approximately 3 seconds per fitting and recognition trial.

Recognition was accomplished by first recovering a 3-D model from the visible-surface range data, and then comparing the recovered mode values to the mode values stored for each of the eight known head models using Equation 11. The known model producing the largest dot product was declared to be the recognized object. The first seven modes were not employed, so that the recognition process was translation, rotation, and scale invariant.

In this experiment 92.5% accurate recognition was obtained. That is, we successfully recovered 3-D models and recognized each of the eight test subjects from each of the five different views with only three errors. Analysis of the recognition results showed that, while the average dot product between different reference models was 0.31 (72%), the average dot product between models recovered from different views of the same person was 0.95 (18°). Recognition was typically extremely certain. All three errors were
from front-facing views, where relatively few discriminating features are visible; remember that only overall head shape, and not details of surface shape, were available to the recognition procedure as only 30 modes were employed.

4.3 Head Recognition from Contours

In a similar head recovery and recognition experiment, we used a few 3-D head contours instead of full range data to see how well our techniques performed in the case of sparse silhouette data. In this experiment, we recovered heads from 2, 3, 4, and finally 5 contours, in order to approximate the information available in an active vision scenario. In each trial, the contours were spaced evenly in rotation. An example of the contours used in this experiment is shown in Figure 5; these contours were taken from the head shown in Figure 3(c).

As in the previous experiment, the recovered heads were compared against the full-detail versions of the reference heads shown in Figure 3, and the model producing the largest dot product was declared to be the recognized object. Heads were compared using the scale, rotation, and translation invariant version of Equation 11.

In our experiments, recognition with two contours averaged 93.75% accurate recognition. Accuracy improved as more contours were added until 96.875% of the heads were correctly identified when 5 contours were used. The results were not as good if the contours did not include the traditional side "silhouette," and performed best when the data's spring attachment was smoothed out more across the surface. Total execution times were slightly greater than those for the full data experiment, averaging 5 seconds per fitting and recognition trial on a Sun 4/330. The greater execution time is attributable to the more careful distribution and smoothing of spring attachment between contours and the underlying deformable model.

4.4 Recognition with Noise

In this experiment, the relatively similar objects shown in Figure 1 were used to generate synthetic data with varying object orientation and varying degrees of noise. From this synthetic data, 3-D models were recovered, and then recognized by comparing the mode values of the recovered objects to those of the original objects. Again, translation, rotation, and scale invariant matching was employed.

In each trial the 3-D orientation of each object was chosen from the entire viewing sphere at random, and the position of visible corners, edges, and face centers was measured, resulting in between 11 and 18 data points. Data points were then corrupted by adding uniformly distributed noise to each point's z, y, and z coordinates. Note that only front-facing 3-D points were used. The correspondence between data points and nodes was assumed to be known, so that performance of the fitting and recognition algorithms could be evaluated independently of the moment method for estimating initial orientation. Execution time averaged 0.1 seconds per trial on a Sun 4/330.

Figure 4 shows results of this experiment. The horizontal axis is the amount of noise, in millimeters, added to the synthetic range data. For comparison, the maximum dimension of each object is approximately 100 mm. The vertical axis is the mean percent accuracy of recognition over 100 trials. Error bars are shown for each level of noise.
As can be seen, a high level of accuracy is maintained up to about 8 mm (approximately 8%) noise, and thereafter declines smoothly. In this experiment almost all errors in recognition resulted from either (1) special views in which discriminating features were not visible, or (2) cases in which noise corrupted the discriminating features sufficiently that objects were confused with each other. Up to approximately 8 mm of noise almost all errors were due to special views, while for greater levels of noise the corruption of discriminating features dominates the error rate. We interpret the results of this experiment to indicate that our fitting and recognition algorithms are quite robust, and degrade gradually in the presence of noise.

5 Conclusion

We have described a closed-form solution for recovering a physically-based 3-D solid model from 3-D sensor measurements. In our current implementation we typically use 30 deformation modes, so that given as few as 11 independent 3-D sensor measurements the solution is overconstrained, and therefore unique except for rotational symmetries and degenerate conditions.

Because the recovered 3-D shape description is unique except for rotational symmetries, we may efficiently measure the similarity of different shapes by simply calculating normalized dot products between the mode values $U$ of various objects. Such comparisons may be made position, orientation and/or size independent by simply excluding the first seven mode amplitudes. Thus the modal representation seems likely to be useful for applications such as object recognition and spatial database search.

The two weaknesses in our current method are initial estimation of object orientation and the need for segmenting data into parts. Currently we use a standard method-of-moments for estimation of object orientation, and a minimum description length technique for segmentation [13]. For simple examples these techniques produce accurate, stable descriptions of shape; however for more complex scenes more sophisticated techniques will be required.

Acknowledgments

This research was made possible in part by the Rome Air Development Center (RADC) and the Defense Advanced Research Projects Agency (DARPA) under contract No. F30602-89-C-0022. We would like to thank the Harvard Face Archive, Cyberware Labs, and The Visual Computing Group for providing range data of human heads.

References