Nonlinear Manifold Learning for Visual Speech Recognition

Christoph Bregler and Stephen Omohundro

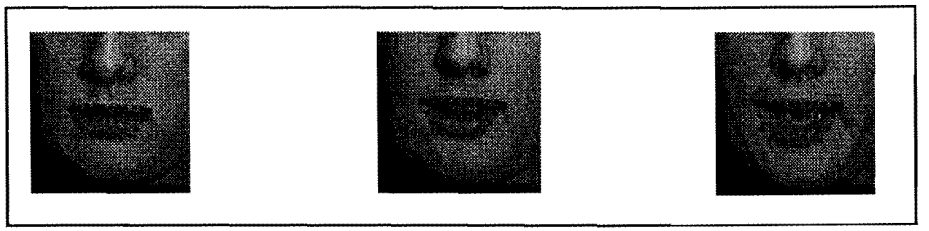
University of California, Berkeley & NEC Research Institute, Inc.

Overview

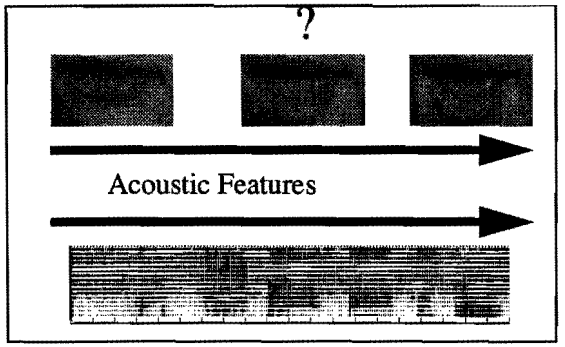
Manifold Learning:	A technique for representing and learning smooth nonlinear manifolds is presented and applied to several lip reading tasks. Given a set of points drawn from a smooth manifold in an abstract feature space, the technique is capable of determining the structure of the surface and of finding the closest manifold point to a given query point.
Applications to Lip Tracking, Image Interpolation, and Feature Extraction.	We use this technique to learn the "space of lips" in a visual speech recognition task. The learned manifold is used for tracking and extracting the lips, for interpolating between frames in an image sequence and for providing features for recognition.
Experiments with a full Speech Recognition System:	We describe a system based on Hidden Markov Models and this learned lip manifold that significantly improves the performance of acoustic speech recognizers in degraded environments. We also present preliminary results on a purely visual lip reader.

Problem

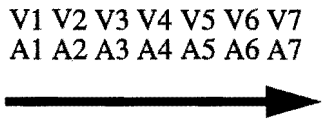
1. Boundary Tracking



2. Image Interpolation

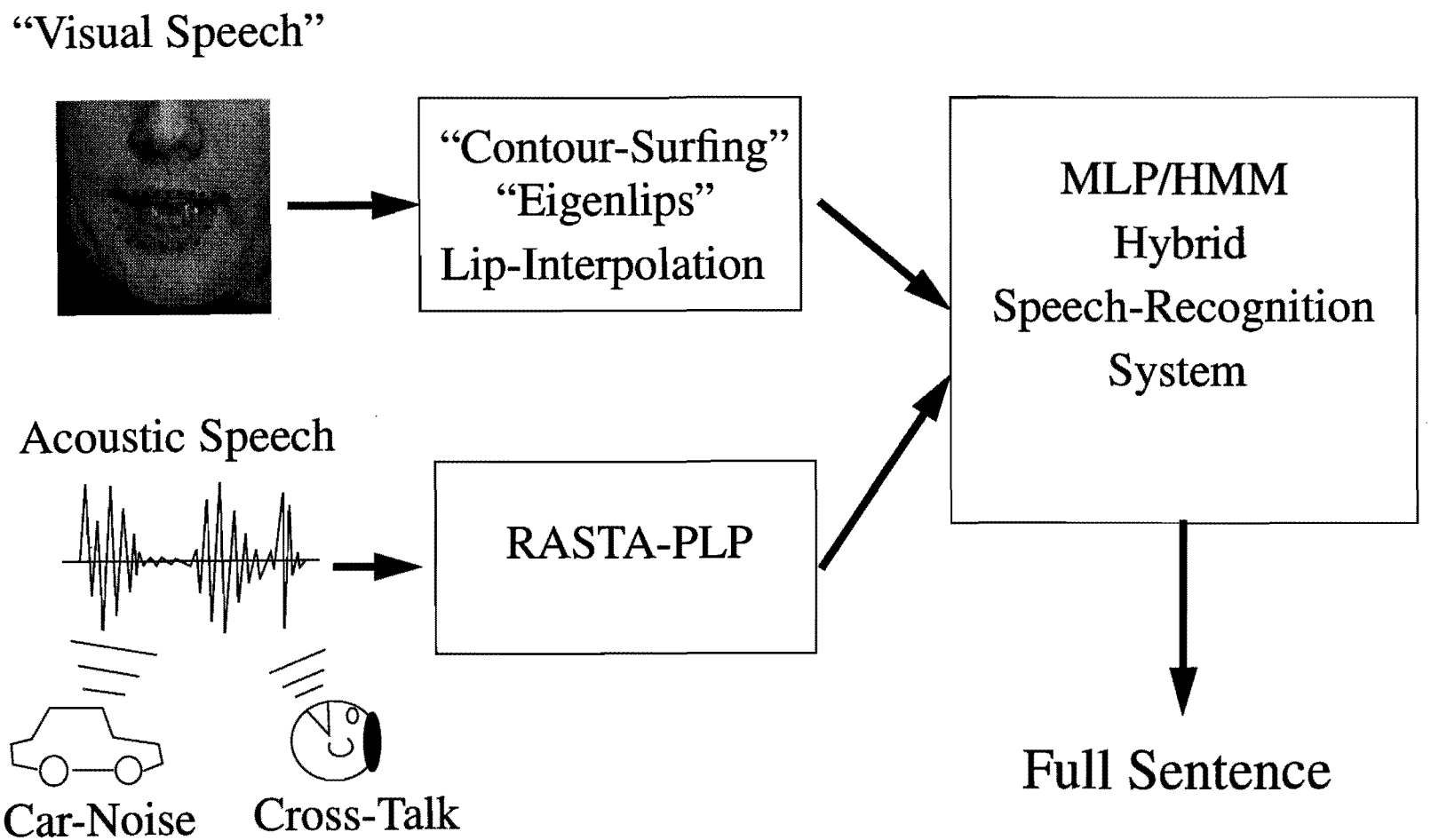


3. Lip Features

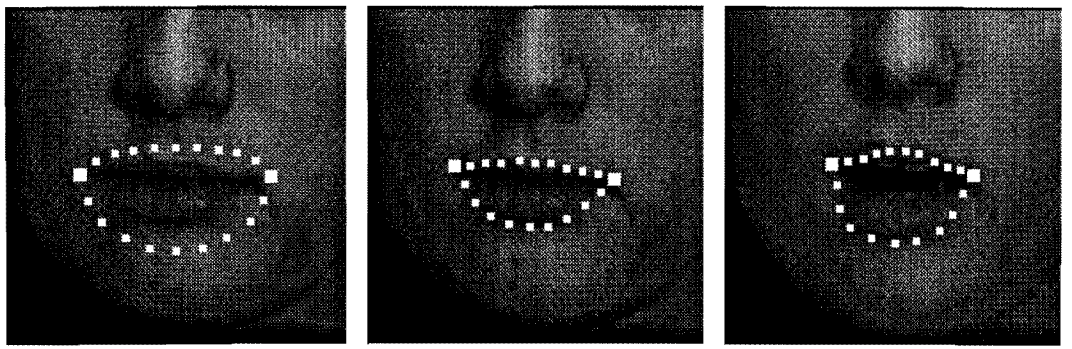


4. Speech Recognition

Full Recognition System

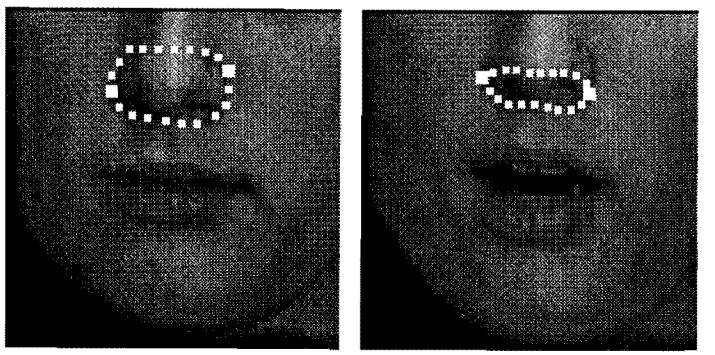


Lip Tracking using Active Contour Models



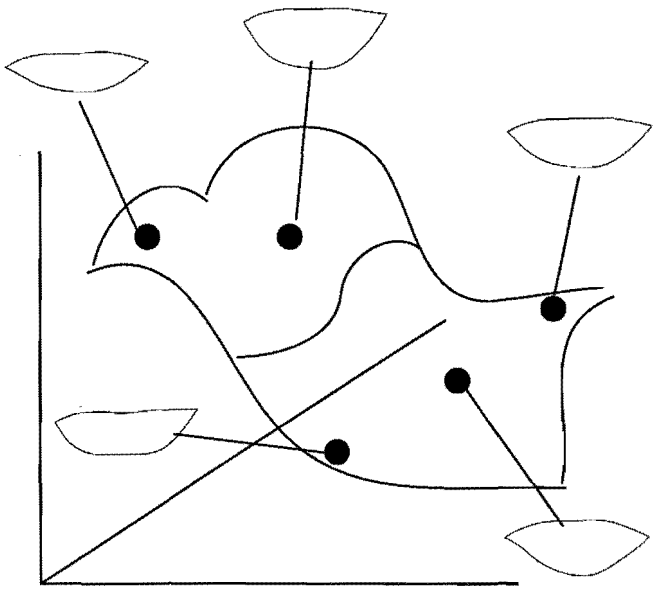
"Snake"-
Tracking

Snake Model =
"Controlled Continuity Spline"



Space of Contours as Constrained Manifold

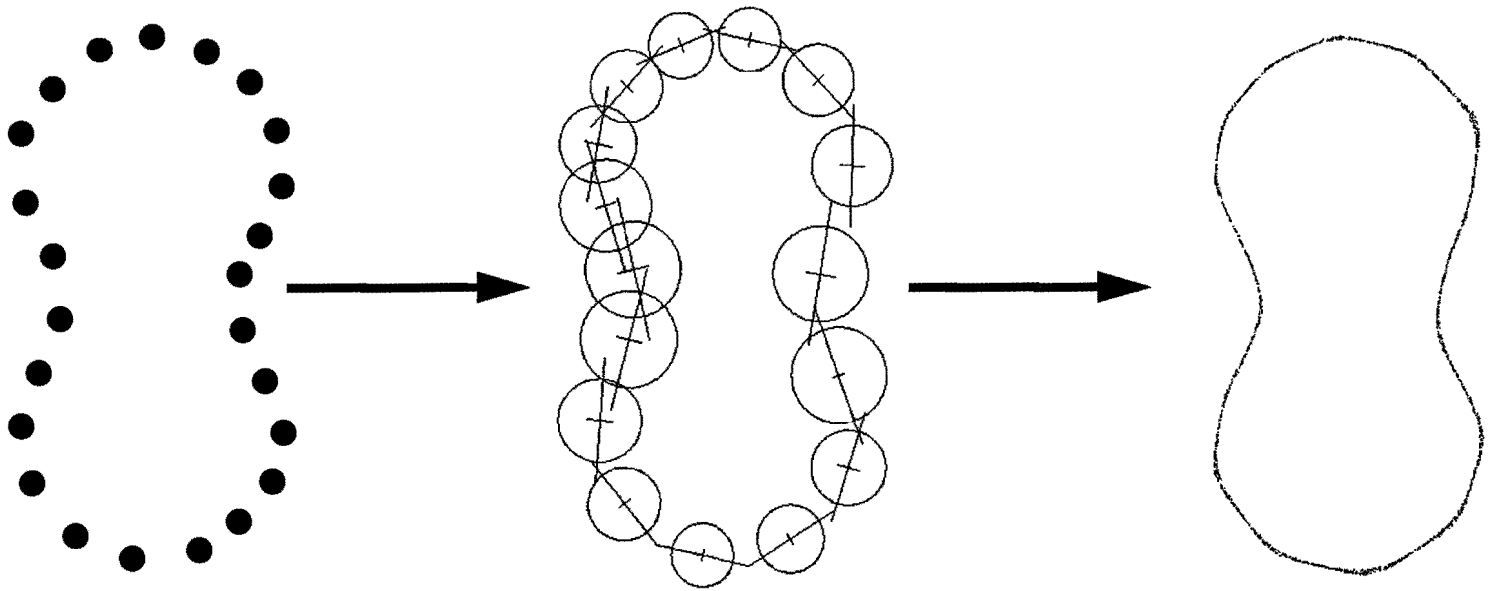
N Dimensional
Features are
Points in N Dimensional
Space.



K Dimensional
Subspace

N Dimensional Feature Space

Abstract Manifold Learning Task

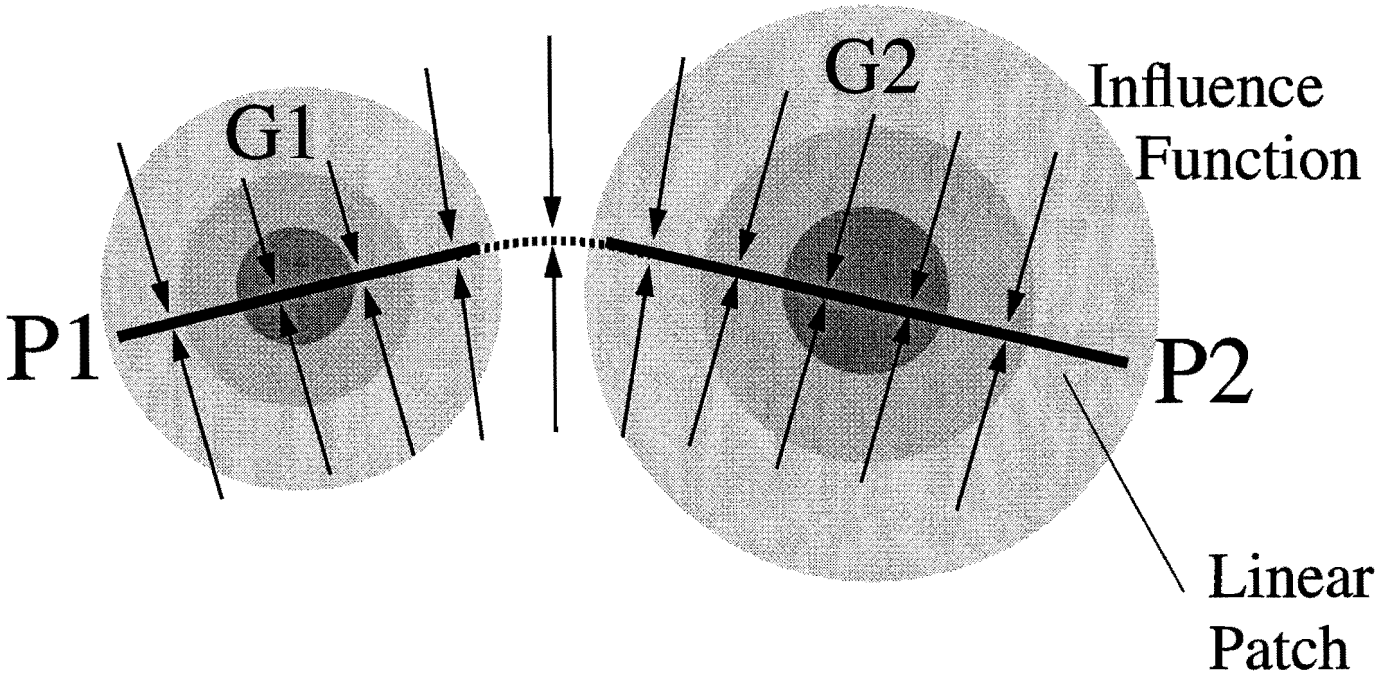


Samples

Mixture of
local adaptive Patches

Learned
Representation

Mixture of Patches Architecture

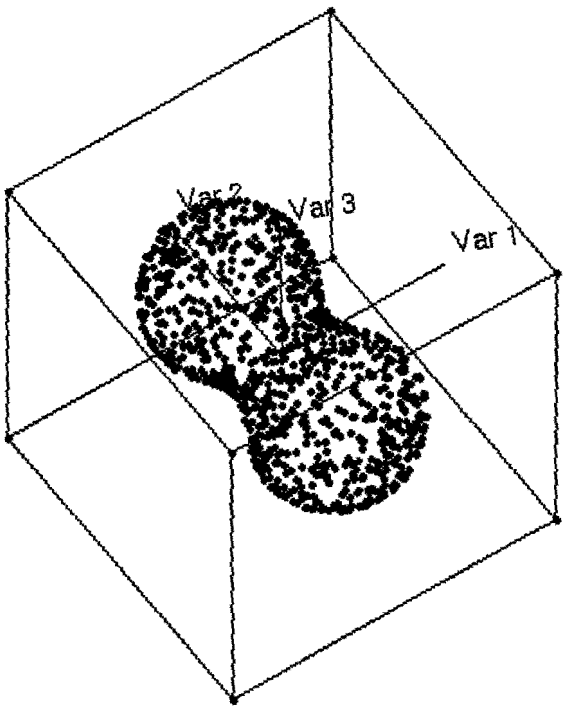


$$P(x) = \frac{\sum_i Gi(x) \cdot Pi(x)}{\sum_i Gi(x)}$$

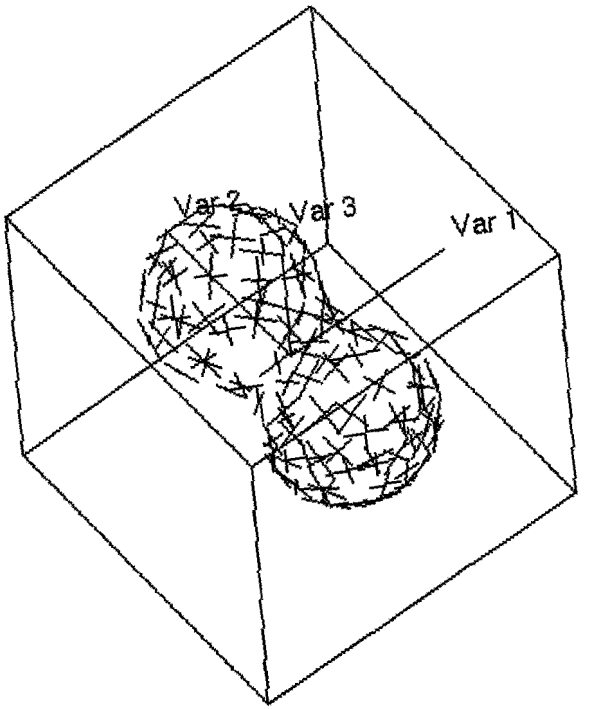
Manifold Learning

- Initial Estimate
  - Cluster Feature Space (K-Means)
  - Local PCA
- Minimize MSE between Training Set and Projection
  - EM Algorithm
  - Gradient Descent
- First Best Model Merging

Synthetic Examples

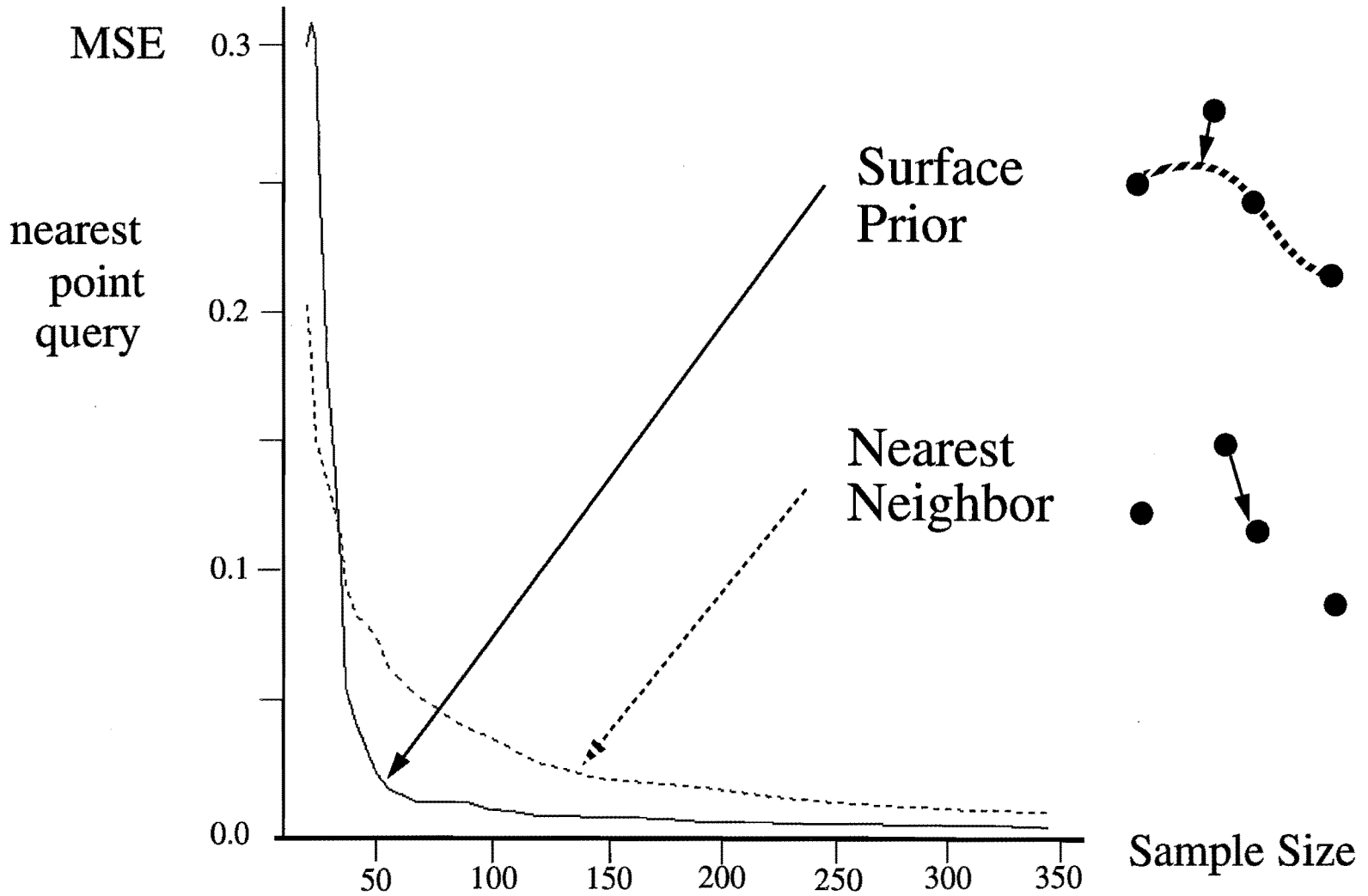


Training Samples

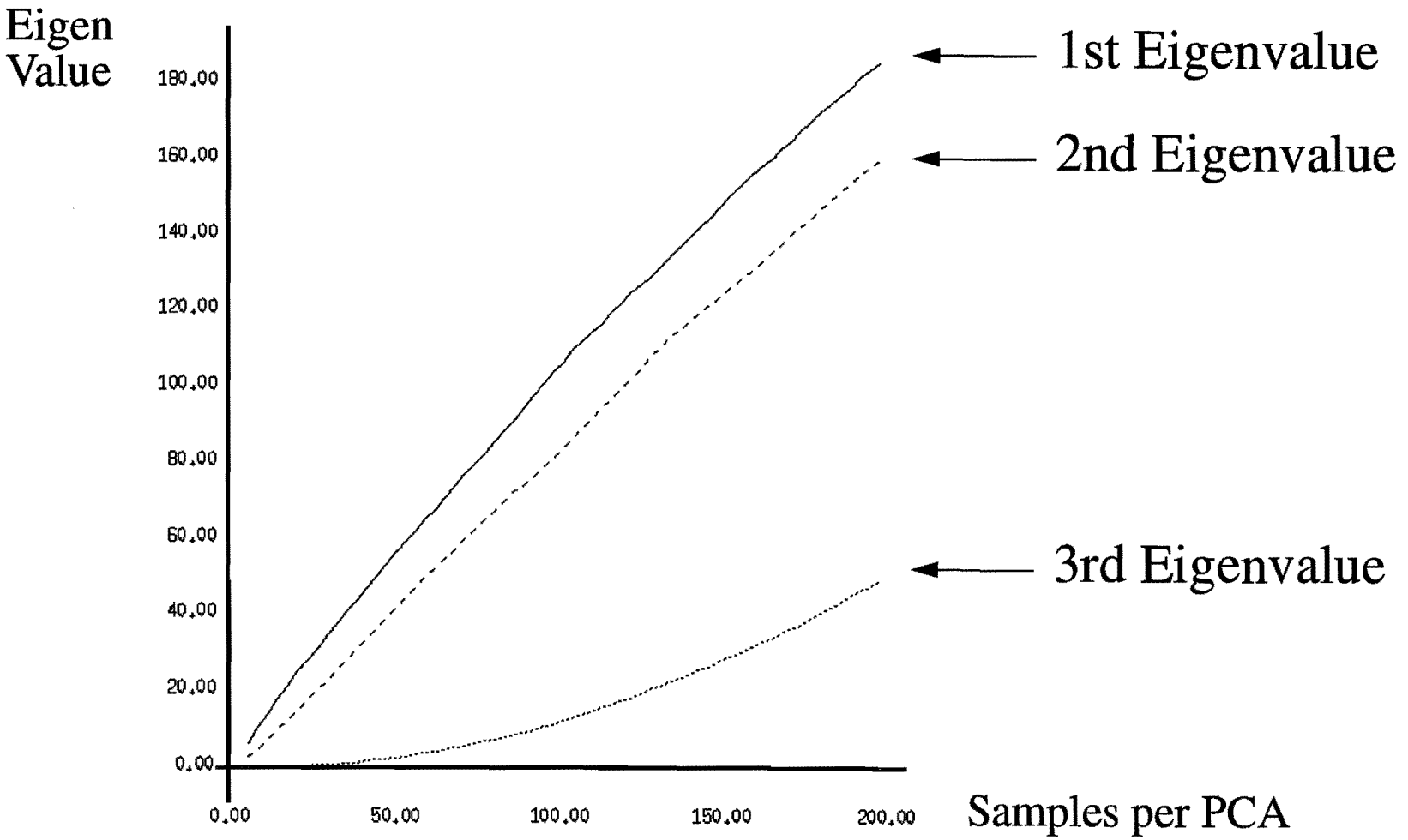


Learned Surface

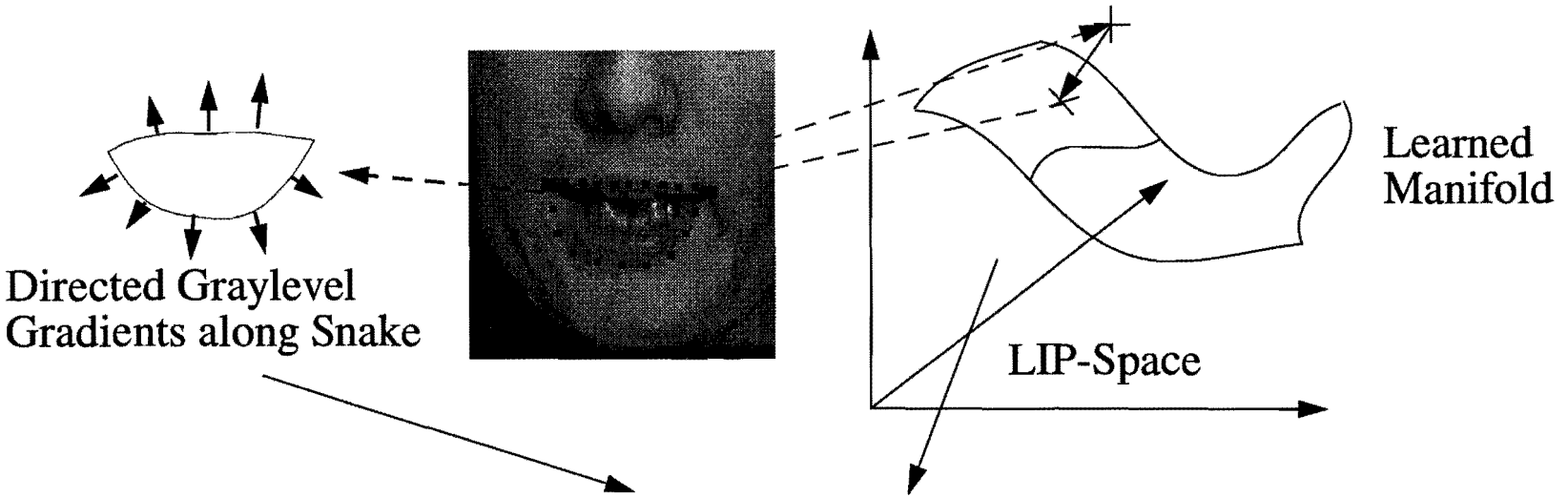
Comparison to Nearest Neighbor



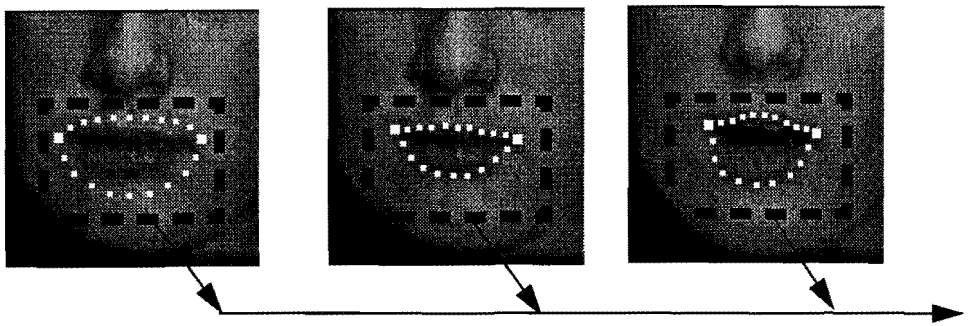
Manifold Dimension ?



Manifold-based Snake Tracking

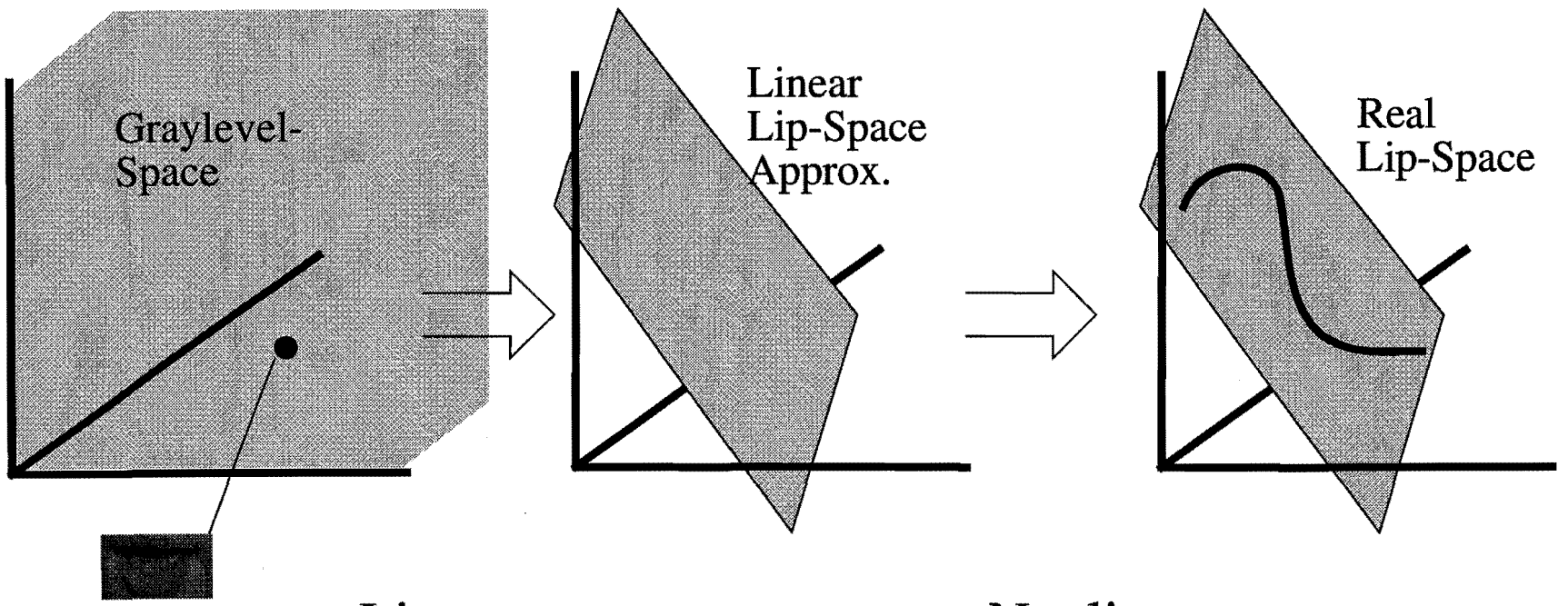


Manifold Snake Energy = Image Term+ Learned Manifold Term



Use Manifold-Snakes to estimate position, scale, and rotation for graylevel matrix extraction.

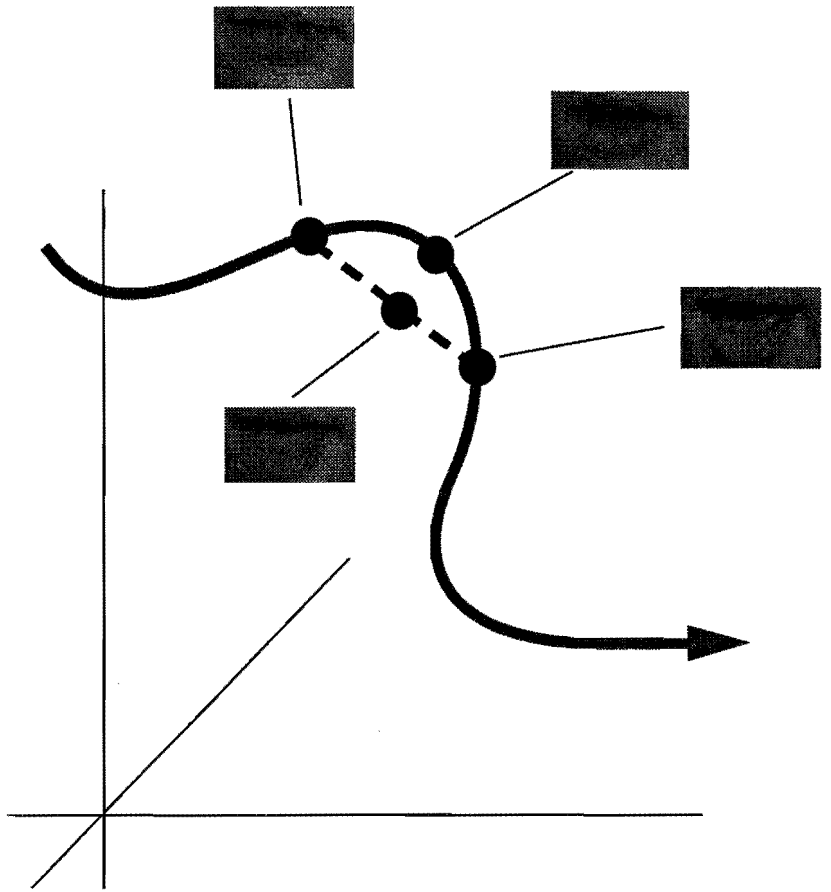
Graylevel Manifolds



Linear
Dimension Reduction

Nonlinear
Dimension Reduction

Linear vs. Nonlinear Interpolation



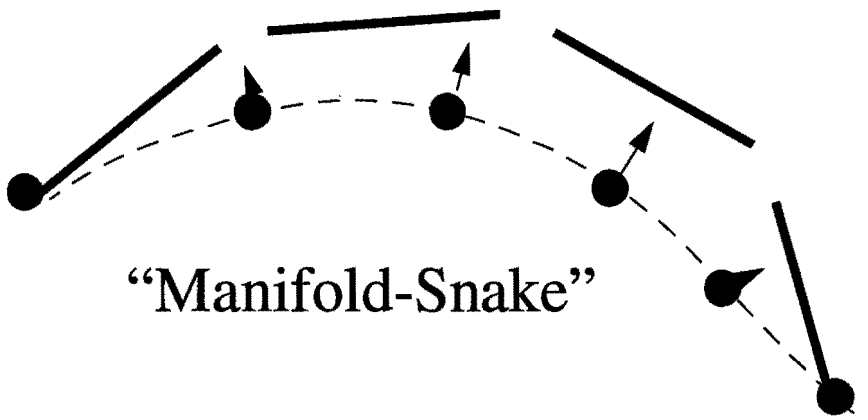
Graylevel
Dimensions

(16x16 pixel = 256 dim. space)

→ Learn
nonlinear subspace
of
"legal" images

→ constrain
interpolated
images to
subspace

Nonlinear Interpolation Technique

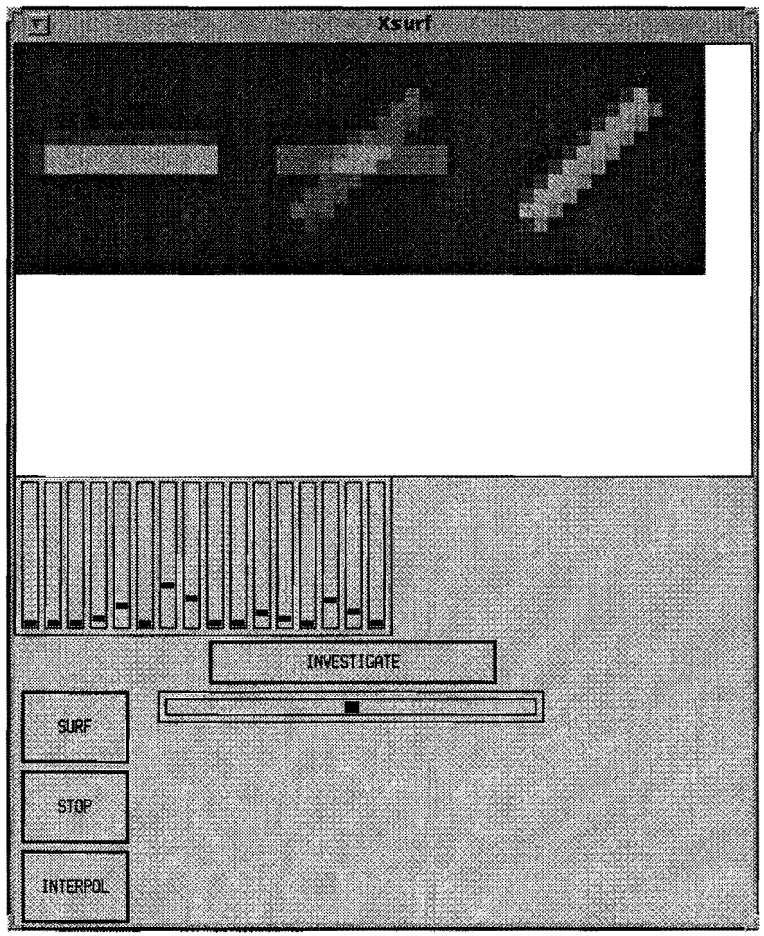


$$E = \sum_i \left| v_{i-1} - 2v_i + v_{i+1} \right|^2 + \text{distance}\langle v_i \rangle$$

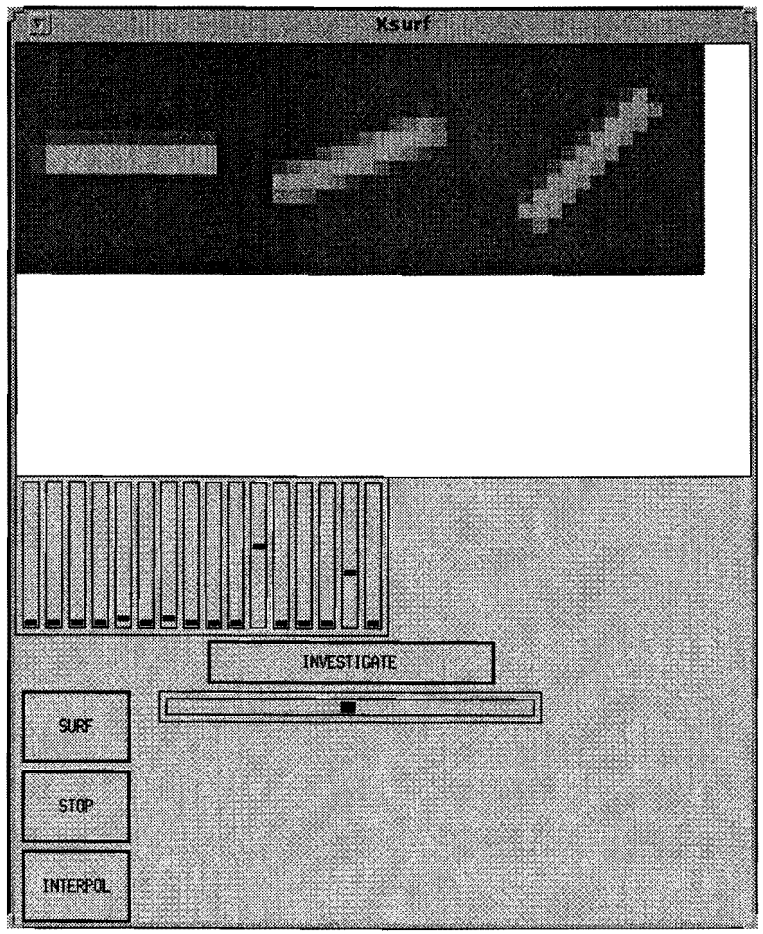
- Begin with sequence of linear interpolated points
- Iteratively move points toward manifold
- Iterations are equal to gradient descent steps using the Manifold-Snake Energy

Artificial Images

Linear Interpolation

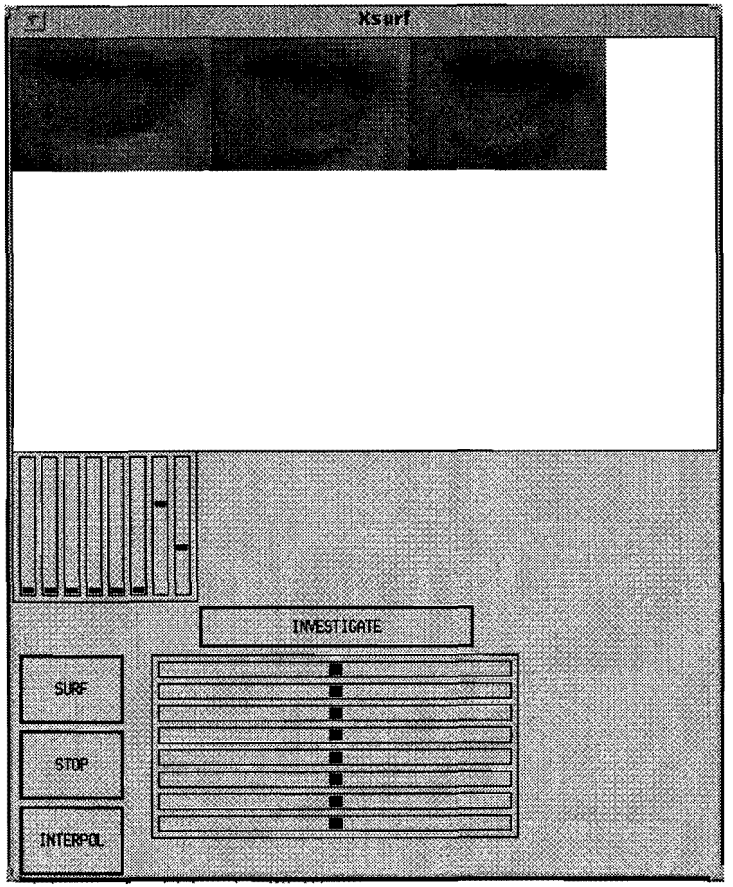


Nonlinear Interpolation

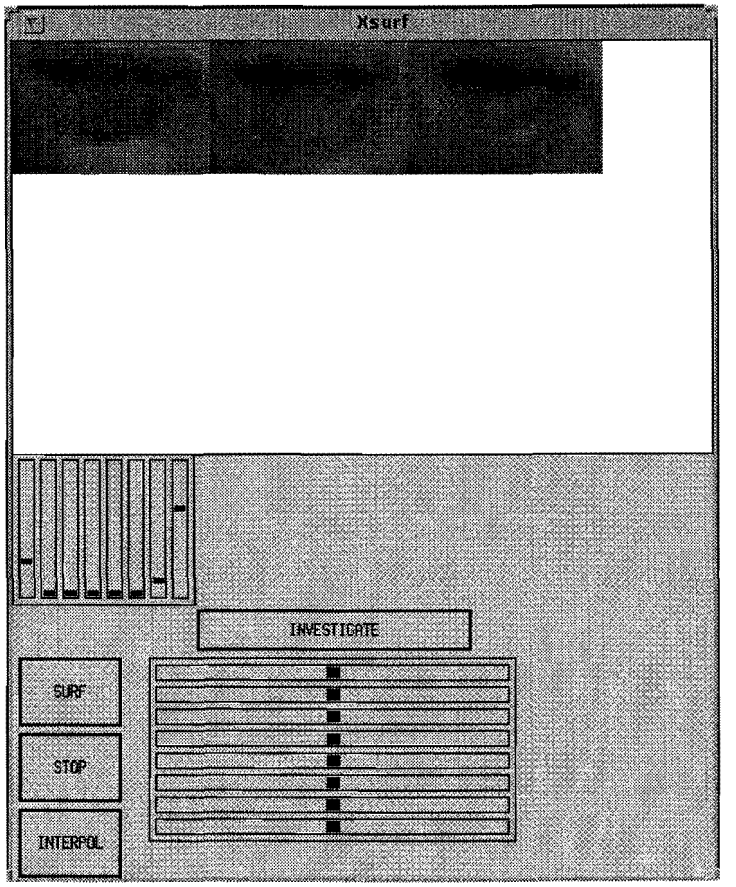


Natural Lip Images

Linear Interpolation

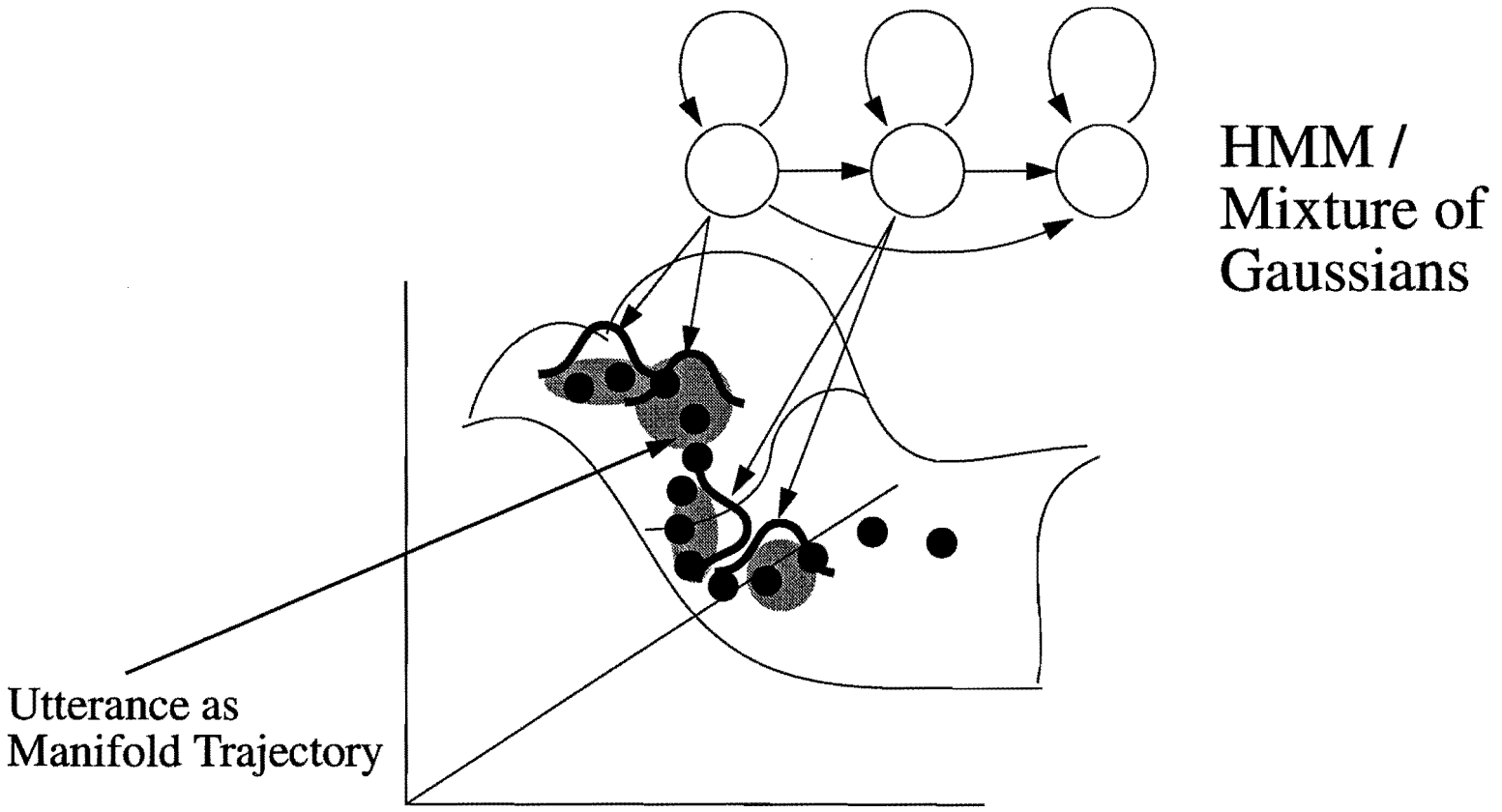


Nonlinear Interpolation

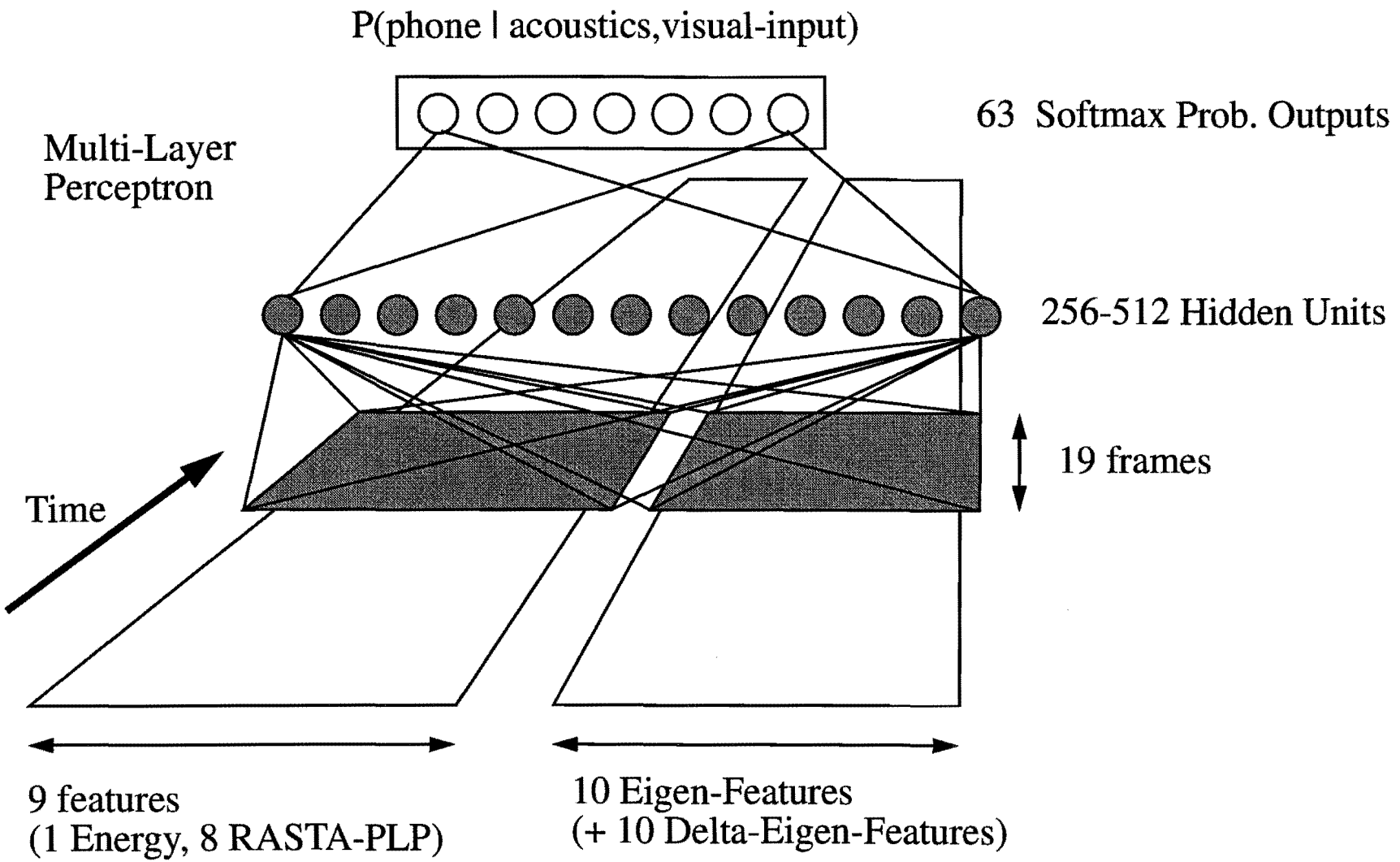


24x24 Images projected in a 32-D linear space and 8-D nonlinear manifold

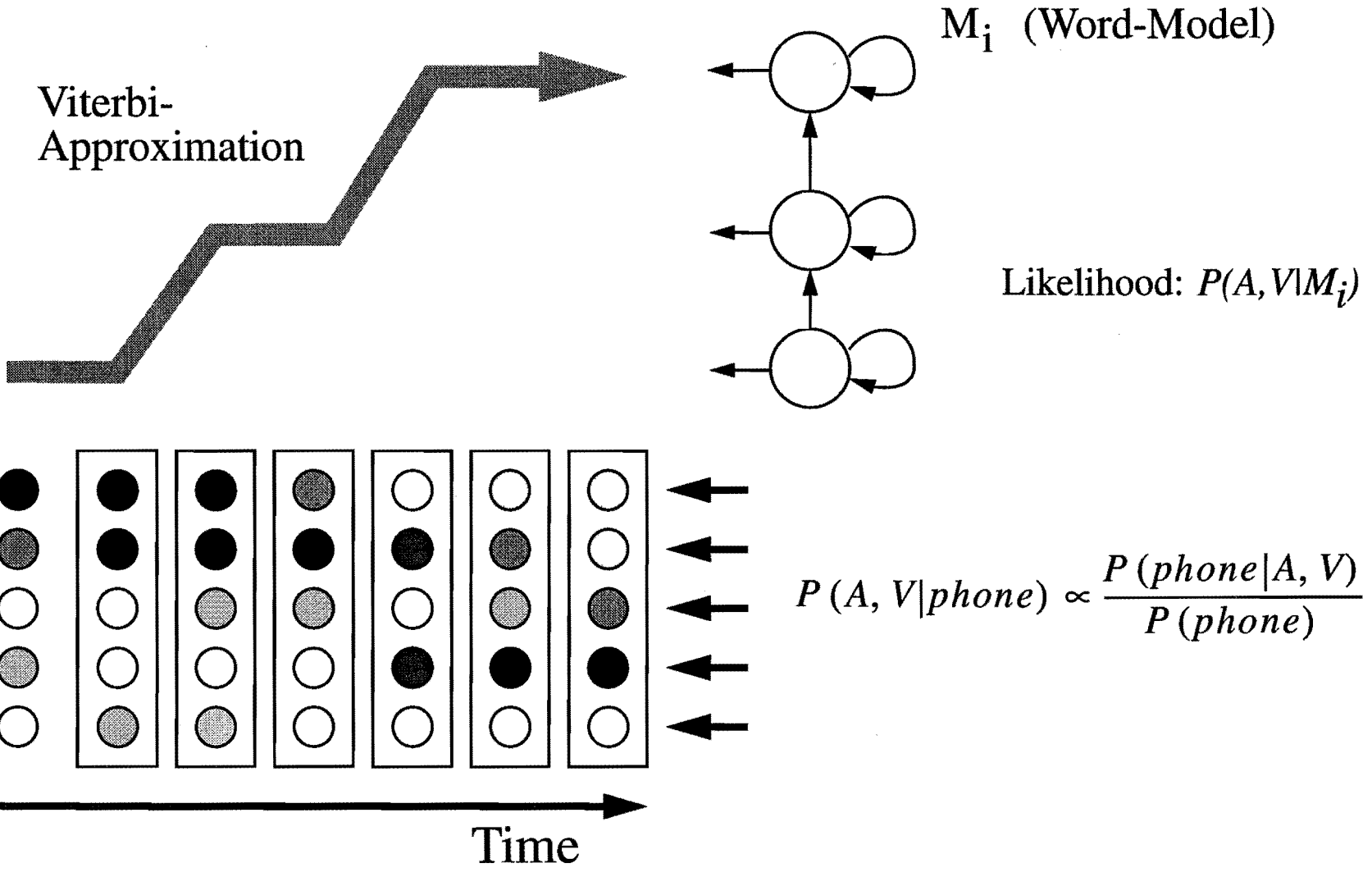
Phone/Word Probability Estimation



Phone Probability Estimation -2-



Hidden Markov Models



$$P(A, V|phone) \propto \frac{P(phone|A, V)}{P(phone)}$$

Recognition Results

Task	Acoustic	Eigenlips	Delta-Lips	Rel.Err.Red.
clean	11.0 %	10.1 %	11.3 %	-
20db SNR	33.5 %	28.9 %	26.0 %	22.4 %
10db SNR	56.1 %	51.7 %	48.0 %	14.4 %
15db SNR crosstalk	67.3 %	51.7%	46.0 %	31.6 %

Table 1: Spelling Task (6 Speakers) Word Error (wrong + insertion + deletion)

Task	Pure Lipreading-Error
Bar-Tender	4.5%

Table 2: Pure Lipreading (1 Speaker, "Bar-Task": 4 Cocktail-Names)

Related Work

Related *Linear* Representations:

Kirby et al., Pentland et al.: Linear Subspaces
Simard: Tangent Distance

Related *Nonlinear* Representation:

Jacobs, Jordan, Nowlan, Hinton: Mixture of Experts
Kambhatla, Leen: Local PCA

Related Interpolation Techniques:

Poggio et al.: RBFs for Image Interpolation

Computer Lipreading History

- *U.S.Patent 3192321 (1965) E. Nassimbene (IBM)*
- *E. Petajan (1984), Dissertation*
- *B. Yuhas, M. Goldstein, T. Sejnowski (1989)*
- *K. Mase, A. Pentland (1991)*
- *O. Garcia, A. Goldschen, E. Petajan (1992)*
- *D. Stork, G. Wolff, K.V. Prasad, M. Hennecke (1992)*
- *P.L. Silsbee (1993), Dissertation*
- *A.J. Goldschen (1993), Dissertation*
- *J. Movellan (1994)*

Summary

Lipreading:

- Improves Speech Recognition Performance significantly
- Full System Solution

Manifold Learning:

- New fundamental technique for learning in geometric domains
- Applicable to variety of different tasks