Nonlinear Manifold Learning for Visual Speech Recognition

Christoph Bregler

Computer Science Division
Soda Hall, U.C. Berkeley
Berkeley, CA 94720
bregler@cs.berkeley.edu

Stephen M. Omohundro*

NEC Research Institute, Inc.
4 Independence Way
Princeton, NJ 08540
om@research.nj.nec.com

Abstract

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. 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. 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.

1 Introduction

This paper describes a new technique that is the basis for a "visual speech recognition" or "lip reading" system. Model-based vision systems currently have the best performance for many visual recognition tasks. For geometrically simple domains, models can sometimes be constructed by hand using CAD tools. Such models are difficult and expensive to construct, however, and are inadequate in more complex domains. To do model-based lipreading, we would like a parameterized model of the complex "space of lip configurations". Rather than building such a model by hand, our approach is to build it using machine learning. The system is given a collection of training images which it uses to automatically construct the models that are later used in recognition.

There are several phases of processing involved in our system. Ultimately, the recognition of the time sequence of images uses Hidden Markov Model technology similar to auditory speech recognition systems. Unlike speech recognition, however, there are extra phases to find, track, and extract the lips from a sequence of images. We will describe how learned models are used to facilitate these tasks.

Some versions of our system do recognition based only on the visual input, while others use both visual and acoustic information. When visual and acoustic information is combined, it is necessary to deal with the fact that the acoustic sampling rate is higher than the visual image rate. We will describe how the learned models are used to interpolate between frames of video.

There is a single abstract learning task that underlies these different taks. We use the expression "nonlinear manifold learning" for the task of inducing a smooth nonlinear surface in a high-dimensional space from a set of points drawn from that surface. This task is important throughout vision because the parameters of visual tasks are often related by smooth nonlinear constraints. Learning such constraints and manipulating them in a computationally tractable way is therefore central to building learning-based visual recognition systems.

The first section of this paper describes the manifold representation and learning algorithm. Next we describe the use of learned manifolds for interpolation. We then present the "lip manifold" that our system learns for visual speech recognition. We show how this is used to improve the performance of a snake-based lip tracker and to interpolate between lip images. We then give recognition performance results for a single speaker based only on visual information. We conclude with more complex experiments on multiple speakers, combined visual and acoustic information, and in the context of a spontaneous speech dialog system.

2 Smooth nonlinear manifold representation and induction

2.1 Motivation

Human lips are geometrically complex shapes which smoothly vary with the multiple degrees of freedom of the facial musculature of a speaker. For recognition, we would like to extract information about these degrees of freedom from images. We represent a single configuration of the lips as a point in a feature space. The set of all configurations that a speaker may exhibit then defines a smooth surface in the feature space. In differential geometry, such smooth surfaces are called "manifolds".

For example, as a speaker opens her lips, the corresponding point in the lip feature space will move

[1] Formerly at ICSI, Berkeley

Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499

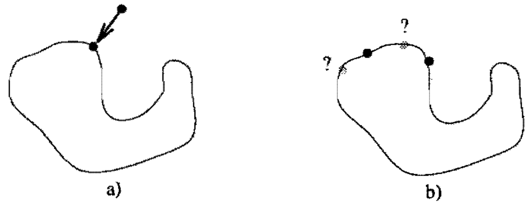

Figure 1: Surface tasks a) Closest point query, b) interpolation and prediction

along a smooth curve. If the orientation of the lips is changed, then the configuration point moves along a different curve in the feature space. If both the degree of openness and the orientation vary, then a two-dimensional surface will be described in the feature space. The dimension of the "lip" surface is the same as the number of degrees of freedom of the lips. This includes both intrinsic degrees of freedom due to the musculature and external degrees of freedom which represent properties of the viewing conditions.

We would like to learn the lip manifold from examples and to perform the computations on it that are required for recognition. We abstract this problem as the "manifold learning problem": *given a set of points drawn from a smooth manifold in a space, induce the dimension and structure of the manifold.*

There are several operations we would like the surface representation to support. Perhaps the most important for recognition is the "nearest point" query: return the point on the surface which is closest to a specified query point (Fig. 1a). This task arises in any recognition context where the entities to be recognized are smoothly parameterized (eg. objects which may be rotated, scaled, etc.) There is one surface for each class which represents the feature values as the various parameters are varied [13]. Under a distance-based noise model, the best classification choice for recognition will be to choose the class of the surface whose closest point is nearest the query point. The chosen surface determines the class of the recognized entity and the closest point gives the best estimate for values of the parameters within that class. The same query arises in several other contexts in our system. The surface representation should therefore support it efficiently.

Other important classes of queries are "interpolation queries" and "prediction queries". Two or more points on a curve are specified and the system must interpolate between them or extrapolate beyond them. Knowledge of the constraint surface can dramatically improve performance over "knowledge-free" approaches like linear or spline interpolation. (Fig. 1b)

2.2 Manifold representation based on the closest point query

In this section we describe a manifold representation based on the closest point query [2]. If the data points were drawn from a *linear* manifold, then we could represent it by a point on the surface and a projection matrix. After the specified point is translated to the origin, the projection matrix would project any vector orthogonally into the linear subspace. Given a set of points drawn from such a linear surface, a principal components analysis could be used to discover its dimension and to find the least-squares best fit projection matrix. The largest principal vectors would span the space and there would be a precipitous drop in the principle values at the dimension of the surface (This is similar to approaches described [9, 18, 17]). A principal components analysis no longer suffices, however, when the manifold is nonlinear because even a 1-dimensional nonlinear curve can span all the dimensions of a space.

If a nonlinear manifold is smooth, however, then each local piece looks more and more linear under magnification. Surface data points from a small local neighborhood will be well-approximated by a linear patch. Their principal values can be used to determine the most likely dimension of the patch. We take that number of the largest principal components to approximate the tangent space of the surface. The idea behind our representations is to "glue" such local linear patches together using a partition of unity.

The manifold is represented as a mapping from the embedding space to itself which takes each point to the nearest point on the manifold. K-means clustering is used to determine an initial set of "prototype centers" from the data points. A principal components analysis is performed on a specified number of the nearest neighbors of each prototype point. These "local PCA" results are used to estimate the dimension of the manifold and to find the best linear projection in the neighborhood of prototype i. The influence of these local models is determined by Gaussians centered on the prototype location with a variance determined by the local sample density. The projection onto the manifold is determined by forming a partition of unity from these Gaussians and using it to form a convex linear combination of the local linear projections:

$$P(x) = \frac{\sum_i G_i(x) P_i(x)}{\sum_i G_i(x)} \qquad (1)$$

This initial model is then refined to minimize the mean squared error between the training samples and the nearest surface point using EM optimization [4] and gradient descent. We have demonstrated the excellent performance of this approach on synthetic examples [3]. A related mixture model approach applied to input-output mappings appears in [7].

3 Using manifold representation for interpolation

This representation is especially suited for nearest point queries. We are interested in using this query to interpolate between two specified points. Geometrically, a linear interpolant moves along the straight line joining two points and will typically not lie within the constraint surface.

In our non-linear approach to interpolation, the point moves along a curve in the learned manifold that joining the two points. This constrains the interpolated point to only "meaningful" values. We

**Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499**

have studied several algorithms for approximating the shortest manifold trajectory connecting two given points [3], but report here only the most sucessful one. We use the term "*Surface-Snake*" to refer to a sequence of n feature space points which are approximately on the surface. An energy function is defined on such sequences of points which prefers curves that better satisfy the three criteria of smoothness, equidistance, and nearness to the surface:

$$E = \sum_i \alpha||v_{i-1}-2v_i+v_{i+1}||^2+\beta||v_i-proj(v_i)||^2 \quad (2)$$

E has value 0 if all v_i are equally distributed on a straight line and also lies on the surface. In general E will not achieve the value 0 if the surface is nonlinear, but the system tries to minimize it.

The interpolation algorithm begins with a straight line between the two query points and performs gradient descent in E to find the optimal solution. For another approach to nonlinear interpolation using a different architecture see [15].

4 Application to visual speech recognition

We use these manifold learning techniques in a system for visual speech recognition. We view certain feature vectors of human lips as points which are constrained to lie on a low-dimensional nonlinear manifold embedded in the lip feature space. This manifold represents all possible lip configurations. While uttering a word or a sentence the "lip feature point" moves along a trajectory on this manifold.

We model these trajectories using Hidden Markov Models (HMMs). The domain of the HMM emission vectors is defined by the lip-manifold. Therefore a specific HMM word model represents the probability distribution over trajectories on the lip-manifold for a given word. We represent the emission probability distributions by a mixture of gaussians or by a multi-layer-perceptron (MLP).

To get the input for the Hidden Markov Model we first find and track the lip position (section 4.1). We then extract the lip image at the selected location and size. It is then encoded as a point in a lip-feature space (section 4.2). When we want to perform combined acoustic and visual recognition, we fuse the visual n-dimensional visual feature vector together with a m-dimensional acoustic feature vector obtained from an acoustic front end (section 4.4). Because the acoustic vectors are produced with a higher frame rate (necessary for good acoustic recognition), we need to interpolate the visual vectors (section 3). This produces a sequence of combined visual-acoustic $n+m$-dimensional vectors as input for the HMM.

The parameters of the HMM are set by the Baum-Welch procedure from a set of example utterances. We train a separate HMM for each word that is to be recognized. Once learned, the HMM's may be presented with a sequence of purely visual feature vectors or a sequence of bimodal visual-acoustic vectors. Each HMM estimates the likelihood that it generated the sequence and the most likely HMM is selected as the class of the utterance. In the pure visual domain we are interested in the recognition performance on the word level (section 4.3). In the visual-acoustic domain we are interested in the improvement that visual information can make over purely acoustic continuous speech recognition (section 4.5).

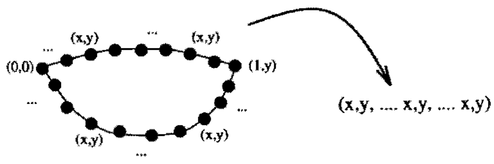

Figure 2: Lip contour coding

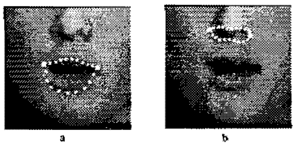

Figure 3: Snakes for finding the lip contours a) A correctly placed snake b) A snake which has gotten stuck in a local minimum of the simple energy function.

4.1 Constraint boundary tracking

To track the position of the lips we use an "active vision" technique related to "snakes" [8] and "deformable templates" [21]. In each image, a contour shape is matched to the boundary of the lips. The space of contours that represent lips is represented by a learned lip-contour-manifold. During tracking we try to find the contour (manifold-point) which maximizes the graylevel gradients along the contour in the image.

The boundary shape is parameterized by the x and y coordinates of 40 evenly spaced points along the contour. The left corner of the lip boundary is anchored at $(0,0)$ and all values are normalized to give a lip width of 1 (Fig 2). Each lip contour is therefore a point in an 80-dimensional "contour-space" (because of anchoring and scaling it is actually only a 77-dimensional space).

The training set consists of 4500 images of 6 speakers uttering random words. The training images are initially labeled with a conventional *snake* algorithm. The standard *snake* approach chooses a curve by trying to maximizing its smoothness while also adapting to certain image features along its length. These criteria are encoded in an energy function and the snake is optimized by gradient descent. Unfortunately, this approach sometimes causes the selection of incorrect regions (Fig. 3). We cull the incorrectly aligned *snakes* from the database by hand.

We then apply the manifold learning technique described above to the database of correctly aligned lip snakes. The algorithm learns a 5-dimensional manifold embedded in the 80-dimensional contour space. 5 dimensions were sufficient to describe the contours with single pixel accuracy in the image. Figure 4

**Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499**

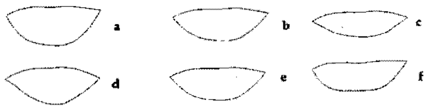

Figure 4: Two principle axes in a local patch in lip space. a, b, and c are configurations along the first principle axis, while d, e, and f are along the third axis.

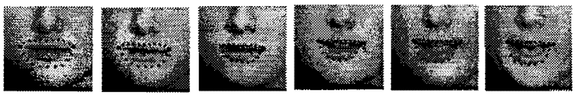

Figure 5: A typical relaxation and tracking sequence of our lip tracker

shows some of the lip models along two of the principal axes in the local neighborhood of one of the patches.

The tracking algorithm starts with a crude initial estimate of the lip position and size. In our training database all subjects positioned themselves at similar locations in front of the camera. The initial estimate is not crucial to our approach as we explain later. Currently work is in progress to integrate a full face finder, which will allow us to estimate the lip location and size without even knowing the rough position of the subject.

Given the initial location and size estimate, we backproject an initial lip contour from the lip-manifold back to the image (we choose the mean of one of the linear local patches). At each of the 40 points along the backprojected contour we estimate the magnitude of the graylevel gradient in the direction perpendicular to the contour. The sum of all 40 gradients would be maximal if the contour were perfectly aligned with the lip boundary. We iteratively maximize this term by performing a gradient ascent search over the 40 local coordinates. After each step, we anchor and normalize the new coordinates to the 80-dimensional shape space and project it back into the lip-manifold. This constrains the gradient ascent search to only to consider legal lip-shapes. The search moves the lip-manifold point in the direction which maximally increases the sum of directed graylevel gradients. The initial guess only has to be roughly right because the first few iterations use big enough image filters that the contour is attracted even far from the correct boundary.

The lip contour searches in successive images in the video sequence are started with the contour found from the previous image. Additionally we add a temporal term to the gradient ascent energy function which forces the temporal second derivatives of the contour coordinates to be small. Figure 5 shows an example gradient ascent for a starting image and the contours found in successive images.

4.2 Lip Image Coding and Interpolation

In initial experiments we directly used the contour coding as the input to the recognition Hidden Markov Models, but found that the outer boundary of the lips is not distinctive enough to give reasonable recognition performance. The inner lip-contour and the appearance of teeth and tongue are important for recognition. These features are not very robust for lip tracking, however, because they disappear frequently when the lips close. For this reason the recognition features we use consist of the components of a graylevel matrix positioned and sized at the location found by the contour based lip-tracker. Empirically we found that a matrix of 24x16 pixels is enough to distinguish all possible lip configurations. Each pixel of the 24x16 matrix is assigned the average graylevel of a corresponding small window in the image. The size of the window is dependent of the size of the found contour. Because a 24x16 graylevel matrix is equal to a 384-dimensional vector, we also reduce the dimension of the recognition feature space by projecting the vectors to a linear subspace determined by a principal components analysis.

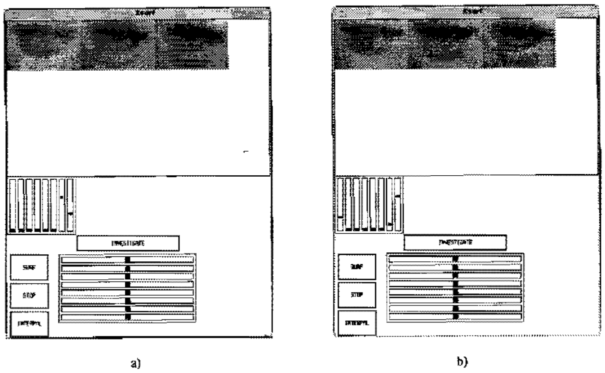

Figure 6: 24x24 images projected into a 32 dimensional subspace: a) linear interpolation b) nonlinear interpolation. (The slider bars represent the current weights for the linear patches which are necessary to produce the interpolated image)

To interpolate missing lip-images, we induce a nonlinear manifold embedded in this lower dimensional subspace. The interpolation is done in the lower dimensional linear space and is also constrained by the learned manifold. Figure 6 shows an example interpolation of lip images in a 32-dimensional linear subspace. Figure 6a shows the linear interpolation, and Figure 6b shows the nonlinear interpolation constrained by an 8-dimensional manifold, using the manifold-snake interpolation technique.

4.3 One speaker, pure visual recognition

The simplest of our experiments is based on a small speaker dependent task, the "bartender" problem. The speaker may choose between 4 different cocktail names[1], but the bartender cannot hear due to background noise. The cocktail must be chosen purely by lipreading. A subject uttered each of the 4 words 23 times. An HMM was trained for each of the 4 words

[1] We choose the words: "anchorsteam", "bacardi", "coffee", and "tequilla". Each word takes about 1 second to utter on average.

Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499

using a mixture of Gaussians to represent the emission probabilities. With a test set of 22 utterances, the system made only one error (4.5% error).

This task is artificially simple, because the vocabulary is very small, the system is speaker dependent, and it does not deal with continuous or spontaneous speech. These are all state-of-the-art problems in the speech recognition community. For pure lip reading, however, the performance of this system is sufficiently high to warrant reporting here. The following sections describe more state-of-the-art tasks using a system based on combined acoustic and visual modalities.

4.4 Acoustic processing and sensor fusion

For the acoustic preprocessing we use an off-the-shelf acoustic front-end system, called RASTA-PLP [6] which extracts feature vectors from the digitized acoustic data with a constant rate of 100 frames per second.

Psychological studies have shown that human subjects combine acoustic and visual information at a rather high feature level. This supports a preceptual model that posits conditional independence between the two speech modalities [11]. We believe, however, that such conditional independence cannot be applied to a speech recognition system that combines modalities on the phoneme/viseme level. Visual and acoustic speech vectors are conditionally independent given the vocal tract position, but not given the phoneme class. Our experiments have shown that combining modalities at the input level of the speech recognizer produces much higher performance than combining them on higher levels.

4.5 Multi-speaker visual-acoustic recognition

In this experiment, the aim is to use the the visual lipreading system to improve the performance of acoustic speech recognition. We focus on scenarios where the acoustic signal is distorted by background noise or crosstalk from another speaker. State-of-the-art speech recognition systems perform poorly in such environments. We would like to know how much the additional visual lip-information can reduce the error of a purely acoustic system.

We collected a database of 6 speakers spelling names or saying random sequences of letters. Letters can be thought of as small words, which makes this task a connected word recognition problem. Each utterance was a sequence of 3-8 letter names. The spelling task is notoriously difficult, because the words (letter names) are very short and highly ambiguous. For example the letters "n" and "m" sound very similar, especially in acoustically distorted signals. Visually they are more distinguishable (it is often the case that visual and acoustic ambiguities are complementary, presumably because of evolutionary pressures on language). In contrast, "b" and "p" are visually similar but acoustically different (voiced plosive vs. unvoiced plosive). Recognition and segmentation (when does one letter end and another begin) have additional difficulties in the presence of acoustical crosstalk from another speaker. Correlation with the visual image of one speaker's lips helps disambiguate the speakers.

Task	Acoustic	AV	Delta-AV	relative 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 w/ crosstalk	67.3 %	51.7 %	46.0 %	31.6 %

Table 1: Results in word error (wrong words plus insertion and deletion errors caused by wrong segmentation)

Our training set consists of 2955 connected letters (uttered by the 6 speakers). We used an additional cross-validation set of 364 letters to avoid overfitting. In this set of experiments the HMM emission probabilities were estimated by a multi-layer-perceptron (MLP) [3]. The same MLP/HMM architecture has achieved state-of-the-art recognition performance on standard acoustic databases like the ARPA resource management task.

We have trained 3 different versions of the system: one based purely on acoustic signals using 9-dimensional RASTA-PLP features, and two that combine visual and acoustic features. The first bimodal system (AV) is based on the acoustic features and 10 additional coordinates obtained from the visual lip-feature space as described in section 4.2. The second bimodal system (Delta-AV) uses the same features as the AV-system plus an additional 10 visual "Delta-features" which estimate temporal differences in the visual features. The intuition behind these features is that the primary information in lip reading lies in the temporal change.

We generated several test sets covering the 346 letters: one set with clean speech, two with 10db and 20db SNR additive noise (recorded inside a moving car), and one set with 15db SNR crosstalk from another speaker uttering letters as well.

Table 1 summarizes our simulation results. For clean speech we did not observe a significant improvement in recognition performance. For noise-degraded speech the improvement was significant at the 0.05 level. This was also true of the crosstalk experiment which showed the largest improvement.

4.6 Large spontaneous speech dialog system

With this evidence that our lipreading technique is able to improve speech recognition performance, we are currently integrating the visual system in a larger spontaneous speech dialog system. The system serves as a restaurant guide for the local area. This project is a testbed for ideas in speech recognition, natural language research and related topics in our research lab. The user interacts with the system by making queries like "I would like to eat Chinese food not far from campus", and the system responds with suggestions or asks for additional information.

Our bimodal database consists of subjects of var-

Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499

ious ethnic and national backgrounds, representing a realistic mix of the current population in the United States. No special attempts were made to reduce office background noise, or to ensure exact head/lip positioning, in order to provide a realistic human computer interaction scenario. Experiments are in progress to train this larger system using the techniques discussed here.

4.7 Related Computer Lipreading Approaches

One of the earliest successful attempts to improve speech recognition by combining acoustic recognition and lipreading was done by Petajan in 1984 [14]. More recent experiments include [10, 20, 19, 5, 16, 12]. Most approaches attempt to show that computer lip reading is able to improve speech recognition, especially in noisy environments. The systems were applied to phoneme classification, isolated words, or to small continuous word recognition problems. Reported recognition improvements are difficult to interpret and compare, because they are highly dependent on the complexity of the selected task (speaker dependent/independent, vocabulary, phoneme/word/sentence recogntion), how advanced the underlying acoustic system is, and how many simplifications were made for the visual task (reflective lipmarkers, special lipstick, or special lighting conditions). We believe that our system based on learned manifold techniques and Hidden Markov Models is so far the most complete system applied to a complex speech recognition task but it is clear that many further improvements are possible.

5 Conclusion and Discussion

This paper can only begin to describe the many applications of manifold learning in vision. We have also not described certain hierarchical geometric data structures that can dramatically improve the performance of these techniques. We have shown how we are using them in the domain of lip reading and that they give significantly improved performance. It would be difficult to build traditional computer vision models of human lips and so the paradigm of building these models by learning is significant. Many lip reading research groups mark a subject's lips with special reflective tape, paint, or lipstick or wire the subject with strain gauges. The techniques described in this paper show that such artifices are unnecessary and that video images may be directly used for visual speech recognition.

Acknowledgments

We would like to thank Jerry Feldman, Yochai Konig, Nelson Morgan, Alex Waibel, and the ICSI Speech Group for their support and helpful discussions.

References

[1] H.A. Bourlard and N. Morgan.*Connectionist Speech Recognition, A Hybrid Approach.* Kluwer Academic Publishers, 1993.

[2] C. Bregler and S. Omohundro, *Surface Learning with Applications to Lip-Reading*, in Advances in Neural Information Processing Systems 6. Morgan Kaufmann Publishers, 1994.

[3] C. Bregler and S. Omohundro, *Nonlinear Image Interpolation using Manifold Learning* in Advances in Neural Information Processing Systems 7. MIT Press, 1995.

[3] H.A. Bourlard and N. Morgan.*Connectionist Speech Recognition, A Hybrid Approach.* Kluwer Academic Publishers, 1993.

[4] A.P.Dempster, N.M.Laird, D.B.Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society B, vol 39.*

[5] A.J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* Ph.D. Dissertation, School of Engineering and Applied Science of the George Washington University, Sep 10, 1993.

[6] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, *RASTA-PLP speech Analysis Technique*, in Proc. Int. Conference on Acoustics, Speech, and Signal Processing, San Francisco 1992.

[7] M. I. Jordan and R. A. Jacobs, *Hierarchical Mixtures of Experts and the EM Algorithm* Neural Computation, Vol. 6, Issue 2, March 1994.

[8] M. Kass, A. Witkin, and D. Terzopoulos, *SNAKES: Active Contour Models*, in Proc. of the First Int. Conf. on Computer Vision, London 1987.

[9] M. Kirby, F. Weisser, and G. Dangelmayr, *A Model Problem in Represetation of Digital Image Sequences*, in Pattern Recgonition, Vol 26, No. 1, 1993.

[10] K. Mase and A. Pentland. *LIP READING: Automatic Visual Recognition of Spoken Words.* Proc. Image Understanding and Machine Vision, Optical Society of America, June 1989.

[11] D.W. Massaro and M.M. Cohen, *Evaluation and Integration of Visual and Auditory information in Speech Perception.* Journal of Experimental Psychology: Human Perception and Performance, 9, 1983.

[12] J.R. Movellan *Visual Speech Recogntion with Stochastic Networks* in Advances in Neural Information Processing Systems 7, MIT Press, 1995.

[13] H. Murase, and S. K. Nayar *Learning and Recognition of 3-D Objects from Brightness Images* Proc. AAAI, Washington D.C., 1993.

[14] E. Petajan, B. Bischoff, D. Bodoff, and N.M. Brooke. *An Improved Automatic Lipreading System to enhance Speech Recognition.* ACM SIGCHI, 1988.

[15] T. Poggio and F. Girosi, *Networks for Approximation and Learning*, Proc. of IEEE, Vol. 78, No. 9, Sep. 1990.

[16] P. L. Silsbee *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition* Ph.D. Dissertation, University of Texas at Austin, May 1993.

[17] P. Simard, Y. Le Cun, J. Denker *Efficient Pattern Recognition Using a New Transformation Distance* Advances in Neural Information Processing Systems 5, Morgan Kaufman, 1993.

[18] M. Turk and A. Pentland *Eigenfaces for Recognition* Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

[19] G.J. Wolff, K.V. Prasad, D.G. Stork, and M.Hennecke *Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration.* in Advances in Neural Information

[20] B.P. Yuhas, M.H. Goldstein, and T.J. Sejnowski. *Integration of Acoustic and Visual Speech Signals using Neural Networks.* IEEE Communications Magazine.

[21] A. Yuille, *Deformable Templates for Face Recognition*, Journal of Cognitive Neuroscience, Volume 3, Number 1, MIT 1991.

Reprinted from: Proceedings of the Fifth International Conference on Computer Vision,
ed. W. Eric L. Grimson, June 1995, IEEE Computer Society Press, Los Alamitos, CA, pp. 494-499