Lip Modeling for Visual Speech Recognition

Ram R. Rao*

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332
rr@eedsp.gatech.edu

Russell M. Mersereau

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332
rmm@eedsp.gatech.edu

Abstract

In this paper, we describe an algorithm for modeling the shape of the mouth, and extracting meaningful dimensions for use by automatic lipreading systems. One advantage of this technique lies in the ability to normalize the model to compensate for scale and rotation. An error function is defined which relates the model to the image, and minimization of the error yields the best fit model. This is similar to deformable templates, but we attempt to perform the minimization in closed form. Visual only recognition was performed with features extracted from the model, and the recognition system achieved 85% accuracy on a two word discrimination task.

1 Introduction

Much research has been conducted to assess the benefits of incorporating visual lipreading information into automatic speech recognition systems. This work has demonstrated improvements in recognition accuracy, especially in the prescence of acoustic noise. Work by Finn and Montgomery [1] and Stork, Wolff, and Levine [2] has shown that the distances quantifying lip shapes form a meaningful parameter set for recognition. In both these studies, however, the distances were obtained from reflective markers, and therefore could not be part of a practical recognition system. Hence, a major problem which these optical recognition systems face is finding robust and efficient methods for extracting visual parameters. Recently, an automatic algorithm for visual feature extraction has been proposed by Prasad, Stork and Wolff [3].

In this paper, we will outline a method for modeling the lips and extracting recognition parameters from the model. Our modeling techniques are similar to deformable templates [4]. One of the key aspects underlying deformable templates is an energy function that relates the template parameters to the image. The minimization of this energy function yields the "best fit" template. Problems with deformable templates arise in deriving an energy function that accurately reflects the match between the template and the image. Furthermore, the minimization can be computationally intensive.

In our method, we choose a small model for the lips which consists of four parabolas. This model was selected since it was determined that four features occur most frequently in lips – the upper edge of the upper lip, and the upper, middle and lower edges of the lower lip. The problem lies in partitioning the pixels of the image to their corresponding edges, and for each partition, finding the parabola which minimizes a mean squared error criterion for that partition. Hence, the mean squared error criterion serves the same purpose as the energy function of deformable templates. Furthermore, we can define the mean squared error in a way that allows for a closed form solution to the problem.

2 Preprocessing

We are mainly interested in the edges of the lip images, so the first step is producing an edge image. Morphological operators can be used, but we choose to use a simple 3 by 3 linear filter for this purpose. Since the lips are a predominantly horizontal feature, we use a filter aligned to extract horizontal edges. This edge image is then averaged with a 3 by 3 window to reduce the effects of noise. One advantage of using a linear filter over a morphological filter is that linear edge detectors result in images with positive and negative values whereas morphological edge detectors

*This work is supported by the U.S. Army Research Office, Contract DAAL03-92-G-0068.

1058-6393/95 $4.00 © 1995 IEEE

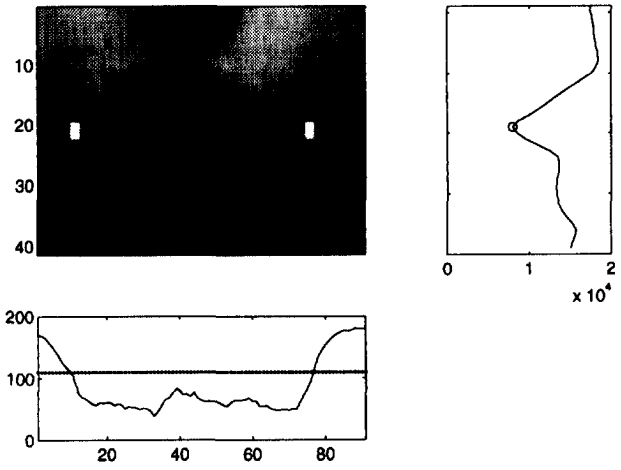

Figure 1: Shown is a sample image, along with the sum of the rows, and the intensity profile along the row with minimum sum.

generally result in positive valued images. Thus, information contained in the sign of an edge can be used for labeling purposes.

Samples of the input image and the edge image are shown in Figures 1 and 2. For the edge image, the negative values are dark and the positive values are bright. Note that the top edge of the upper lip is a light to dark transition, vertically, as is the middle edge of the lower lip.

3 Modeling Procedure

Suppose we have a set of points,(x, y), and wish to find the coefficients of the parabola, ax^2+bx+c, that minimizes the mean squared error, $E[(y-(ax^2+bx+c))^2]$. The coefficients a,b and c should satisfy:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} E[x^4] & E[x^3] & E[x^2] \\ E[x^3] & E[x^2] & E[x] \\ E[x^2] & E[x] & 1 \end{bmatrix}^{-1} \begin{bmatrix} E[x^2y] \\ E[xy] \\ E[y] \end{bmatrix} \quad (1)$$

One modeling strategy could have an algorithm pick the peaks and valleys corresponding to each edge, and use these points to define a, b and c. This, however, would require a robust peak picking and labeling strategy. It would also ignore the value associated with each pixel along the edge.

Instead, our first step is to derive rough initial estimates for the coefficients of the parabolas corresponding to the top of the upper lip, and the middle of the lower lip. We start by finding the vertical position of the center of the mouth. This is done by examining

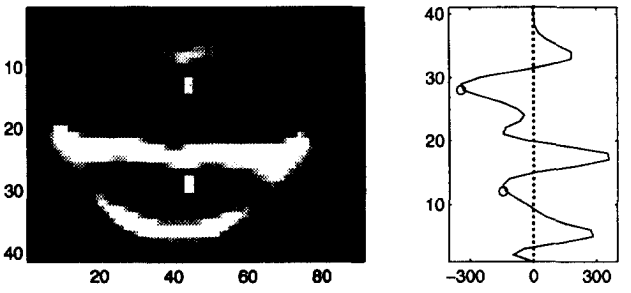

Figure 2: Shown is the edge image obtained from sample image. Valleys from vertical profiles are chosen as initial estimates for points on a particular edge.

one of three statistics in the original mouth image: the sum of each row, the minimum value in each row, or the maximum value minus the minimum value in each row. The vertical position of the center of the mouth can be defined to be the minimum point of the first two statistics or the maximum point of the third. We use the first statistic which provides good results. It is, however, susceptible to lighting variations due to shadows and irregular illumination. The row of pixels passing through the center of the mouth is then analyzed to find the corners of the mouth. If a threshold is set at the average of the maximum and minimum values in the row, the leftmost and rightmost pixels with values under this threshold are defined to be the corners of the mouth (see Figure 1).

We next calculate which columns are one-quarter, one-half, and three-quarters the distance between the corners of the mouth. For each of these columns we find a point on the top of the upper lip, and the middle of the lower lip by peak picking along the columns of the edge image (see Figure 2). This yields a set of five points per edge (the 2 corners of the mouth, and the 3 additional points). Using these five points, we can estimate a, b, and c for the top of the upper lip, and the middle of the lower lip using Equation 1.

Since this method will result in crude estimates of the two parabolas, we would like to be able to refine our estimates. To accomplish this, we look at each of the parabolas individually. We dilate the parabola with a 9 by 9 square, and use this to partition our edge image. Within this partition, points with intensity less than a prescribed threshold are removed, and the remaining points are given a probability proportional to their intensity. The points in the partition and their corresponding probabilities are used to compute the expectations necessary for Equation 1, and the coefficients of the parabola are recomputed. Thus, the "best fit" parabola is attracted to the high intensity

		Out			
		Initial		Modeled	
		wow	mom	wow	mom
I	wow	93%	7%	100%	0%
n	mom	34%	66%	31%	69%

Table 1: Recognition Results

	Initial	Modeled
Intracluster distance – wow	466	136
Intracluster distance – mom	798	331
Intercluster distance	944	395

Table 2: Cluster Statistics

edge pixels. This procedure of partitioning and reestimation is iterated until the intersections of the two parabolas remain at the same locations. Normally, 3 to 5 iterations yields good performance. We can now associate the intersection of the two parabolas with the corners of the mouth.

The top of the lower lip is then defined by searching for the valleys above the middle of the lower lip. Likewise, the bottom of the lower lip is found by searching for the valleys below the middle of the lower lip. These points are used in an equation similar to (1) except the minimization is constrained such that these parabolas pass through the corner of the lips. Figure 4 shows sample results obtained from our modeling procedure.

4 Recognition Experiments

A simple recognition experiment was performed to test the performance of our modeling system. Seven utterances of the words "wow" and "mom" were captured, and the images were cropped to a 100 by 80 pixel region containing the mouth. Each utterance consisted of 28 frames. For each frame, two sets of parameters were measured. The first set consisted of horizontal and vertical separations derived from the initial estimation procedure. The second set of parameters was derived after these initial estimates were revised through the modeling procedure. The first modeled feature was the horizontal distance between the corners of the mouth. The second was the maximum vertical separation between the top of the upper lip and the middle of the lower lip. Thus, each utterance was represented by vectors of dimension 28 by 2.

Dynamic time warping was used to derive a distance between each pair of utterances. The DTW algorithm allowed for insertion into one token or the other, but two consecutive insertions were not allowed.

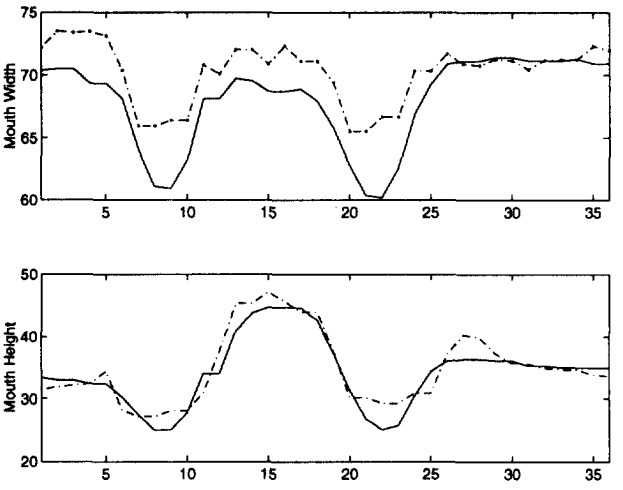

Figure 3: Time warped tokens for the words, "wow" (solid) and "mom" (dashed).

The optimal warping path was evaluated using the Viterbi algorithm, and the total mean squared error distortion between time warped tokens was accumulated.

The recognition experiment consisted of considering each of the 49 possible pairs of templates, and classifying the remaining 12 tokens by comparing against the templates. Recognition was done on the initial feature set and the modeled feature set. The results of this experiment are shown in Table 1.

First, it can be seen that there is an 80% overall recognition rate for the initial set of features, and an 85% overall recognition rate for the modeled features. This shows that visual information is useful for recognition purposes. Looking at the sample tokens in Figure 3, we can see that although the vertical separation trajectories for the words "wow" and "mom" are similar, there is a significant difference in the horizontal separation trajectories.

Next, the recognition rates are seen to be asymmetric. There are more misclassifications of the word "mom". This can be explained by looking at the intracluster distances for the two words as shown in Table 2. The average distance between tokens of the word, "wow", is much smaller than the average distance between tokens of the word, "mom". Furthermore, the intracluster distance of the word "mom" is quite close to the average intercluster distance.

Finally, although there was not a tremendous difference between the recognition rates achieved by the initial and the modeled parameters, it was shown that modeling reduces the intracluster distances by a fac-

tor of 2 or 3. This shows that modeling provides a more reliable parameter set. The effects on recognition rates would probably be more noticeable with a larger vocabulary set.

5 Conclusion

This modeling procedure works well for most of the images tested. It consistently finds edges and adjusts to their shape. Problems arise when the initial estimates cause the model to be attracted to an improper edge (such as one caused by teeth). Problems can also arise when the model is attracted to two edges at the same time.

Finally, the procedure is not scale independent. When estimated parabolas are dilated with an 9 by 9 square, the size of the image is constrained to be within a certain range. If the lips are too small, many edges will be caught in the new partition, and the new estimates will be incorrect. If the lips are too large, then the dilation will not include enough new pixels to allow for fast convergence. This can be corrected by increasing the size of the dilating element, but this also increases the computation.

In some sense, our modeling system bridges the gap between deformable template based approaches to feature extraction, and syntactic (or peak picking) based approaches to feature extraction. We start with a certain amount of syntactic information and refine our estimates with a minimization procedure. Our procedure works extremely well if each individual edge can be isolated and labeled properly, and it has the ability to compensate for small errors made in the labeling stage.

Finally, it was seen that this modeling procedure produces excellent recognition results for a two word discrimination task, and produces a feature set that is more reliable than those obtained through the initial peak picking procedure.

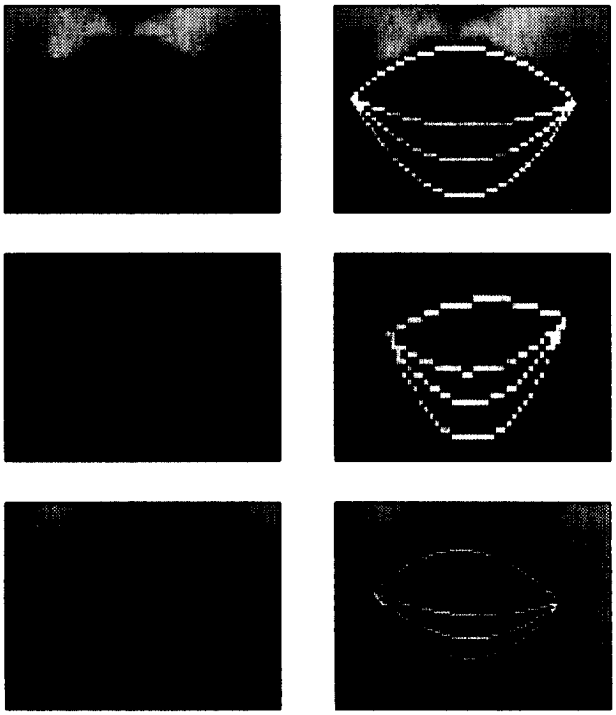

Figure 4: Sample results

References

[1] K. Finn and A. Montgomery, "Automatic optically-based recognition of speech," *Pattern Recognition Letters*, vol. 8, no. 3, pp. 159–164, 1988.

[2] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *International Joint Conference on Neural Networks*, pp. 285–295, 1992.

[3] K. V. Prasad, D. G. Stork, and G. J. Wolff, "Preprocessing video images for neural learning of lipreading," Tech. Rep. CRC-TR-9326, Ricoh California Research Center, September 1993.

[4] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 99–111, 1992.