A MULTIPLE DEFORMABLE TEMPLATE APPROACH FOR VISUAL SPEECH RECOGNITION

Devi Chandramohan
Peter L. Silsbee

Dept. of Electrical and Computer Engineering
Old Dominion University

ABSTRACT

In this paper, we propose an improved deformable template algorithm for modeling the shape of a talker's mouth. We use a two step approach which begins by classifying mouth images into broad categories. The classification procedure yields both a set of template parameters (in effect, a unique template) and a set of initial conditions. The second step is to allow the deformable template to converge using standard techniques. The multi-model approach is significantly more flexible than single-model approaches and consistently provides better solutions. We present examples of single and multiple template solutions which support this statement. In a small recognition experiment, recognition of consonants improved from 16% to 33%, based only on visual information, when multiple templates were used.

1. INTRODUCTION

The use of deformable templates to locate oral cavity features for visual speech recognition has recently become quite popular. These templates generally consist of a topologically constrained set of geometric features (e.g., a set of linear segments constrained to form a simple closed contour) which is matched to an image using some set of criteria. Typically, contours of the deformable template are matched to edges in the image. Templates often have "default" configurations towards which the solutions are biased. Matching the image to the template is a constrained optimization problem which can be solved by gradient descent. However, deformable templates, while quite flexible, can be sensitive to several factors, including parameter initialization and lighting variations. Furthermore, it is not always possible for a single template to accurately describe the entire range of possible images.

In this paper, we present a two step deformable template approach which uses multiple templates and an automatic initialization procedure to alleviate these problems. The first step consists of a coarse classification of the image which yields both a choice of template and an initial set of parameters for that template. This is followed by a standard solution (fitting) of the template.

We have experimented with several different sets of templates. The simplest set contains only an open-mouth template and a closed-mouth template. All templates have the same topological structure — each is represented as a simple collection of joined line segments. They differ only in their default configurations and in the set of biasing "springs" which tend to force the templates towards their default configurations.

The model selection procedure involves a set of idealized (archetype) images. Each template is associated with one such image, which dictates the default configuration of that template. The image being analyzed is compared to each archetype image; the comparison yields both a matching score and a template position. The template associated with the best matching archetype is selected, and initialized according to the position from the image match. Gradient descent is then used to fine tune the template parameters.

The use of a crude match followed by fitting of the template offers the following advantages:

- The deformable models can be less general, and therefore easier to fit to specific mouth shapes.
- The matching process provides a very good initial estimate for the model parameters. Convergence is fast and more likely to reach the correct "energy" minimum.

2. DEFORMABLE TEMPLATES FOR VISUAL SPEECH RECOGNITION

Deformable templates have been widely used for visual speech recognition [3, 4, 5, 6, 8, 10, 12, 13]. This class of models is well suited for analysis of objects whose contours undergo continuous nonrigid deformation. Typical deformable template models consist of a parameterized set of curves, a set of geometric constraints, and a set of rules for fitting the curves to features in an image. The constraints and rules are generally encapsulated in an "energy" or "penalty" functional which assigns a numerical value to every possible configuration of the template. The penalty depends on the image as well as the template parameters.

The penalty functional typically assigns two types of penalties. High penalty values are assigned to configurations in which template segments do not lie near image features of interest. Configurations which are far (in some sense) from a default configuration also receive high penalties. These latter penalties arise from the geometric constraints of the model, which serve to regularize the solution process.

Design of deformable templates often becomes a balancing act between accomodating variability on the one hand, and allowing spurious solutions on the other. If a subject is capable of undergoing extreme deformations, then the template should be able to undergo the same deformations without a high penalty — but this in turn means that, for a given image, there are likely to be many incorrect solutions with low penalties.

3. MULTIPLE TEMPLATES

3.1. Pixel Based Approaches

Another class of systems uses pixel based approaches to visual speech recognition. Most of these are based on neural networks, with image pixels as inputs (e.g. [1, 7, 11]). Silsbee [9] used a direct matching approach, which we extend in this paper. An advantage of such approaches is that they are less prone to making unrecoverable errors in a "feature extraction" stage [2]. A disadvantage is that they can be quite susceptible to variability in the image acquisition process.

The method described here is a hybrid method in which an initial, coarse classification is performed based on pixel matching, and a deformable template is selected and solved based on the results of the coarse classification.

3.2. Image Classification

The image classification procedure is essentially the same as outlined in [9]. Each image first undergoes two preprocessing operations designed to reduce variability due to lighting conditions. We do not describe the preprocessing steps here; see [9] for details.

The matching algorithm is as follows. We begin with a set of K "archetype" images $\{A_1, \ldots, A_K\}$, which represent a broad range of oral shapes. Synthetic images could be used as archetypes; however, we have chosen them from the set of training images.

For every image frame I being analysed the system chooses the archetype A_k which minimizes a mismatch measure.

The mismatch is calculated hierarchically: for each $N \times N$ image, a resolution pyramid of smaller images ($N/2 \times N/2$, $N/4 \times N/4$, etc.) is formed, where a pixel at each resolution level is the average of four neighboring pixels at the next higher resolution. We define the lower resolution images

$$I^{(n)}(i,j) \quad = \quad \text{ave}\{I(l,m) \ : \ 2^n i \leq l < 2^n(i+1), \quad (1)$$

$$2^n j \leq m < 2^n(j+1)\}$$

Where $I(i,j)$ refers to the pixel value in the i^{th} column and the j^{th} row. Beginning at the lowest resolution, the average absolute difference between pixels is calculated for several different possible alignments. If one image is shifted by p pixels vertically and q pixels horizontally, the local error is defined

$$E^{(n)}(A_k, I, p, q) = \sum_{i=i_0}^{i_n} \sum_{j=j_0}^{j_n} \frac{|A_k^{(n)}(i,j) - I^{(n)}(i-p, j-q)|}{(M^{(n)} - p)(M^{(n)} - q)} \quad (2)$$

where $i_0 = \max(0, p)$, $j_0 = \max(0, q)$,
$i_n = M^{(n)} - \max(1, 1-p)$, $j_n = M^{(n)} - \max(1, 1-q)$, and the size of $I^{(n)}$ is $M^{(n)} \times M^{(n)}$ pixels. Evaluation of this quantity for each (p, q) pair and each archetype image yields a score and an alignment for each image. The alignment corresponding to the lowest score is saved, converted to the appropriate coordinates for the next higher resolution image, and used as a starting point for the search at that resolution. Once the best alignment is found at the highest resolution, the image is classified according to the lowest scoring archetype. The search space is limited at each resolution to just a few pixels about the starting point. The exact choice of search space represents a tradeoff between computation time, ability to handle extreme translations within the images, and a risk of spurious feature matches.

3.3. Template Fitting

The image classification step returns two pieces of information. The first is the class to which the image belongs, which, in turn, is associated with a particular deformable template. The second is the alignment, which can be used to provide the template with a good initial estimate of its parameters.

The template used in our system is shown in Fig. 1. The solid lines in the figure represent template segments which are attracted to particular types of edges — all versions of the template seek configurations such that the interior of the contour is dark, and the exterior is light. Also shown (dashed lines) is the set of possible *springs*. These may or may not be present, depending on the model; they serve to bias the template to a particular default configuration.

Once a particular model has been selected, the parameters (the coordinates of the twelve points which define the template) are initialized to the default positions associated with that model, taking the alignment information into account. The template is then allowed to converge to a local minimum of the penalty function using gradient descent.

4. EXAMPLES

The primary purpose of this section is to provide a comparison of template solutions for single and multiple template approaches. Fig. 2 shows template outlines after convergence, superimposed on the image analysis, for a single template ap-

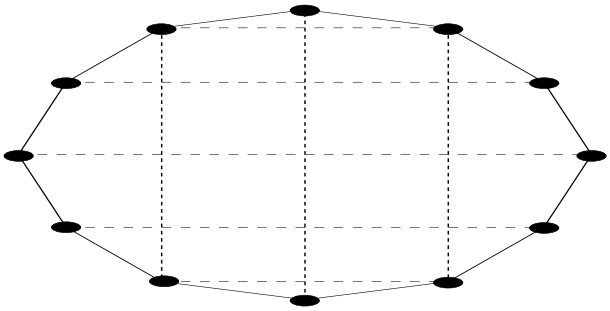

Figure 1: The generic template used in the system. Solid lines represent template segments, dashed lines represent springs. All template models share the same set of segments; models differ in which springs are active and in their default configurations.

# templates	1	6
% correct	16	33

Table 1: Percent correct classification for different sets of templates.

proach. Many of the templates show poor fits to the image. This is due partly to poor initialization. A more important effect is that there is no appropriate default configuration for the variety of mouth shapes seen. A template which works well for a variety of open mouth shapes is not likely to work well when the mouth is nearly closed, for example. Fig. 3 shows the same images, with template outlines superimposed after solution using a multiple template approach. The template segments clearly fit the lip edges much better when multiple templates are used.

Of course, the ultimate criterion for success is recognition of the utterances. Although we have primarily focused on validating the templates by visual inspection of images such as those shown in Fig. 3, we have also performed a few recognition experiments. The task consisted of 22 consonants uttered in an /a/-C-/a/ context. Six tokens of each consonant were used for training and four for testing. Each utterance was acquired as a 30 frame per second video sequence (lengths ranged, for the most part, from 20 to 30 frames). Each frame was 80×80 pixels of 8-bit grayscale.

The features used for the recognition experiments are simply estimates of the height and width of the oral cavity, derived from the template segment-end coordinates after convergence.

Table 1 shows the recognition rates, based on visual data only, for 1 and 6 templates. In these pilot experiments, at least, there is a significant improvement in recognition when multiple templates are used.

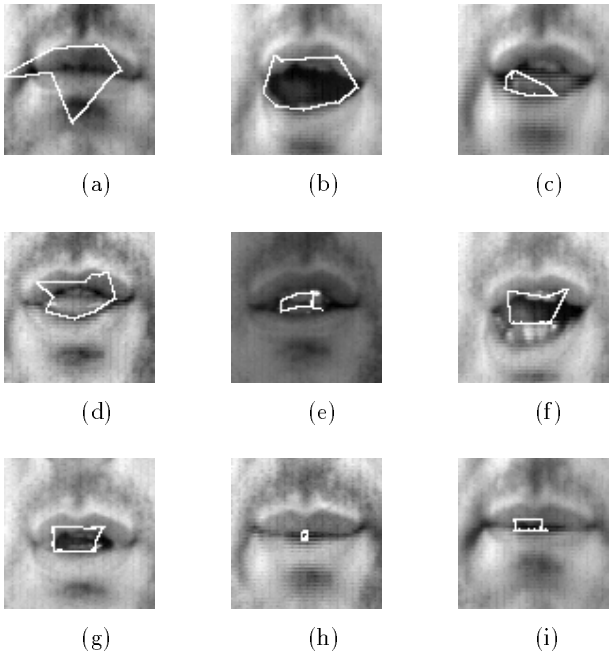

Figure 2: Examples of mouth images and template solutions using a single template. Poor initial estimates and insufficient constraints can cause incorrect solutions.

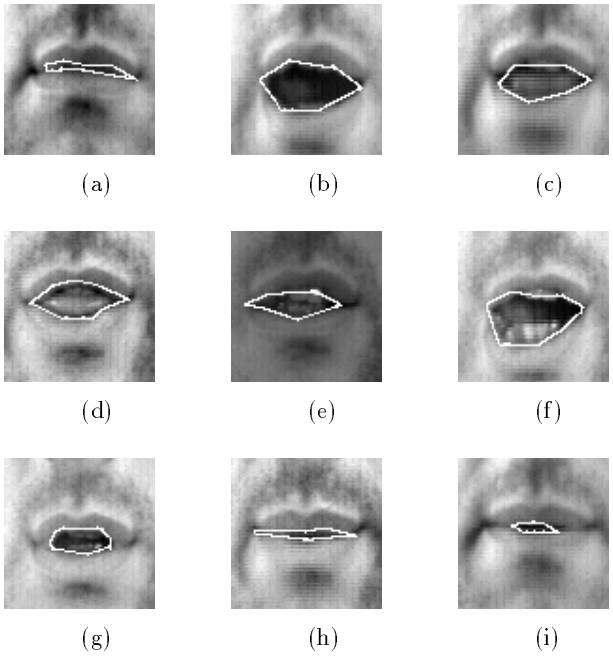

Figure 3: Multiple model solutions for the images of Fig. 2. Solutions clearly provide a better fit to the outline of the oral cavity than does the single template of Fig. 2.

5. DISCUSSION

We have presented a two step approach for fitting deformable templates to mouth images. The method possesses several advantages relative to single template techniques. Initialization is greatly improved, yielding better solutions in significantly less time. Better solutions are also obtained since the templates can be more finely tuned to a smaller class of images. Because each template has the same topological structure, features can be obtained in the same way from every template.

It should be noted as well that the initial classification of images does *not* represent a "hard decision" which could lead to unrecoverable errors. In fact, all the templates are capable of assuming the same configurations; even if the model and initialization are chosen poorly, there is a strong probability that a reasonable fit can be obtained.

Much can be done to refine this approach. A tradeoff exists between the number of models, computation time, and quality of solutions. The matching and alignment procedure currently consumes a significant portion of the overall computation, and could be made more efficient by narrowing the search to just a few archetypes at high resolutions.

Currently, the models vary only in the default positions and the presence or absence of certain springs. It is also possible to provide springs of differing strength, to further refine the models. We also expect to relax the constraint that all models should have the same topology. In fact, any set of models from which a fixed set of parameters can be derived could be appropriate for this framework.

6. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IRI-9409851.

REFERENCES

1. Christoph Bregler, Stephen M. Omohundro, and Yochai Konig. A hybrid approach to bimodal speech recognition. In *Twenty-Eighth Annual Asilomar Conf. on Signals, Systems, and Computers*, pages 556–560, November 1994.
2. Christoph Bregler, Stephen M. Omohundro, Jianbo Si, and Yochai Konig. Towards a robust speechreading dialog system. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
3. Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile. 2D deformable models for visual speech analysis. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
4. Barney Dalton, Robert Kaucic, and Andrew Blake. Automatic speechreading using dynamic contours. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
5. Marcus E. Hennecke, K. Venkatesh Prasad, and David G. Stork. Using deformable templates to infer visual speech dynamics. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 578–582, Pacific Grove, CA, November 1994. IEEE, IEEE Computer Society Press.
6. Juergen Luettin and Neil A. Thacker adn Steve W. Beet. Active shape models for visual speech feature extraction. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
7. Javier Movellan and George Chadder. Channel separability in the audio-visual integration of speech: Implications for engineers and cognitive scientists. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
8. R. R. Rao and R. M. Mersereau. Lip modelling for visual speech recognition. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 587–590, Pacific Grove, CA, November 1994.
9. P. L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, University of Texas, 1993.
10. Peter L. Silsbee. Motion in deformable templates. In *First IEEE Intl. Conference on Image Processing*, volume 1, pages 323–327. IEEE, November 1994.
11. David G. Stork, Greg Wolff, and Earl Levine. Neural network lipreading system for improved speech recognition. In *Intl. Joint Conf. on Neural Networks*, pages 285–295, 1992.
12. Michael Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
13. Alan L. Yuille, David S. Cohen, and Peter W. Hallinan. Facial feature extraction by deformable templates. Technical Report 88-2, Harvard Robotics Laboratory, 1988.