Using Deformable Templates to Infer Visual Speech Dynamics

Marcus E. Hennecke
Dept. of Electrical Engineering
Stanford University
Stanford, CA 94305
marcush@leland.stanford.edu

K. Venkatesh Prasad
Ricoh California Research Center
2882 Sand Hill Road #115
Menlo Park, CA 94025-7022
prasad@crc.ricoh.com

David G. Stork
Ricoh California Research Center
2882 Sand Hill Road #115
Menlo Park, CA 94025-7022
stork@crc.ricoh.com

Abstract

The visual image of a talker provides information complementary to the acoustic speech waveform, and enables improved recognition accuracy, especially in environments corrupted by high acoustic noise or multiple talkers. Because most of the phonologically relevant visual information is from the mouth and lips, it is important to infer accurately and robustly their dynamics; moreover it is desirable to extract this information without the use of invasive markers or patterned illumination. We describe the use of deformable templates *for speechreading, in order to infer the dynamics of lip contours throughout an image sequence. Template computations can be done relatively quickly and the resulting small number of shape description parameters are quite robust to visual noise and variations in illumination. Such templates delineate the inside of the mouth, so that the teeth and the tongue can also be found.*

1 Introduction

Whereas the speech chain, in its most general form, involves the production and perception of *both* the acoustic (A) and visible (V) components of speech, automatic speech recognition (ASR) research has traditionally ignored the role of the V component. Recently, however, several groups have incorporated visual subsystems into their ASR systems (Petajan, Bischoff & Bodoff [3], Yuhas, Goldstein, Sejnowski & Jenkins [6], Pentland & Mase [2], Stork, Wolff & Levine [4], Wolff, Prasad, Stork & Hennecke [5]). From such work, and that of several other groups, it is clear that whether by Hidden Markov Models, neural networks, or statistical pattern recognition approaches, visual information can improve speech recognition. However, since most of this research has employed special markers on the face, or chromatic lipstick, standard grayscale video input cannot be used easily. Hence it is now time to address more seriously the image feature extraction problem, and its unique aspects relevant to speech recognition.

We present here visual preprocessing that enables the use of grayscale images while at the same time incorporating *dynamical* processes in a natural and computationally efficient way. Our method employs *deformable templates* — parameterized descriptions of the object to be tracked (Yuille, Cohen & Hallinan [7]). In our work, the object to be tracked is the talker's mouth and the template consists of parabolas and quartics which follow the outer and inner edges of the upper and lower lips. The method can be used for a number of other attributes, such as jaw rotation (chin position), but here we present only our results on lips. The final shape parameters themselves provide input features to a recognition engine.

Deformable templates have several advantages over *snakes* (Kass, Witkin & Terzopoulus [1]) and some other contour finding schemes, and this makes them well suited for application to speechreading:

- The computations required can be done relatively quickly.
- Information from many pixels is averaged throughout space and over time to give the best fit, including those along the edges of the template. Fits are thus quite robust.
- Heuristic information — such as maximum rates of deformation, physical constraints such as (approximate) spatial symmetry, etc. — can easily be incorporated into a global "fit" or cost function, increasing the likelihood of a correct fit.
- Deformable templates help in segmenting the mouth opening from the rest of the image, and

1058-6393/95 $4.00

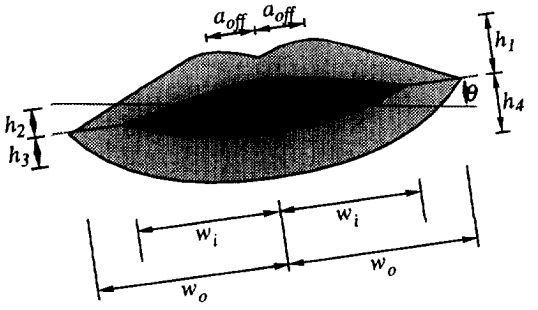

Figure 1: The mouth template consists of two parabolas for the inside edges and three quartics for the outside edges. Including the center coordinates and the tip angle there are a total of 12 parameters describing the shape, location and orientation of the template.

thereby aid in the detection of the tongue and teeth. These features are, of course, useful for visual speech recognition as well.

2 The deformable template

A deformable template can be understood as a model of an object. By tuning a set of parameters, the template can be deformed to match the object in some optimal way, as defined by a cost function. If the cost function is at a minimum we say that a fit has been found and the parameters describe the object.

Figure 1 shows the template we used for the lips. The template is designed so that it is able to appromixate a large number of mouth shapes as closely as possible while at the same time keeping the number of parameters to a minimum. After some experimentation the template shown in the Figure has proven useful.

The template consists mainly of two types of curves: parabolas and quartics. They are centered around the coordinates (x_c, y_c) and rotated by the tip angle θ, which are found while adapting the template to the mouth. The center coordinates and the tip angle define the mouth coordinate system. Within this shifted and rotated coordinate system the parabolas and quartics are constrained to be symmetric, and to intersect the axes at $(-w, 0)$, $(0, h)$, and $(w, 0)$. A parabola of height h and width $2w$ is described by the following equation:

$$y = h(1 - \frac{x^2}{w^2}),$$

and the equation for the corresponding quartic is:

$$y = h(1 - \frac{x^2}{w^2}) + 4q(\frac{x^4}{w^4} - \frac{x^2}{w^2}).$$

The parameter q in the equation for the quartic determines how far the quartic deviates from a parabola. It serves as an additional shape parameter describing the deformation of the corresponding edge.

We now have all the tools to assemble the mouth template:

1. The center coordinates (x_c, y_c) and the tip angle θ define the location and orientation of the mouth coordinate system. All lengths and measurements are in reference to that coordinate system and the mouth is assumed symmetric to its ordinate.

2. Two parabolas make up the inside edges of the mouth. Each parabola has its own height h_2 and h_3, but they share the same width w_i.

3. The outside edge of the lower lip is modeled by a quartic. It has been our experience that the flexibility that the additional parameter provides allows a quartic to adapt itself to a larger variety of talkers and to track the lower lip more accurately than a simple parabola.

4. For the outside edge of the upper lip two quartics are used. Their centers are offset by $\pm a_{off}$ from the ordinate and they are constrained to share the same parameters. This ensures symmetry of the template.

3 Cost function

Adapting the template to match an object involves changing the parameters to minimize a certain cost function. There are no restrictions on the choice of the cost function, however, if one uses gradient descent to find the minimizing set of parameters, the function needs to be continuous and a gradient has to exist. Usually the cost function is defined in the following way:

1. One or more *potential fields* are computed from the image. They usually represent certain local features such as edges, valleys or peaks. These fields can be blurred to ensure continuity for the gradient descent algorithm.

2. The cost function is then a sum of integrals over these potential fields either along or in between

the curves of the templates. For example, if the curves of the templates represent the edges of the object, one term could be the integral over the edge field along the curves of the template.

3. In addition, heuristics and higher level information can be incorporated via penalty terms which constrain the parameters to stay within certain reasonable limits.

Because our application involves processing a large number of images, it was our goal to design a cost function that is both effective and computationally simple. In our experiments we found that a large proportion of the time was spent on computing the potential fields. In contrast, the penalty terms of the cost function were almost negligible in terms of the complexity of the algorithm. Thus, our cost function consisted of the following:

Potential fields The edges of the lips are among the most salient features of the mouth. Since the edges are mostly horizontal, we used the vertical gradient of the image as potential field (the vertical gradient is most sensitive to horizontal edges). For the kernel, we used the horizontal 3×3 Prewitt kernel:

$$\frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}.$$

The advantage of using this gradient is not only that it is most sensitive to the relevant edges, it also differentiates between "positive" and "negative" edges. A positive edge is one where the image intensity is higher above the edge than below it, and vice versa for a negative edge. A template that makes use of this difference will not confuse the upper and the lower lip.

In order to improve long range interactions, the gradient was blurred by a 5×5 exponential kernel. The kernel is maximal at the center and trails off exponentially with the distance from the center.

Integrals The cost function then consists of four curve integrals, one for each of the four lip edges. Taking into account positive and negative edges, it takes on the following form:

$$\begin{aligned} E &= -\frac{c_1}{|\Gamma_1|}\int_{\Gamma_1}\Phi_e(\vec{x})ds - \frac{c_2}{|\Gamma_2|}\int_{\Gamma_2}\Phi_e(\vec{x})ds \\ &\quad + \frac{c_3}{|\Gamma_3|}\int_{\Gamma_3}\Phi_e(\vec{x})ds + \frac{c_4}{|\Gamma_4|}\int_{\Gamma_4}\Phi_e(\vec{x})ds \end{aligned}$$

where the c_i are coefficients that allow one to place different weights on the curves, Γ_i are the four curves making up the mouth template, $|\Gamma_i|$ are the lengths of the curves and $\Phi_e(\vec{x})$ is the potential field.

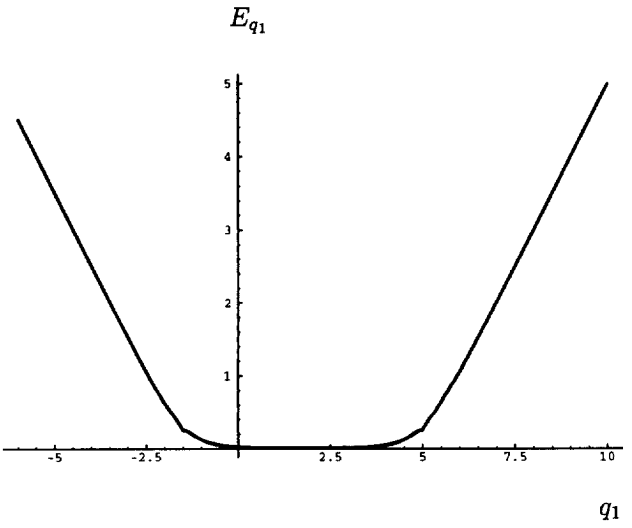

Figure 2: The penalty term for the parameter q_1. If the parameter is contained within the specified range, the penalty term is negligible. As the parameter moves closer to the limits the penalty term grows larger, pushing the parameter back to within its valid range.

Penalty terms In addition to the curve integrals, the cost function also consists of a number of penalty terms. There are two types of penalty terms, one introduces spatial constraints, the other temporal constraints. The spatial constraints make sure that the parameters of the template stay within reasonable limits. Figure 2 shows how we included spatial constraints in the cost function. A penalty term was added to the cost function for each spatial constraint to keep the template parameters within reasonable bounds. These bounds were determined heuristically and sometimes depended on other parameters. For example, one constraint ensured that $h_1 > h_2$. This means that the lower bound for h_1 depends on h_2 and the upper bound for h_2 depends on h_1.

Besides the spatial constraints we also included one temporal constraint. For each image sequence the average thickness of the lips was determined. The spring forces were added to keep the lip thickness close to the mean. This does allow variations of the thickness as they occur in utterances such as /ba/. There is also some physical motivation to use spring forces as they roughly model the lip deformations. The term for the

Figure 3: Fitting of a template to the mouth. Left: The initial state placed the template in the center. Middle: after several iterations reducing a global cost E the template has moved closer to the mouth. Right: In the final state the template approximates the mouth.

temporal constraint then looks like this:

$$E_{temp} = k_{12}((h_1 - h_2) - \overline{(h_1 - h_2)})^2 + k_{34}((h_3 - h_4) - \overline{(h_3 - h_4)})^2$$

where k_{12} and k_{34} are the spring constants and $\overline{(h_1 - h_2)}$ and $\overline{(h_3 - h_4)}$ are the mean thickness of the lips averaged over several frames.

Adapting the deformable template to a particular image means finding a set of parameters that minimizes the cost function. In our work, we used gradient descent to find a (possibly local) minimum. After the parameters have been updated to minimize the cost function, they can be scaled and directly entered into the speech recognition engine. The center coordinates and the tip angle do not convey significant speech related information and are not used (although they might indicate stress or emphasis). Hence, a total of nine parameters are extracted for recognition: $w_i, w_o, a_{off}, h_1, q_1, h_2, h_3, h_4, q_4$.

4 Procedure

The focus of our work is the talker's mouth. Before the initial templates can be found, another image processing algorithm determines the region of interest (ROI) around the mouth. We have shown robust automatic ROI detection of mouths elsewhere (Wolff et al., [5]), but in our work here the region was determined manually.

Finding the ROI was key to placing the template in close vicinity of the mouth. Figure 3 demonstrates how a deformable template adapts to an image. In the figure we used a simplified template consisting only of four parabolas which share the same width. The fitting process starts with a generic set of parameters that place the template near the mouth. The template is placed in the center of the ROI and oriented horizontally ($\theta = 0$). The initial heights and width

Figure 4: Three frames from the utterance /po/. The left frame shows the mouth in its initial resting position, in the next frame the lips have reached maximum closure and in the third frame maximum opening.

of the parabolas have been determined heuristically to match an "average" mouth shape, scaled to two-thirds the width of the ROI. After several gradient descent iterations, the template has started to move closer to the mouth and approximate its shape. In the final state the template has fit itself to the shape of the mouth.

In our experiments we found that the magnitude of the cost function varied significantly with speakers and lighting conditions. Thus, we made termination of the template matching dependent on the total change of parameters. If it dropped below a certain threshold, adaptation was stopped. Alternatively, if the gradient descent had exceeded a certain number of iterations, the adaptation was also stopped.

For measuring speech dynamics we have to look at image *sequences*. The procedure for image sequences is very similar to the one described above. For the first image (frame 1), the template is placed near the mouth and then allowed to fit itself around the mouth. The final state for frame 1 is then used as initial state for the next image, frame 2. This helps ensure temporal continuity.

After running the adaptation process over the entire image sequence, the parameters of the final states are scaled and normalized and can then be used for recognition.

5 Results

During development of the algorithm several image sequences were used to find an optimal template, potential field and cost function. The sequences were taken from one speaker in three different lighting conditions and another speaker in a fourth lighting condition. Because the algorithm has been optimized to those sequences it does quite well on them (see Figure 4). Tracking of the widths and heights of the lips was excellent. Only the lower edge of the lower lip

posed some problems. This edge had the highest degree of variability between speakers and lighting conditions and even within the same image sequence. Depending on the particular angle of the incoming light and the position of the jaw the edge could switch from negative to positive and vice versa. Often the chin midpoint provided a much stronger attractor for the template than the edge of the lip.

We also tested the algorithm on a set of ten speakers and five image sequences each. The sequences consisted of five repetitions of the same consonant-/a/ utterance. The utterances used were /ba/, /da/, /fa/, /la/, and /ma/. Overall, we tested adaptation of the template on 50 image sequences with a total of 250 utterances.

We observed that tracking of the heights of the lips was usually very satisfactory and remarkably stable even under unfavorable lighting conditions. Sometimes the teeth provided distractors for the inside edges. Only in three sequences did the template move away from the mouth.

While the heights were tracked well, determining the widths of the lips was much harder. This is especially the case for people with 'non standard' mouth shapes, i.e., shapes that require parameter sets far from those used during development of the algorithm.

6 Conclusions

We have shown that deformable templates provide a useful way to find and track the lip movements of a talker. Designing the global cost function is crucial in guaranteeing the success of our algorithm. With a good cost function, tracking can be done with high accuracy while keeping the computational requirements at reasonable levels.

In future work we plan to increase the accuracy of the width estimates and to make the template more robust against the teeth. We are also investigating new features such as the presence of the tongue and lip protrusion.

Acknowledgements

We would like to thank Greg Wolff for useful discussions and support.

References

[1] Michael Kass, Andrew Witkin & Demetri Terzopoulus. (1988) "Snakes: Active contour models," *International Journal of Computer Vision* pp. 321-331.

[2] Alexander Pentland & Kenji Mase. (1989) "Lip reading: Automatic visual recognition of spoken words," *Proc. Image Understanding and Machine Vision, Optical Socieity of America*, June 12-14.

[3] Eric D. Petajan, B. Bischoff & D. Bodoff. (1988) "An improved automatic lipreading system to enhance speech recognition," *ACM SIGCHI-88* pp. 19-25.

[4] David G. Stork, Greg Wolff & Earl Levine. (1992) "Neural network lipreading system for improved speech recognition," *Proc. IJCNN-92* Vo. II, pp. 154-159.

[5] Greg Wolff, K. Venkatesh Prasad, David G. Stork & Marcus Hennecke. (1994) "Lipreading by neural networks: Visual preprocessing, learning and sensory integration," *Proceedings of the Neural Information Processing Systems-6* J. D. Cowan, G. Tesauro and J. Alspector (eds.) Morgan Kaufmann, pp. 1027-1034.

[6] Benjamin P. Yuhas, M. H. Goldstein, Jr., Terrance J. Sejnowski & R. E. Jenkins. (1988) "Neural network models of sensory integration for improved vowel recognition," *Proc. IEEE* **78**(10), pp. 1658-1668.

[7] Alan L. Yuille, David S. Cohen & Peter W. Hallinan. (1989) "Feature extraction from faces using deformable templates," *CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Washington DC, IEEE Computer Society Press, pp. 104-109.