Lip tracking using pattern matching snakes

Mark Barnard , Eun-Jung Holden, Robyn Owens
Department of Computer Science & Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley, W.A. 6009, Australia.
E-mail:{markb, eunjung, robyn}@cs.uwa.edu.au

Abstract

Lip reading is an important tool used by hearing impaired people to increase their understanding of spoken language. Linguistics experiments show that the features of speech that are degraded by noise in audio systems are the features of speech that are most distinct visually. An automatic visual speech recognition system should therefore be a beneficial complement to an audio speech recognition system when the audio system is used in a noisy environment.

An automatic lipreading system requires a robust method to track lips from the image sequence, regardless of the variations of lip shapes, colours and movement amongst speakers. One of the most common lip tracking methods is the use of active contour models or snakes to detect outer lip contours. This paper proposes a robust and adaptable lip tracking method that uses a combination of snakes and a 2D template matching technique. The snake, an energy minimising spline, is driven by 2D template matching techniques to find the expected lip contour of a specific speaker. Our experiments show that the technique can track the unadorned lips in various colours and shapes of speakers, including the lips of a bearded speaker.

1. Introduction

Speech recognition is not purely auditory. When a listener can see the speaker, visual information is used in the speech recognition process. The contribution of this visual information to overall speech recognition was first reported by McGurck and MacDonald [7].

Noisy environments seriously degrade the performance of audio speech recognition systems. The nature of this degradation can be seen in figure 1 with data obtained by Kryter [6]. In this experiment the confusion between consonants was noted as the signal to noise ratio was lowered by 6 decibels.

Consonant cluster	Manner of articulation
t, k, p, f, th, s, sh	unvoiced
m, n	nasal
d, g, b, v, dh, z, zh	voiced

Figure 1. Effects of noise on audio phoneme recognition. Consonant clusters show those consonants that are confused with -6 dB noise level. At this level of noise only the manner of articulation can be used to categorise the consonants.

Consonant cluster	Place of articulation
th, dh	dental
f, v	labio-dental
sh, zh, z, s	alveolar/palatal
p, b, m	bilabial
w	labio/velar
t, d, n, k, g, j, r, l	not visible

Figure 2. Visual confusion between different consonants. Consonants confused by untrained subjects. Subjects placed phonemes correctly into these groups more than 70 percent of the time.

These results show that the most distinctive audio feature of phonemes in a high noise environment is the manner of articulation. The most distinctive feature visually is the place of articulation. Figure 2 shows six visually contrastive consonant groups [9].

The feature of phonemes that is most easily lost in noise is the place of articulation, and thus is the most distinctive feature visually. An automatic visual speech recognition system therefore provides a useful complement to an au-

dio speech recognition system in noisy or corsstalk environments.

Lip reading has been an important communication tool for the hearing impaired. Speech is composed of individual speech sounds, called phonemes, and when spoken, they show specific movements in the mouth shape including the lips, tongue and the appearance of teeth. However, there are 43 phonemes in the English language, while there exist only 28 different mouth shapes that separate them [2]. For example, 'd' and 't', or 'f' and 'v' produce the same mouth shape. Therefore, the art of lip reading for humans is context sensitive: it consists not only in visually recognising mouth shapes, but also mentally recognising key elements to predict the word, as well as further recognising key words to predict the sentence.

Automatic lip reading, moreover, is difficult for both the visual feature extraction and the speech recognition processes. Visual feature extraction requires a robust method of tracking the speaker's lips through a sequence of images and a representation of the inner mouth appearance. Lip tracking is not a trivial task because there is variability between people in skin colour, lip colour, lip width, and the amount of lip movement during speech, as well as variability in the environment such as lighting conditions. Any method used to track lips during speech must be adaptive to the movement of the lips from frame to frame, but also stable enough not to be affected by the appearance of the teeth and tongue.

A number of methods have been proposed for extracting lip contours from images. One of the most common methods of visual feature extraction is the active contour models or snakes which are parameterised energy minimising splines that converge to an object contour within an image. The snake technique was first introduced by Kass, Witkin & Terzoupoulos [5] and use an energy equation that maximised smoothness in the spline while finding the maximum gradient in the image contour. A serious problem with the snake is that some high level process, usually the user, must place the initial snake points close to the feature of interest. This problem arises because the snake will converge to the closest minimum in the image. In lip tracking, however, the appearance of the teeth and tongue generates a large intensity gradient and causes the snake to diverge from the outer lip contours. Thus researchers have used various modification of snakes [11][1], mainly by employing learned models of the lips to constrain the snake.

We have developed a lip tracking system that uses a combination of the snake [10] and a 2D template matching technique. In our system, a 1D deformable template (snake) is drawn onto the outer lip contour by using 2D template matching to find the expected lip contour within the search neighbourhood. The initial contour patterns are extracted from the manual selection of the mouth region from the first image, then the expected patterns are gradually updated throughout the sequence.

2. Methods of Visual Speech Feature Extraction

Yuille *et al*. [11] use shape templates with snakes in order to extract lip contours from an image. The lips are described by a parameterized shape template. This shape template models an object within the image. By adjusting the parameters the model can be made to deform to fit the object in the image. The shape of this template is based on prior knowledge of the shape of the lips.

Their shape templates consist of parabolas to describe the upper and lower lip shapes. These deformable shape templates then interact dynamically with the image through an energy function that draws the shape template onto salient features by altering the parameters of the parabolas. Preprocessing is done on the image to produce fields containing features of interest such as edges, valleys, or peaks. The shape template then interacts with these fields through the energy function. Shape templates have the advantage that the internal potentials in the parabolas will tend towards already known lip shapes as the energy equation is minimized. This minimizes the probability of the shape template settling on the wrong minimum, as is possible with snakes. A drawback of the shape template method is the limited flexibility of using predefined lip shapes.

The most successful lip reading system to date is developed by Bregler and Omohundro [1]. They also use learned lip shapes with the snake algorithm by using the technique of nonlinear manifolds. The configurations of the lips are represented as points in a feature space and the set of all possible lip configurations is a surface or manifold in this space. From training data a set of points in configuration space is produced and the dimension and structure of the manifold on which those points lie is induced. These training data are collected using manually controlled snakes. To track unknown lip shape, snakes are used that are controlled using the learned manifold. An initial crude estimate is back projected from the lip manifold to the image. There is then one iteration around the snake points and the grey level gradient is estimated. After each iteration the snake points are projected back to the lip manifold, so only legal lip shapes are considered throughout the iteration. The snake is constrained to only converge to these legal lip shapes. Legal lip shapes are determined from training data collected using manually controlled snakes.

3. Method

The systems described above use modified snakes that are controlled through the use of learned lip shapes. We use

a technique that adapts to different lips without any prior training. The hypothesis is that snakes can be controlled by using two dimensional pattern templates of the lip edge contour instead of the image gradient. The pattern templates should provide stability, while the snakes should provide the dynamics to move with the lips and adjust to new shapes.

3.1. Initialisation

The first image of the input sequence is assumed to be the image of a speaker in neutral mouth position, that is a closed mouth. Currently, initialisation involves the manual selection of a mouth region that is defined by the top, bottom, left and right corners of the mouth from the first image.

The initial pattern templates that are the surface image patches for each snake point are automatically extracted from the selected mouth region. These templates are then gradually updated throughout the sequence in order to determine the expected contour patches towards which the snake points move. The upper and lower lip widths of the speaker are also detected from the first image for the purpose of inner lip tracking. The inner lip contour edge varies a great deal during speech due to the appearance of the teeth and tongue. Thus we use lip widths that are extracted from the neutral (closed) mouth position to find the inner lip contour throughout the sequence.

3.2. Lip model

The mouth contour is modeled mathematically by using elliptical segments. The dynamics of lip movement and the biological lip shape of the upper and lower lips are different, thus the curvature and the expected movement range of the upper and lower lips are considered separately. This is achieved by modeling the outer lip contour as a combination of two semi-elliptical shapes, similar to the model used by Yuille et al. [11].

The initial inner snake points are also modeled as two semi-elliptical segments defined within a rectangular region where the corners of the inner and outer lips are coincident. Lip widths W is defined by the neutral mouth position, thus for n snake points, $W = (w_1, \ldots, w_n)$. The modeling process is shown in Figure 3.

3.3. Modification of the snake

2D pattern templates of the snake points are updated by using a weighted average of the initial pattern template and the template extracted from the previous image of the sequence.

The initial snake points for each image are determined by modeling the lip contour using two ellipses within the mouth dimensions detected in the previous frame.

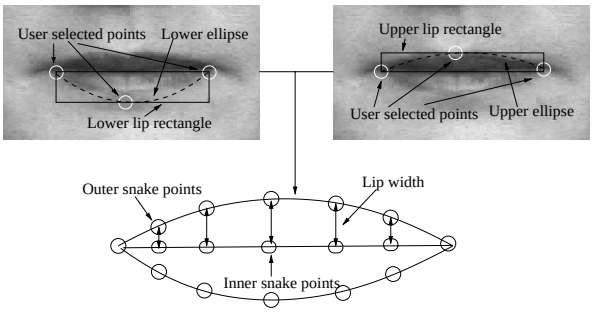

Figure 3. Modeling lip contours using two rectangles. The ellipses for the upper and lower outer lip edge are determined from these two rectangles

Our snake algorithm adapts the algorithm of Williams and Shah [10]. In this algorithm the overall energy term is reformulated as

$$E = \int (\alpha(s)E_{cont} + \beta(s)E_{curve} + \gamma(s)E_{image})ds, \quad (1)$$

where the parameters α, β and γ are used to control the relative importance of each term.

Given n snake points in a single frame, $s_1 \ldots s_n$ where $s_i = (x_i, y_i)$, the continuity term is defined as $E_{\text{cont}} = \overline{d} - d_i$, where $d_i = |s_i - s_{i-1}|$ and $\overline{d}$ is the average of d_i. This term ensures that the snake points will not be drawn together along the snake contour but will remain approximately equidistant.

The curvature term is defined as

$$E_{curv} = \left[\frac{\Delta x_i}{d_i} - \frac{\Delta x_{i+1}}{d_{i+1}}\right]^2 + \left[\frac{\Delta y_i}{d_i} - \frac{\Delta y_{i+1}}{d_{i+1}}\right]^2, \quad (2)$$

where Δx_i is $x_i - x_{i-1}$ and Δy_i is $y_i - y_{i-1}$. When this term becomes too large at a point, the curvature is very large so β for that point is set to zero allowing the snake to bend around a corner. The gradient threshold must be above a threshold to ensure that the snake is on, or very close to, the desired feature before a corner is developed.

For image energy, snakes normally use image edge strength to search for the object surface. This, however, is not suitable for lip tracking because the appearance of the teeth and tongue would distract the snake from the outer lip contour. Thus, we use 2D pattern matching to draw the snake onto the expected mouth surface, by defining the image energy as the two dimensional correlation between a two dimensional patch taken from the image and the expected template for the specified snake point.

3.4. Pattern matching

Lip shapes change dynamically through the image sequence. It is observed that the mouth corners require the templates to be updated while the other snake points along the lips can be tracked by using the initial templates. TWe use the following formula to update the mouth corner templates.

$$T_{exp\_temp} = 0.5 \times T_{cur\_temp} + 0.5 \times T_{init\_temp}.$$

The pattern matching snake is illustrated in Figure 4.

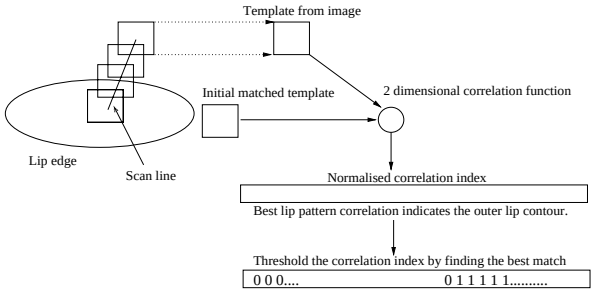

Figure 4. The 2D template matching snake algorithm. This is designed to smooth out noise in the image by using a two dimensional template with correlation matching.

For each snake point a scan line is generated along the normal of the lip edge, see figure 5.

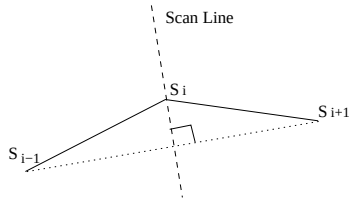

Figure 5. Each scan line is normal to a chord running between the previous snake point and the next snake point.

Along each scan line sampled at m discrete points, 2D templates centred on these sampled points are used to move the snake point. A two dimensional correlation is computed between these templates and the expected template using the following correlation function, where A and B are two dimensional matrices and C is a correlation index between 0 and 1:

$$C = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}}. \quad (3)$$

The correlation indices form a 1D array with the size of the sample points, m. This index array is normalised with the value of the maximum correlation index, and thresholded. The array then has a step edge in the position of the lip contour. We convolved this array with a first derivative Gaussian mask to smooth the step edge and used this as the image energy.

3.5. Inner Lip Tracking

Currently, inner lip tracking is done by using the initial measurement of lip width. The inner lip point is found by taking a point that is the distance of the initial lip width along the scan line of the corresponding outer lip point. However, accurate inner lip tracking would require the snake to move by using a combination of lip width and edge strength.

3.6. Head Tilting

The problem of subjects tilting their heads from side to side during speech is also taken into account. The degree of tilt from the horizontal axis of the mouth is determined by the mouth corner snake points found, and thus the ellipses that form the snakes are rotated accordingly for subsequent tracking.

4. Experiments

We have implemented a prototype of the pattern matching snakes using MATLAB 5.2 image processing toolkit.

The parameters used for the inner and outer snakes in our implementation are shown in figure 6

The biological features of the mouth are considered. This has been done in the selection of the snake parameters. At the mouth corners the values of α and β are reduced to 0.2 in both snakes to allow sharp corners to form. The outer snake image parameter γ was set to 2.5 and the lip width parameter set to 0.

In tracking the lips it is essential to accurately track the mouth corners. The position of the mouth corners determines the angle that the mouth is tilting. This then determines how the upper and lower rectangles are constructed, thus determining the starting point for the next snake iteration. If the position of the mouth corners is incorrect this will affect the position of all following snake points. In order to track accurately the mouth corners the template size

Snake parameter	Value
α	0.7
β	1
γ	2.5
ϵ	0
Number of points	20
Pattern template size	20×20
Upper lip scan line length	21
Lower lip scan line length	31

Figure 6. Snake parameters used in experiments with pattern matching snakes.

at the corners was increased to 40×40. Other snake points use the template size of 20×20.

The motion of the lips in general is in the direction of the normal to the lip edge. The scan lines for the outer and inner lip snakes are normal to the previous snake contour. During speech the lower lip has a greater range of movement and in order to account for this the scan line length of the lower lip points was 21 pixels while the length for the upper lip points was 11 pixels. The mouth corners have a much smaller range of movement, however this movement can be either vertical or horizontal. In order to deal with this, the points at the mouth corners scan in a 7×7 square.

5. Results

The subjects for these experiments were all native speakers of Australian English. The tracking system was tested on 6 males and females. One male speaker has darker skin colour and another male has some beard growth. All speakers had unadorned lips, thus there was a range of different lip shapes and colours amongst the speakers. They were asked to stay stationary during speech. The speakers articulated the sequence of numerals from "one' to "ten" without any pause between the numerals. The sequences were captured through a video camera (Sony Hi8) from a single viewing point. These sequences were then stored as a series of 24-bit colour bit map images with resolution 384×284. However, we converted them into greyscale, and all processing is performed on the greyscale images. The pattern matching snake successfully tracked all of the sequences of the 6 speakers. Each sequence consists of over 200 images.

Figure 7 shows the lip tracking result of two subjects, one with dark complexion and the other with a beard. These experiments represent probably the most troublesome cases for other lip trackers. For the dark complexioned speaker, the contrast between skin and lip intensities is not obvious. For the bearded speaker, there was a false contour close to the lips, namely the contour between skin and beard. The

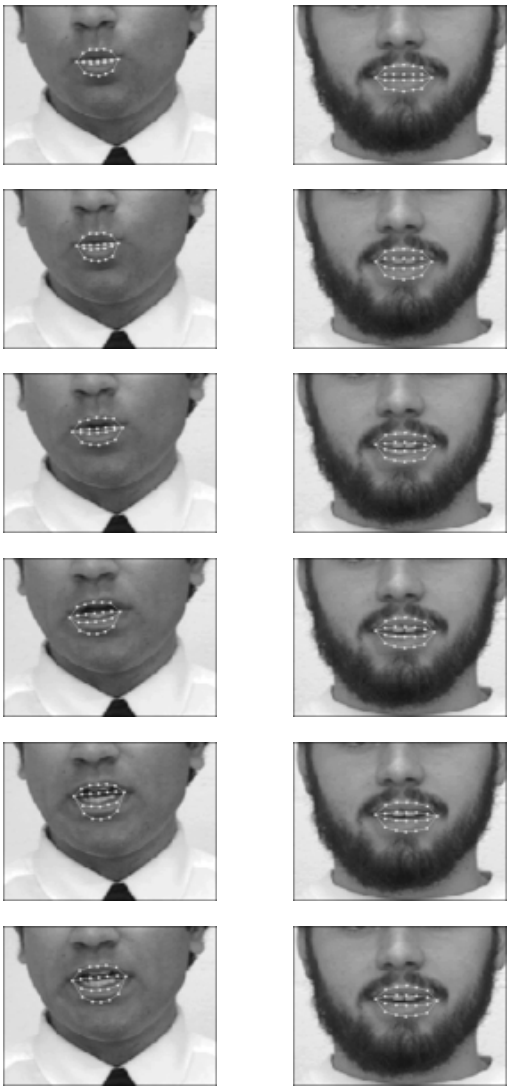

Figure 7. Sequence of six image showing lip tracking while the mouth is moving.

pattern matching snake, however, is not affected by these constraints because it searches for the expected contour patterns of the specific speaker. The figure also shows that the speaker's head moved slightly even though they were asked to stay stationary during speech. The system successfully tracked the tilted mouth.

Figure 8 shows three different speakers used in our experiment. The complexions, lip colour and shapes of the speakers were very different, and also the teeth and tongue were visible during speech. The system however was not affected by these, successfully tracking all of the sequences during experiment. The pattern matching snake is also not

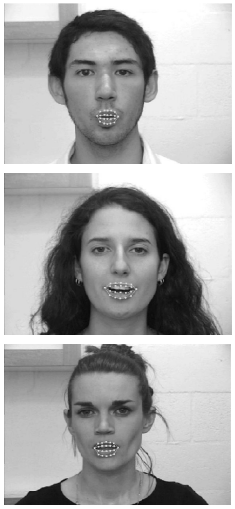

Figure 8. Single frame taken from sequences of lip tracking on a range of different subjects.

affected by noise because of the smoothing effect of the two dimensional templates used in the image energy calculations.

6. On-going Development

Tracking the mouth is a difficult task since the mouth shape changes rapidly during speech. Our lip tracking technique uses a combination of the active contour models and a 2D template matching technique where the energy minimising spline is driven onto the expected lip contour by a 2D pattern matching technique. Our experiments show that this technique is reliable and adaptive to speakers with various mouth shapes and colours.

For practical use of the template matching snake, we need to automatically collect the initial templates. We plan to adapt the technique of Okata *et al*. [8], which uses the Gabor wavelet response to detect the facial features for their face recognition system.

A lip reading system not only requires the mouth shape dimensions that are extracted from the lip tracker but also requires a representation of the appearance of teeth and tongue during speech. For this problem we have developed a technique that combines Cepstral analysis and higher order local auto-correlation feature extraction [4].

The lip tracker assumes the speaker faces the camera directly, allowing for a slight rotation of the face within the image plane. The speakers, however, naturally follow conversation cues and move the head in 3D. 3D movements such as turning and nodding of the head affect the lip tracker and cause it to extract incorrect mouth dimensions. To deal with this problem, we have developed a 3D head tracker, which extracts the 6 degrees of freedom of the head orientation from a single view image sequence, and corrects the incorrect mouth dimensions [3].

References

[1] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 494–499, Los Alamitos, CA, June 1995. IEEE Computer Society press.

[2] A. Greenwald. *Lipreading made easy*. Alexander Graham Bell Association for the Deaf, 1984.

[3] E. J. Holden, G. Loy, and R. Owens. Accommodating for 3d head movement in visual lipreading. In *Proceedings of IASTED International conference on Signal and Image Processing*, pages 166–171, 2000.

[4] E. J. Holden and R. Owens. Visual speech recognition using cepstral images. In *Proceedings of IASTED International conference on Signal and Image Processing*, pages 331–336, 2000.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Proc. of IEEE First International Conference on Computer Vision*, pages 259–269, 1987.

[6] K. D. Kryter. *The effects of noise on man*, chapter 2, pages 48–52. Acedemic Press, 1970.

[7] D. McGurck and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264, December 1976.

[8] J. Okada, K.and Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. Malsburg. The bochum/usc face recognition system and how it fared in the feret phase iii test. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman Souwe, and T. Huang, editors, *Face Recognition: From Theory to Applications*, pages 186–205. Springer-Verlag, 1998.

[9] B. K. Walden, R. A. Prosek, M. A. A, C. K. Scherr, and C. J. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20:130–145, 1977.

[10] D. J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding*, 55(1):14–26, 1991.

[11] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.