Robust Lip Tracking using Active Shape Models and Gradient Vector Flow

Simon Lucey, Sridha Sridharan and Vinod Chandran
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
e-mail: s.lucey@qut.edu.au, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

Abstract

Lip movements and configurations provide useful information which can be utilized to improve automatic speech and speaker recognition. However the use of this visual information requires accurate lip tracking algorithms. A new technique is outlined that is able to estimate the outer lip contour under noisy conditions accurately. It is based on the combination of gradient vector flow (GVF) force fields and active shape models (ASM). Results are presented for speech reading using a speaker dependent database. The effectiveness of this algorithm on noisy edge maps generated from chromatic segmentation is also demonstrated.

Keywords: gradient vector flow, lip tracking, speech reading, active shape models

1 Introduction

An audio-visual speech processing (AVSP) algorithm using lip movements must be able to accurately track the labial contour despite variations owing to lighting, camera noise and the speaker's appearance. The importance of the labial contour for AVSP applications has been well documented [1, 7, 8]. Techniques useful for lip tracking include active contours [3, 4, 5] and deformable templates [12]. Deformable templates are difficult to use for representing fine details in the labial contour. Active contour models (ACM) have been used extensively for lip tracking as they provide finer detail about the outer labial contour.

In this paper, we address the problem of extracting the labial contour from a pre-processed 'edge-map'/potential image (i.e. a binary image describing the outline of an object) with the edge map \( f(x,y) \) being described as,

\[
f(x,y) = \begin{cases} 1 & T(I(x,y)) \geq Th \\ 0 & \text{else} \end{cases}
\]

where \( T \) is some edge enhancement operator, \( Th \) is a threshold and \( I(x,y) \) is the original intensity image. The extraction of the labial contour from the potential image \( f(x,y) \) has to be performed in a manner that has low sensitivity to the initialisation of the estimated contour and any noisy artifacts present in the potential image. When an image taken from the mouth region of interest (ROI) is pre-processed to gain a potential image, the contour around the labial outline can contain unwanted visual artifacts from noise and/or contain broken lines. Using an edge detector alone, however good, will not separate the outer labial contour from other structures in the image. More prior knowledge on the allowable shapes of lips needs to be brought to bear on the problem. Previously, active contour models (ACM) or 'snakes' have been used to provide syntactic restrictions in lip shape with good results [1]. However, active contour models have some problems associated with them when being used to track lips from a potential image. Firstly, the syntactic restrictions they provide for shape deformation are quite general so that for noisy potential images the resultant fitted contour may itself be noisy. Secondly, the potential force fields derived from the potential image, which tell the contour in which direction to move, have problems associated with the initialisation of the estimated contour when noise (i.e. unwanted visual artifacts or broken lines) is present.

We present a new class of potential forces, based on gradient vector flow (GVF) fields, that can evade some of the problems caused by noisy potential images. Gradient vector flow (GVF) fields [9] are both insensitive to initialization and have an ability to move into concave boundary regions. These fields are used in conjunction with active shape models (ASM) which provide a way to vary a contour based on pre-trained syntactic information about allowable labial contour deformations. Using both GVF fields and ASM a technique has been developed that can reliably converge to the correct lip contour outline whilst maintaining a valid lip shape under adverse conditions.
2 Active shape models and potential images.

Active shape models (ASM) have proved to be very good at providing a model for the deformation of lip contours and in turn provide an accurate way of tracking a speaker's lips [7]. ASMs have the advantage of providing a priori knowledge about typical deformation of lips from a training set of labeled lips.

The lip contour \( x \) can be described by \( n \) points

\[
x = (x_0, y_0, x_1, y_1, \ldots, x_{n-1}, y_{n-1})
\]

This contour can be approximated using principle component analysis (PCA) [7] by,

\[
x \approx \bar{x} + Pb
\]

where \( \bar{x} \) is the mean of the training feature vectors, \( P \) the matrix of the first few column eigenvectors of the covariance matrix which correspond to the largest eigenvalues and \( b \) a vector containing the weights for each eigenvector. The vector \( b \) can be used as a compact and decorrelated approximate representation of the original contour vector \( x \) in which the main modes of variation have been preserved.

Active shape models when used for lip tracking have been predominantly implemented within a multidimensional energy minimisation framework [7] that actually differs from the original approach given by Cootes, Hill, Taylor and Haslam [2]. Cootes suggested that potential images could be used to calculate suggested movement for each model point such as those used with active contour models [4]. This potential image describes how likely each point in the image is to be the model point and is usually described via an edge map which is created using a pre-processing step described in Section 3.

Adjustments to the position of each point can then be derived from the potential force field generated from the potential image. Convergence can be achieved in fewer iterations with the model constrained to vary within valid lip shapes as dictated by the ASM.

3 Extraction of an potential image from the mouth ROI.

The mouth ROI image that is used for visual feature extraction can be a grayscale or colour image. Grayscale images suffer from lighting problems and it has been reported that the variation of the gray levels around lips is small [1]. Colour images of the mouth ROI have greater class distinction between human lips and the predominant skin background allowing for far more accurate modeling of the labial contour when compared to their grayscale counterparts. However, using colour as a feature has several problems. Firstly, the colour representation of a person obtained by a camera is influenced by ambient light and background. Secondly, different cameras produce significantly different colour values, even for the same person under the same lighting conditions [10]. These variabilities make the generation of edge-maps through chromatic features problematic, with the resultant binary images often containing noisy artifacts.

In this paper we have chosen to use a technique of thresholding using the ratio of red to green pixel values in a RGB image to produce a binary image of candidate lip pixels as described in [1, 6, 8]. A binary image is created by finding a threshold in \( \frac{r}{g} \) space, as described by Chiu [1], that best segments the much redder lip pixels from the skin pixels as is demonstrated in Figure 1 (b). After some morphological operations to get rid of spurious pixels a edge map \( f(x, y) \) is created by applying a standard edge kernel to the binary image as demonstrated in Figure 1 (c).

![Figure 1: Demonstration of how potential image is created.](image)

4 Gradient vector flow.

The generation of a suitable potential force field \( v(x, y) \) from a potential image \( f(x, y) \) can be error-prone. First, the initial contour must, in general, be close to the boundary or else it is likely to converge to a wrong result. Second, most potential field forces have problems progressing into boundary concavities which can sometimes restrict a contour from being fitted accurately to a potential image.

Recently, a new class of potential forces has been proposed that overcomes these problems. These fields, called gradient vector flow (GVF) fields, are dense vector fields derived from images by minimizing a certain energy functional in variational framework [9]. When used with active contour models they have been shown to be insensitive to initialization and have an ability to move into boundary concavities. The theory behind GVF can be found in [9] which gives a thorough overview of the process. A gradient vector flow field can be defined as \( v(x, y) = (u(x, y), v(x, y)) \) that minimises the energy functional

\[
E = \iint \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + \nabla f^2 \|\nabla f\|^2 \, dx \, dy
\]

(4)
Where $\mu$ is a regularization parameter governing the tradeoff between the first term and the second term in the integrand. This produces the desired effect of keeping $v$ nearly equal to the gradient of the edge map when it is large, but forcing the field to be slowly-varying in homogenous regions as can be demonstrated in Figure 2. The solution to Equation 4 is found via a numerical implementation which is outlined in [9].

![Lip potential image](image)

![GVF force field](image)

![Normalised GVF force field](image)

Figure 2: Process of calculating normalised GVF force field.

To make the process of fitting a contour via the potential force field $v$ as linear as possible it is convenient to normalise the magnitude of the fields such that the force field contains only directional information as demonstrated in Figure 2(c). This normalisation process was undertaken for all our tests.

5 Calculating movement for each model point.

The method we used for calculating the adjustments to shape parameters based on the GVF and ASM are very similar to the technique used by Cootes in [2]. Given an initial estimate of the positions of a set of model points which we are attempting to fit to an lip image, and the GVF potential force field $v(x)$ which points to the proposed outer labial contour, we need to estimate a set of adjustments which will move each point toward a better position while maintaining a valid lip shape. These adjustments can be calculated for each model point which can be denoted as

$$dx = (dx_0, dy_0, dx_1, dy_1, ..., dx_{n-1}, dy_{n-1})$$  \hspace{1cm} (5)

Where $n$ denotes the number of points representing the contour. Before deforming the ASM itself we have to first find the approximate centre of the lips $(x_c, y_c)$,

$$x_c = (x_c, y_c)$$  \hspace{1cm} (6)

The need for calculating $x_c$ is due to the ASM of the lips being trained in such a way that the centre of the labial contour is at the origin. This was done to ensure the ASM modeled only the allowable lip shape variation and not translational variation. As an initial estimate of the lip shape the mean lip shape of the lips being trained in such a way that the centre of the labial contour is at the origin. This was done to ensure the lips being trained in such a way that the centre of the labial contour is at the origin.

The approach for finding the approximate centre of the labial contour is as follows:-

1. Calculate adjustment vector $dx$ from the GVF force field $v(x + x_c)$;
2. Calculate centre adjustment vector $dx_c = (dx_c, dy_c)$ where $dx_c = \frac{1}{8} \sum_{i=0}^{n-1} dx_i$ and $dy_c = \frac{1}{8} \sum_{i=0}^{n-1} dy_i$;
3. Update $x_c$ by new centre adjustment vector such that $x_c^{(t+1)} = x_c^{(t)} + sdx_c$ where $s$ is step size;
4. Repeat steps 1-3 for $n$ iterations;

For our experiments we chose a step size $s$ of one pixel and performed the above steps for twenty iterations as this assured convergence. Once the initial shape estimate is positioned correctly we can then make adjustments to each model point within a ASM framework so as to give an optimal fit to the potential image.

We aim to adjust the shape parameters so as to move the points from their current locations $x$ and $y$ as close to $x + dx$ as can be arranged whilst still satisfying the shape constraints of the ASM. In [2] it was shown that the optimal way to calculate adjustments to the shape parameters $db$ of an ASM described by the weights $b$ in a least squares sense is

$$db = P^T dx$$  \hspace{1cm} (7)

Using the result obtained in Equation 7 we can deform the ASM contour via the following steps

1. Calculate adjustment vector $dx$ from the GVF force field $v(x + x_c)$;
2. Calculate $db$ as per Equation 7;
3. Update $b^{(t+1)} = b^{(t)} + sdb$
4. Get new estimate of $x \approx x + Pb$;
5. Repeat steps 1-4 for $n$ iterations;

Again for our experiments we chose a step size $s$ of one pixel and performed the above for forty iterations as this assured convergence.

5.1 Qualitative performance of algorithm.

The fusion of ASM based contour deformation and GVF gives excellent tracking performance in a number of trying conditions. This is particularly true when the labial contour outline in the potential image is obstructed by noise.

Typically there are two types of noise present in segmented mouth ROI images that cannot be treated effectively through conventional means:-

1. Binary image of lips with unwanted artifacts attached to the binary lip cluster as seen in Figure 3(a);
2. Binary image of lips with missing lip pixels in the binary lip cluster as seen in Figure 4(a);

Using GVF force field in conjunction with an ASM of the lips, fits a contour to the model that gives an excellent estimation of the labial contour circumventing the noisy artifact present in the potential images shown in Figures 3(a) and 4(a) the final results of which can be seen in Figures 3(d) and 4(d).

![Figure 3: Demonstration of robust contour fitting on a potential image with unwanted lip pixel artifacts.](image)

![Figure 4: Demonstration of contour fitting on potential image with missing lip pixels.](image)

6 Lip tracking results.

Comparison of different lip tracking algorithms is difficult, one method of judging the quality of the extraction is to evaluate the effectiveness of the contours when used in a specific application. Luettin [7] has recently proposed that an accurate measure of the quality of visual features is indicative of how well it performs in a given lip reading application (ie. the recognition of visemes or words). Through automated lip reading we are able to quantitatively gauge the effectiveness of this lip tracking algorithm.

We tested this algorithm on our own speaker dependent database to evaluate its effectiveness for a speaker dependent word recognition application. We chose to test our algorithm initially within a speaker dependent framework so as to gauge the effectiveness of the GVF with a speaker dependent ASM of the lips due to its natural insensitivity to visual noise. From our own experiments we found speaker dependent ASMs are more robust in noisy circumstances, because they have fewer modes of variance in their shape, in contrast to speaker independent ASMs which have many more modes of variance making them further prone to the effects of visual noise. In this paper it was not our purpose to quantitatively compare our algorithm to other lip tracking algorithms. We thought it more instructive to get a qualitative measure of how effective the algorithm is within a speaker dependent word recognition application. To this end we have recorded our own database ensuring we have large amounts of reasonable resolution speaker dependent visual data captured under typical lighting conditions for the purposes of word recognition.

We chose ten words for the recognition task where the data was collected from a single speaker using SGI 02 workstation and Panasonic VSK0537 digital camera. The recordings had the following characteristics:

- ten words from ‘one’ to ‘ten’;
- each word has 16 examples;
- video captured at 25 fps;
- captured at standard 720x576 PAL resolution;
- video recorded under normal lighting and audio levels;

Using HTK ver 2.2 [11] the continuous video sequences were automatically transcribed into their respective digits using a HMM recognition technique on the synchronous audio with all silence segments being removed. The audio was not used for the tests conducted subsequently for word recognition from lip features. Each video sequence then had the eyes and mouth manually located with the distance between the eyes being used as a measurement of scale. The mouth region of interest (ROI) for each speaker was extracted with the eye distance being used to normalise for scale thus giving a 140x120 intensity image of the mouth an example of which can be found in Figure 1 (a).

Due to the small size of the training/testing set recognition tests were performed using the ‘leave-one-out’ method ie: fifteen utterances were used for training and one for testing for each individual digit. The whole procedure was repeated sixteen times. The actual word recognition system was made very simple so as to gauge
the effectiveness of the lip features and not the recognition system itself. A left to right mixture five state HMM was used for the word models. The resultant contour features from the lip tracking algorithm in this paper are of too high a dimensionality \( d \) (ie. for testing \( d = 100 \)) to be used in a HMM recogniser due to singularities. Therefore, the weights \( b \) from the ASM, as described in Equation 3, were used as visual features due to their low dimensionality (ie. for testing \( d = 7 \)). For further performance improvement the recogniser was fed a vector of static ASM weights concatenated with their respective delta weights calculated from adjacent frames. The recognition results can be seen in Table I.

<table>
<thead>
<tr>
<th>ONE</th>
<th>TWO</th>
<th>THREE</th>
<th>FOUR</th>
<th>FIVE</th>
<th>SIX</th>
<th>SEVEN</th>
<th>EIGHT</th>
<th>NINE</th>
<th>TEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

Table I: Confusion matrix of word recognition results with overall recognition accuracy of 95 %.

7 Conclusions

A technique for tracking lips using GVF force fields in conjunction with ASM an trained on typical outer labial contours, has been presented. Experiments were conducted on our database so as to gauge the effectiveness of the tracking system under trying conditions. Results have shown the tracking system to be quite effective with an overall recognition accuracy of 95 %. The system also demonstrated an insensitivity to the visual noise qualitatively described in Section 5.1, with the ability of making an estimate of the labial contour. The confusion matrix shown in Table I demonstrates the effectiveness of our technique, within a speech reading context, as the HMM based automated speech reading engine was able to distinguish individual digits in all but a few instances.

The algorithm in its current form has been only tested for speaker dependent situations. Some of our future work shall concentrate on formulating a speaker independent algorithm for use in wide spread AVSP applications. Our system acts as a good post processing tool for extracting a labial contour from a potential image. The syntactic information used within our technique along with its robust nature can make up for deficiencies currently existing in chromatic based lip tracking techniques.

The technique still has some problems associated with it in the creation of the GVF force fields. Some of our future work shall concentrate on expediting the formulation of the GVF force fields so as to allow this system to be used in a real time system within AVSP applications.

References