Fifth International Symposium on Signal Processing and its Applications,
ISSPA '99, Brisbane, Australia, 22-25 August, 1999
Organised by the Signal Processing Research Centre, QUT, Brisbane, Australia

CHROMATIC LIP TRACKING USING A CONNECTIVITY BASED FUZZY THRESHOLDING TECHNIQUE

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
s.lucey@qut.edu.au, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

ABSTRACT

This paper describes a new robust method for extracting visual labial information based on chromatic information. Using chromatic temporal information for a speaker and a special fuzzy based thresholding algorithm a stable technique of extracting a speaker's lips is proposed that overcomes many of the problems associated with previous adaptive threshold techniques. Experiments performed thus far show encouraging results while being computationally inexpensive thus lending the entire process well to a robust real time application.

1. INTRODUCTION

Effective robust lip tracking techniques suitable for real-time applications have so far eluded the visual processing community. Many attempts have been made to try and create an effective system for the tracking of a speaker's lip movements. Some of the latest techniques incorporate chromatic information taken from a speaker's lips [5, 1, 3]. These techniques have been effective when being used for lip tracking, but require time consuming iterative techniques to optimise the lip contour estimates. Continuing the work initiated in [5] a syntactic approach to lip tracking has is proposed based on the chromatic information taken from a speaker using the ratio of red to green intensities $\frac{R}{G}$. This technique [5] used a simple adaptive thresholding technique to separate the lips from its surroundings. The Bayesian approach used in [6] involving the EM algorithm has problems in finding accurate thresholds and is quite computationally expensive.

The problem of selecting an appropriate threshold to separate the foreground and background of an image is an age old problem in image analysis. In this paper we propose a new technique based on the existing thresholding framework to refine and improve the lip tracking procedure and overcome some of the deficiencies associated with previous algorithms. Using a threshold selection algorithm based on a fuzzy measure it has been possible to find an appropriate threshold in a fast and reliable manner. The incorporation of the temporal information from progressive image frames is used in order to calculate a stable threshold for separating lip pixels from the predominantly skin based background. This technique is used in conjunction with a threshold relaxation method of identifying lip pixels based on connectivity. The combination of these techniques enable the accurate capture of a speaker's lips while still being suitable for use in a real time framework.

2. LIP TRACKING

2.1. Chromatic Characteristics of Lip ROI

Using traditional thresholding techniques as in [5, 1] a fixed threshold is usually found to differentiate between the different pixel classes. A major problem with most thresholding techniques is that they make the assumption that lip pixels have a fixed statistical distribution and can always be separated from the skin background. This assumption is valid for identifying some lip pixels but it is often the case that lip shape is lost if the threshold is too high or there may be a lack of lip pixels if the threshold is too low. This is due to the skin and lip pixel distributions being overlapped under normal lighting conditions. The task of the lip tracking algorithm is then to find the threshold which minimizes the probability of a pixel being labeled in the wrong *skin/lip* class.

Wark [6] recently attacked this problem using a Bayesian approach to find the threshold at which the probability of incorrectly classifying a pixel is minimised. The expectation maximisation algorithm (EM) was used to get the distributions at priori. This technique proved adequate in extract-

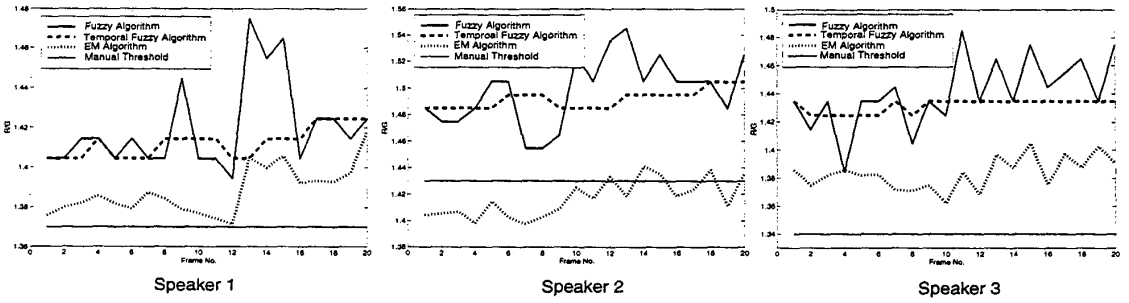

Figure 1: Thresholding results over 20 frames and 3 speakers
(The temporal fuzzy algorithm exhibits the least fluctuation from frame to frame)

ing lip contours in a lot of cases. However this thresholding technique tended to fluctuate considerably from frame to frame in the same image sequence with the estimated threshold usually a fair distance from it's ideal value. This caused problems in the lip tracking process with an image sequence's lip tracking ability being highly dependent on the accuracy of the threshold found.

To combat these problems we have developed a more elegant solution that doesn't require strict thresholds. Using a technique of fuzzy thresholding we established a technique of obtaining a threshold that identified definite lip pixels in a stable fashion. Then using syntactic information of lip shape the threshold is relaxed so as to extract the full lip contour.

2.2. Fuzzy Thresholding Algorithm

Many image processing problems that use a threshold to binarize the image, face problems due to the overlapping of the two class probability densities (ie. fuzziness). This type of problem lends itself quite well to fuzzy logic. Several fuzzy model based methods have been developed in the past to overcome these difficulties due to the fuzziness of the image data [7]. Most of these techniques have required for each possible threshold value in the range $[0, L-1]$, the membership value of every pixel as to compute the compactness or index of area coverage measure. For a real time application these methods can require long computing time to compute a optimal threshold value. Huang and Wang [2] have recently developed a fuzzy thresholding algorithm which makes use of the image pixel value histogram but does not need deal with each individual pixel.

We adapt the method proposed in [2]. Firstly, before implementing the lip tracking procedure the histogram for the $\frac{R}{G}$ intensities in the lip ROI has to be found and placed in a acceptable framework to be used with Huang and Wang's method. These distributions are vital to the entire lip tracking algorithm as they provide a way of adaptively selecting lip $\frac{R}{G}$ intensity thresholds based on some basic assumptions about the distribution. Using this distribution as a basis for classifying pixels by itself, without any preprocessing is fraught with problems. This is due to the camera being used to grab the digital images containing noise along with noisy lighting conditions. These problems manifest themselves in the distribution of $\frac{R}{G}$ values by giving the distribution a very jagged and discontinuous appearance. This in turn does not lend the distribution very well to be used as a multimodal statistical classifier. To alleviate the poor distributions received from such images a number of steps can be taken. Firstly we pass a two-dimensional median filter over the lip ROI to reduce camera and lighting noise. Then we pass median filter over the derived $\frac{R}{G}$ intensity image of lip ROI so as to reduce numbers of spurious intensity pixels. Finally we set all intensity pixels below 0.5 to 0.5 as they are too green. As well as all intensity pixels above 2.5 to 2.5 as they are too red.

Using a fuzzy approach the resultant $\frac{R}{G}$ intensity image can be considered as an array of fuzzy singletons corresponding to image pixels, each having a membership value associated with a certain property of the pixel. The $\frac{R}{G}$ intensity image can be set as an discrete integer intensity image f with range $[0, L-1]$.

$$f(x,y) = \tfrac{1}{2}(L-1)\tfrac{R}{G}(x,y) - 0.5 \qquad (1)$$

where the actual image can be represented as

$$I = f(x,y), \mu_I(f(x,y)) \qquad (2)$$

In Huang and Wang's method, the normalised pixel value histogram $h(z)$ which represents the percentage of pixels having intensity values z over the total number of pixels has it's membership function dened as

$$\mu_I(z) \begin{cases} \frac{1}{1+|z-m_1(T)|/D} & \text{if } z \le T \\ \frac{1}{1+|z-m_2(T)|/D} & \text{if } z > T \end{cases} \quad (0 \le z \le L-1)$$

where m_1 and m_2 are the averages of the intensity levels of the pixels in two classes and are given by

$$m_1(T) = \sum_{z=0}^{T} zh(z) / \sum_{z=0}^{T} h(z) \qquad (3)$$

Figure 2: Three tested speakers.

and

$$m_2(T) = \sum_{z=T+1}^{L-1} zh(z) / \sum_{z=T+1}^{L-1} h(z) \quad (4)$$

with $D = (L-1)$ being chosen in this application so as to ensure the membership values stay in a range suitable and stable for use with Huang and Wang's entropy equation as described in [2].

Huang and Wang made use of a entropy measure to quantitatively measure the uncertainty between the two classes as

$$J_{HW}(T) = \sum_{z=0}^{L-1} h(z) S_e(\mu_I(z)) \quad (5)$$

where Shannon's entropy equation is defined as

$$S_e(\mu) = -\mu \ln \mu - (1-\mu)\ln(1-\mu)(0 \leq \mu \leq 1) \quad (6)$$

The optimal threshold is chosen to minimize $J_{HW}(T)$, such that

$$T^*_{HW} = \arg_{0 \leq T \leq L-1} \min J_{HW}(T) \quad (7)$$

Using these equations it is possible to solve the classification problem using a fast computation algorithm as found in [7].

2.3. Comparison of results

The thresholding techniques described in previous sections were tested over three speakers, as seen in Figure 2, taken from the M2VTS database [4]. To begin with, the ideal threshold values in $\frac{R}{G}$ space were manually selected for each speaker. The image sequences were tested using the traditional Bayesian EM algorithm and Huang and Wang's fuzzy algorithm. A comparison between the EM algorithm and the fuzzy approach showed that the overall mean threshold for the fuzzy algorithm across twenty frames is considerably higher than the EM algorithm and manual thresholds for all three speakers. This can be attributed to the entropy based cost function of the fuzzy algorithm in Equation 5 that separates pixels into two classes whose means are as well separated as possible with the two class variances as small as possible. This encourages the second higher lip pixel class to be as well separated as possible. Unfortunately from testing it was found that neither algorithm was stable enough to be used accurately as a thresholding algorithm over all frames.

2.4. Temporal approach

As seen in the results in Figure 1 it is very difficult for an adaptive thresholding algorithm to successfully find an ideal threshold from a single frame taken from a continuous image sequence. The fluctuations seen in Figure 1 are at odds with our assumption that for a given image sequence, the threshold separating lips from the face background is relatively constant. These fluctuations in threshold values can be attributed to a number of unwanted attributes in the captured images such as camera and lighting shadows. However, the main problem is the size of the sample set being used to calculate the threshold. Due to the small set of data in the lip ROI (typically 3000 pixels) one image frame cannot give enough class distinction. To alleviate this problem a temporal profile of the $\frac{R}{G}$ histogram was used. This type of approach is ideal for the previously mentioned fuzzy threshold algorithm where the histogram $h(z)$ can be updated every frame with the new $\frac{R}{G}$ intensities to adaptively improve the accuracy of the *skin/lips* decision threshold. For our scheme we chose a training stage of five frames, as it would only take a negligible amount of training time (ie. at 25 frames/sec requires $\approx$ 0.2 seconds) and provided quite a stable and accurate threshold when compared to the manual thresholds. Due to the linear discriminant nature of the algorithm the threshold found was assured of being higher than the ideal threshold as seen in Figure 1. This lent the algorithm well to a threshold relaxation scheme.

2.5. Connectivity based thresholding

A problem that can be seen in all these results is that the adaptive threshold found is higher than the manually found ideal threshold. This can be attributed to a number of factors especially the size of the mouth with respect to the lip ROI. If the size of the mouth varies from frame to frame then the number of lip pixels changes thus changing the distinction between the *skin/lips*. There is no clear cut way to combat this problem as the algorithm has no way of knowing how many lip pixels are set a priori. An elegant solution is to introduce a relaxing procedure into the thresholding procedure based on connectivity. Pixels above the predened threshold can be labeled as definite pixels in region R_1. Pixels lying between the adaptive threshold T and a heuristic point between T and m_1, *(in this application* $m_1 + (T - m_1)/2$ was chosen) labeled as lying in region R_2.

Using this framework for the identification of lip pixels, a simple scheme based on connectivity was derived to get a true representation of lip shape. Working on the assumption that all pixels identified in region R_1 are lip pixels or some other non-skin pixel, a simple binary image can be created an example of which can be seen in Figure 3(a). This binary

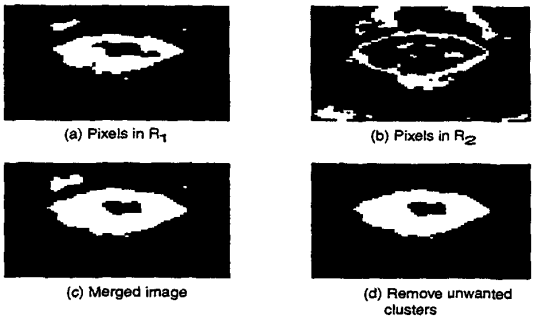

Figure 3: Binary image of lips using connectivity based thresholding.

image depending on the threshold used does not resemble the shape and form desired in a binary image of the lips as it is excluding many of the lip pixels existing in the image. By obtaining a second binary image of all pixels lying in region R_2 it is possible to make a much cleaner and accurate binary representation of a speaker's lips by merging the two binary images.

The merging process is based on using primary and secondary binary images. By using the set pixels in the primary image R_1 as an indication of where the lips are, a new image can be created by setting the pixels in the primary image on that are set in the secondary image R_2 and have a 8 neighbor connectivity to a set pixel in the primary image. This process can be repeated for n iterations but testing has shown $n = 2$ to suffice. The result of this technique can be seen in Figure 3(c). Finally the image can be cleaned up using morphological processing and connectivity based clustering. Removing all clusters except the largest results in the image seen in Figure 3(d).

3. LIP SHAPE PARAMETERS

Once a good estimate of the lip region had been formed, we wished to extract the outer labial information. This was accomplished by fitting separate semi-ellipses to the upper and lower lips through a simple procedure of finding the corners and height of the lips. This results in the lip contours as seen in Figure 4. Fitting semi-ellipses to the outer lips was chosen as it is computationally inexpensive to fit and forces the extracted shape to be lip like in appearance. A simple second order system of the form in Equation 8 was used to place dynamic constraints on the lip shape.

$$G(s) = \frac{K\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{8}$$

Set points along the perimeter of the lip contour are forced to obey the dynamic constraints enforced by the second order system. The damping ζ and natural frequency ω_n were

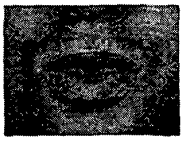

Figure 4: Final contoured lips.

chosen based on the frame rate of the captured lip sequence and their aesthetic appeal in tracking the lips in a manner similar to the natural dynamics of the lips.

4. CONCLUSIONS

An improved chromatic lip tracking algorithm has been proposed based on a simple thresholding technique. Using a fuzzy based approach a technique for calculating a stable threshold across an image sequence has been demonstrated. A method of using the connectedness of pixels along with an adaptive threshold has been demonstrated to accurately gain a true representation of a speaker's lips. The proposed algorithm is computationally inexpensive and can be implemented into a viable real time multimodal speech processing application.

5. REFERENCES

[1] S. Horbelt. Automatic Lipreading on the Basis of Image Sequences to support Speech Recognition. Master's thesis, Univeristy Erlangen-Nuremberg, April 1995.

[2] L. K. Huang and M. J. Wang. Image thresholding by minimizing the measure of fuzziness. *Pattern Recognition*, 28:41–51, 1995.

[3] M. Lievin and F. Luthon. Unsupervised Lip Segmentation under Natural Conditions. In *ICASSP' 99*, pages 3065–3068, Phoenix, Arizona, March 1999.

[4] S. Pigeon. The M2VTS database. Laboratoire de Telecommunications et Teledection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium, 1996.

[5] T. Wark and S. Sridharan. A syntatic approach to automatic lip feature extraction for speaker identication. In *ICASSP' 98*, pages 3693–3696, May 1998.

[6] T. Wark and S. Sridharan. A hybrid chromatic-parametric approach to automatic real-time lip tracking. Unpublished Report, May 1999.

[7] H. Yan Z. Chi and T. Pham. *Fuzzy Algorithms*. World Scientic, River Edge, New Jersey, 1996.