A SUITABILITY METRIC FOR MOUTH TRACKING THROUGH CHROMATIC SEGMENTATION.

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

In recent history the use of chromatic segmentation has come very much into vogue for mouth tracking. Our recent work has endeavored to show under what conditions and representations chromatic segmentation works. Results are presented showing for some members of the population chromatic segmentation does not work satisfactorily irrespective of recording conditions. A suitability metric is proposed that can give a quantitative measure on how well chromatic mouth tracking will work for a given subject.

1. INTRODUCTION.

The effective automatic location and tracking of a person's mouth is a problem that has proven very difficult in the field of computer vision. The term *mouth tracking* is used to include *lip tracking* as, the lips are a component of the mouth which contains other vital cues describing the mouth (ie. tongue, teeth, oral cavity). The lips however, act as an invaluable feature for tracking the mouth as in many cases the labial area gives a very good line of demarcation between the mouth and the face background. The outer labial contour of the mouth has very poor grayscale distinction when compared against its skin background making the segmentation of the mouth a difficult problem [1].

For mouth tracking, complex models that incorporate a priori knowledge of a mouth's shape and texture into an adaptive high dimensional energy minimisation problem, such as the active shape model implementation of Luettin [4] or the active appearance models used by Matthew [5], have been used. Such approaches have a number of problems with them, as they can be computationally expensive, have problems with convergence, require large amounts of pre-tracked data and offer no guarantees that the complex models are statistically stationary across a wide population.

Chromatic features have been shown useful for segmenting a person's mouth from that person's primarily skin background. Such techniques take advantage of the premise that mouth pixels, particularly those of the lips, are much redder than the paler skin background pixels they coexist in. These techniques have the advantage of being fast as they are simply pixel based and don't require any syntactic information to restrict the mouth shape during the segmentation stage. 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 [8]. Finally, there is the question of whether there is enough class distinction between the mouth and skin pixel chromatic representations to make such a segmentation plausible.

Previous work has shown that such distinction exists in many cases but the quantitative results for the use of chromatic mouth segmentation has been unclear. In many cases, in house data bases have been used with minimal subject variance with the results of which, not being readily comparable to pre-tracked data or other published or documented algorithms. Analysis has been carried out using the large and widely available audio visual M2VTS database [6] on the suitability of pixel based chromatic mouth segmentation across a reasonable number of subjects. Automatic tracking results for the database already exists [7] and we have also hand tracked a particularly large portion of the database for quantitative analysis of the chromatic segmentation procedure. In this paper we propose a metric to give a quantitative measure on the suitability of pixel based chromatic mouth segmentation given a subject's mouth and background distributions.

We present results in this paper that suggest, even under excellent lighting conditions, some members of the population lack sufficient chromatic class distinction for successful pixel based chromatic mouth segmentation. These experiments were conducted using the two most common chromatic representations for skin and lip segmentation [1, 7, 8] with both representations being insensitive to luminance.

2. FORMULATION OF PROBLEM.

The segmentation stage can be modeled as a two class problem where a pixel x taken from the mouth ROI can belong to either class ω_m or ω_b for mouth pixels and background pixels respectively. This can be expressed in terms of a decision rule, using Bayes theorem [3] as

$$\frac{p(x|\omega_m)}{p(x|\omega_b)} \begin{array}{c} \omega_m \\ > \\ < \\ \omega_b \end{array} \frac{P_b}{P_m} \tag{1}$$

where $p(x|\omega)$ is the conditional density function and P is the *a priori* probability. For simplicity $\frac{P_b}{P_m} = 1$ for our initial analysis for each class.

3. CHROMATIC REPRESENTATIONS

Recent work in the field of real time face tracking has used a normalised chromatic space model to characterize human faces [8]. Normalised chromatic space can be defined in Equation 2 based on an image in RGB space. It has been shown that human skin obeys an approximate normal-like clustering distribution in normalised chromatic space [8].

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B} \tag{2}$$

Work has been done by Sanchez [7] using normalised chromatic space to segment the mouth with results being presented using the M2VTS database. The ratio of red to green intensities ($\frac{R}{G}$) [1] have been used as a chromatic feature for mouth segmentation. This feature has been used on the pretense that they can accurately discern between redder and paler pixels whilst being relatively independent to fluctuations in luminance. Our work shall concentrate on comparing normalised chromatic $[r, g]$ and $\frac{R}{G}$ space in terms of stochastic complexity and class distinction.

4. TESTING DATA.

Of the 37 subjects in the M2VTS [6] database 36 were used with the subject ‘pm’ being excluded because of his beard. The first 3 shots of the database were used gain ample training data for segmentation analysis. For each subject and shot in the database the frames 1 to 100 were tracked at 10 frame intervals. This resulted in over 1000 pre-tracked frames with approximately 10 pre-tracked frames per subject per shot. The mouth ROI chosen for segmentation was based on the subject’s eye separation distance d_{eye}, with a $(3d_{eye}) \times (4d_{eye})$ box centred at the mouth centre.

As mentioned previously our human trackers, who were employed to manually track the mouth and eyes of the M2VTS subjects, reported much difficulty in labeling the outer labial contour. This highlights a fundamental problem associated with mouth tracking as the region between the skin and mouth can be uncertain. This suggests why approaches based on edge finding often fail with more success being found in segmentation techniques which treats the mouth much more like a texture. This boundary uncertainty can be somewhat reduced using post-processing that place temporal and syntactic restrictions on the resultant shape so as to reduce errors from the segmentation process [1].

5. CLASSIFIER DESIGN.

Theoretically a Bayes classifier is optimal [3] as it is the best classifier which minimizes classification error. However, it is very difficult to create such a classifier as it requires the true conditional density functions and true a priori class probabilities for each possible class. Practically this is infeasible as we can only ever get estimates due to sample size limitations and computational complexity. However, parametric classifiers can be used to approximate such density functions. Normally a chromatic segmentation classifier’s complexity is limited by the amount of training data available to it. This limits most classifiers to unimodal topologies or require the employment of other nonparametric techniques that require less training data. For our analysis we have used a vast amount of pre-tracked training data allowing one to investigate classifier topologies of increasing complexity. A well known type of classifier, which allows for increased model complexity, is a Gaussian Mixture Model (GMM). GMMs have benefits over other parametric classifiers as they give conditional density estimates which can be applied directly to a Bayesian framework as found in Equation 1.

A GMM models the probability distribution of a statistical variable z as the sum of Q multivariate Gaussian functions,

$$p(z) = \sum_{i=1}^{Q} \alpha_i N(z; M_i, \Sigma_i), \tag{3}$$

where $N(z; M, \Sigma)$ denotes a normal distribution with mean vector M, covariance matrix Σ and α denoting the a priori probability of class i. The parameters of the model (α, M, Σ) can be estimated using the Expectation Maximization (EM) algorithm [2].

Mixtures		Recognition error (%)	
Mouth	**Background**	$\frac{\mathbf{R}}{\mathbf{G}}$	$[\mathbf{r}, \mathbf{g}]$
1	1	8.73	8.80
2	1	9.14	9.39
4	1	9.34	9.56
2	1	10.41	11.39
2	2	10.87	11.95

Table 1: Average error rates for different GMM classifier topologies and chromatic features.

6. SEGMENTATION RESULTS.

Colour constancy [1] was a major problem in the mouth segmentation process even though the entire M2VTS database was recorded under similar lighting conditions for all subjects and shots. GMMs had to be built up for each subject and shot in the database as poor performance was received when models were generalised across subjects or shots.

The segmentation experiments were carried out across the entire pre-tracked database. For each shot of each subject 7 of the frames were used as training data to create the GMM mouth and background class density function estimates. The remaining 4 frames were used to test the resultant classifier. A number of GMM topologies were investigated with the results being given in Table 1 for normalised and $\frac{R}{G}$ chromatic space. The individual subject error rates are presented in Figure 1 for the two best error rates in Table 1, namely the unimodal topologies for normalised and $\frac{R}{G}$ chromatic space. The error rates were obtained by calculating the ratio of pixels, excluding pixels lying in the inner mouth cavity, not lying in their appropriate pre-tracked mouth/background areas.

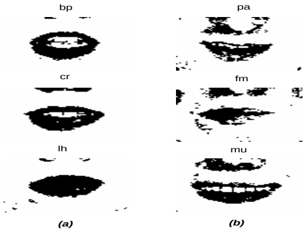

Figure 3: Sample segmented images taken from M2VTS database for shot 1. Columns (a) and (b) represent some well and poor segmented images respectively.

6.1. Calculating a suitability metric.

Given a subject's lip and background chromatic conditional density functions a quantitative measure is required to state whether such a segmentation is plausible without having to test the models against physical data. From the results in Table 1 one can see that the mouth and background distributions for both normalised and $\frac{R}{G}$ chromatic space are approximately unimodal with further stochastic complexity only degrading performance.

The Bhattacharyya distance μ is a convenient measure of the separability of two normal distributions [3] and gives an approximation of the upper bound of the Baye's error ε_{μ} between two unimodal distributions,

$$\varepsilon_{\mu} = \sqrt{P_1 P_2} \exp^{-\mu} \tag{4}$$

where P_1 and P_2 are the a priori probabilities of the two classes which can be assumed to be equal for our purposes. The Bhattacharyya distance μ can be decomposed into the summation of two terms.

$$\begin{aligned} \mu &= \mu_M + \mu_\Sigma \\ \mu_M &= \tfrac{1}{8}(M_2 - M_1)^T \left(\tfrac{\Sigma_1+\Sigma_2}{2}\right)^{-1} (M_2 - M_1) \\ \mu_\Sigma &= \tfrac{1}{2} ln \frac{|\frac{\Sigma_1+\Sigma_2}{2}|}{\sqrt{|\Sigma_1||\Sigma_2|}} \end{aligned} \tag{5}$$

where M_1 and M_2 are the means of the two classes and Σ_1 and Σ_2 are the covariance matrices of the two classes.

If we inspect the Bhattacharyya distances in Figure 2 we can see a definite correlation between ε_{μ} and the practical error rates received in Figure 1. A threshold was deduced empirically to be $\mu = 0.6$ approximately correlated to the 10% error threshold received in our practical results. If the Bhattacharyya distance μ between the mouth and background classes lies above this threshold then it can be assumed their is sufficient class distinction to segment the mouth effectively. The $\frac{R}{G}$ chromatic representation was used for calculating the suitability metric as the representation received similar performance to normalised chromatic space and was of a lower dimensionality making calculation easier.

7. DISCUSSION AND CONCLUSIONS.

It should be noted that object detection based purely on a chromatic signature is error prone at the best of times. What we are trying to establish in this paper is a suitability metric to act as a guide for when such a simplistic segmentation is possible. Due to there often being no definite boundary between the mouth and background the error results received in Figure 1 have to be analysed with some trepidation. Empirically, we found that an error rate above 10% results in a segmented image with minimal mouth shape information, examples of which can be seen in Figure 3. This poor class distinction can be attributed to a number of causes,

[1] Colour constancy refers to the ability to identify a surface as having the same color under considerably different viewing conditions [8].

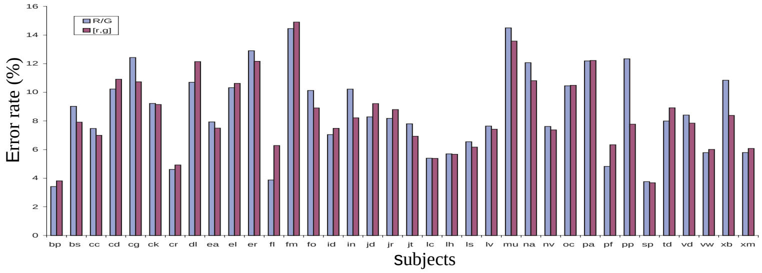

Figure 1: Segmentation error rates across the 36 subjects of the M2VTS database.

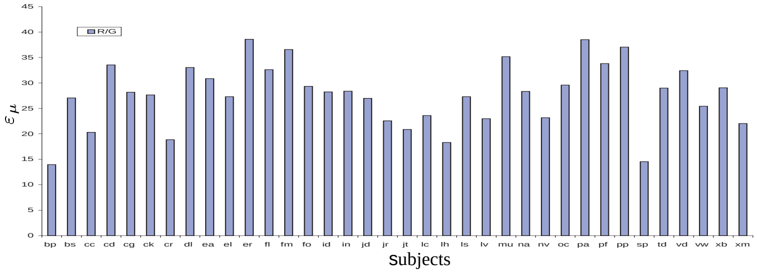

Figure 2: Upper Bayes limit across the 36 subjects of the M2VTS database.

- minimal lip area being displayed by subject or during certain mouth configurations (ie. very open mouth or pressed lips),
- low class distinction between classes with the mouth shape being corrupted by erroneous pixels (ie. shadows from mouth and nose).

These problems cannot be readily resolved within a simple pixel based segmenting procedure. Extra information must be brought to the problem. This is where our distinction between lip and mouth tracking is important as certain mouth configurations and subjects require other visual cues (ie. teeth and oral cavity) than the labial area for successful mouth tracking. Our suitability metric based on the Bhattacharrya distance between two unimodal distributions can act as a guide for when such extra cues are required.

In this paper we have shown that there is a marked difference in class distinction between mouth and skin pixel chromatic representations across a number of subjects. We have also demonstrated that both normalised and $\frac{R}{G}$ chromatic representations have approximately equal performance at the same stochastic complexity, but $\frac{R}{G}$ chromatic space has advantages due to its smaller dimensionality. A suitability metric has been described that can act as a guide for when chromatic pixel based mouth segmentation may be effective. Our future work shall try and improve segmentation when the mouth cannot be segmented based on the labial area alone.

8. REFERENCES

[1] G. Chiou and J. Hwang. Lipreading from Color Video. *IEEE Transactions on Image Processing*, 6(8):1192–1195, August 1997.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society*, 39:1–38, 1977.

[3] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Inc., 24-28 Oval Road, London NW1 7DX, 2nd edition, 1990.

[4] J. Luettin, N. Thacker, and S. Beet. Visual Speech Recognition using Active Shape Models and Hidden Markov Models. In *ICASSP 96'*, Atlanta, GA, May 1996.

[5] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. Bangham. Lipreading using Shape, Shading and Scale. In *AVSP'98*, 1998.

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

[7] M. U. Ramos Sanchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In *Int. Conf. Audio and Video based Biometric Person Authentication*, pages 69–76, Crans Montana, Switzerland, 1997.

[8] J. Yang and A. Waibel. A Real-Time Face Tracker. In *Proceedings of WACV'96*, pages 142–147, Sarasota, Florida, USA, 1996.