

Adaptive Mouth Segmentation using Chromatic Features

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
slucey@ieee.org, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

Abstract

The automatic segmentation of the mouth from its facial background is a very difficult computer vision problem due to the low grayscale distinction between classes. Recently chromatic based segmentation has enjoyed some popularity for the purposes of mouth tracking due to its ability to distinguish between the two classes. Such systems have to be highly adaptive due to problems with colour constancy. In this paper a technique for adaptive segmentation is investigated using an unsupervised clustering technique incorporating the expectation maximisation (EM) algorithm across a variety of chromatic features. Results are presented from the M2VTS database across a number of subjects.

Keywords: chromatic mouth segmentation, mouth tracking, colour constancy

1 Introduction

Segmentation is a common approach in computer vision to track the outline or shape of an object. However, with such an approach the accuracy or usefulness of the tracked object is directly related to how well the object has been segmented from its background. Segmentation of the mouth from its facial background is a very difficult problem due to low grayscale variation around the mouth (Chiou and Hwang, 1997). Chromatic pixel based features have shown to be useful for segmenting a person’s mouth from the primarily skin background (Chiou and Hwang, 1997; Lievin and Luthon, 1999; Lucey et al., 2000; Ramos Sanchez et al., 1997; Tian et al., 2000). 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.

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 (Yang and Waibel, 1996). Finally, the class models describing the chromatic distribution of pixels in both the mouth and background classes can vary from person to person. All these effects can be grouped together under the banner of a problem known in computer vision circles as *colour constancy*. Colour constancy refers to the ability to identify a surface as having the same colour under considerably different viewing conditions. The colour constancy problem requires a classifier, using chrominance as a feature for segmentation, to be as *adaptive* as possible. The term *adaptive mouth segmentation* is used in this paper to describe the task of segmenting the mouth from the surrounding skin background in an unsupervised manner such that an a priori parametric description of either class is not required.

The problem of low grayscale distinction between the mouth and its background has been previously addressed by complex models that incorporates 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 Luetttin et al. (1996) or the active appearance models used by Matthews et al. (1998). Such approaches have a number of problems with them as they can be computationally expensive, have problems with convergence, are highly non-linear, require large amounts of pre-tracked data and may require the models to be re-trained for new subjects. Conversely, simple segmentation techniques that contain minimal a priori knowledge of the mouth’s shape or texture are advantageous as they can be fast, owing to them being pixel based, and do not require any syntactic information to restrict the mouth shape during the segmentation stage.

Although chromatic mouth segmentation has enjoyed considerable success, most techniques have required the mouth and background class distributions to be known a priori (Chiou and Hwang, 1997; Ramos Sanchez et al., 1997; Tian et al., 2000) through manual tracking. Such a restriction can make the process

of mouth tracking through chromatic segmentation a practically infeasible task as new class distributions have to be manually found when a new subject is encountered or if there is a lighting change. Global class distributions taken from many subjects and environments tend to perform poorly on individual subjects due to the problems with colour constancy and the distributions being too general (Lievin and Luthon, 1999).

A soft clustering (Fukunaga, 1990) approach is employed to automatically deduce the mouth and background distributions from samples of pixels taken from a subject’s mouth ROI using a maximum likelihood criteria. Results are presented that can segment a subject’s mouth from its background with no a priori knowledge except the use of the initial starting clusters for the mouth and background classes. The M2VTS (Pigeon, 1996) database is used to present results for unsupervised mouth segmentation for a wide number of subjects.

In this paper a number of issues are investigated. Firstly, we investigate if mouth segmentation can be improved by including features describing the chromatic localised second order statistics of neighboring pixels. Secondly, mouth segmentation performance was tested for three chromatic representations in terms of pixel segmentation error rate, labial contour degradation and classifier complexity. An investigation of what chromatic representations perform best during unsupervised segmentation is also undertaken. The term *mouth tracking* is used in this paper to include *lip tracking*. The lips are a component of the mouth which contain other vital cues describing the mouth (i.e. tongue, teeth, oral cavity). The lips however, act as an invaluable feature for tracking the mouth as in most cases the labial area gives a very good line of demarcation between the mouth and the face background.

This correspondence is organised as follows. Firstly, in Section 2 the segmentation problem is formulated within a Bayesian decision context. Section 3 deals with what types of chromatic representations are to be tested with a new variant of a chromatic feature being introduced that takes into account the localised second order statistics of adjacent pixels. Section 4 outlines how and why the M2VTS (Pigeon, 1996) database was used for our tests and what error metrics were used for quantitative evaluation. Sections 6 and 7 present results for mouth segmentation over various classifier topologies and chromatic features for the supervised case, where both mouth and background distributions are known a priori, and the practical unsupervised case, where no a priori knowledge is given. Section 8 gives a brief discussion and some concluding remarks.

2 Formulation of problem

The segmentation stage can be modelled 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 and background pixels respectively. This can be expressed in terms of a decision rule, using Bayes theorem (Fukunaga, 1990) as

$$\frac{p(x|\omega_m)}{p(x|\omega_b)} > \lambda \quad (1)$$

where $p(x|\omega)$ is the conditional density estimate and λ is the decision threshold. In this paper we are dealing with both supervised and unsupervised segmentation approaches, for the unsupervised case there is no viable way of accurately calculating λ . For simplicity $\lambda = 1$ was used for our initial analysis.

3 Chromatic representations

Recent work in the field of real time face tracking has used a normalised chromatic space model to characterize human faces (Yang and Waibel, 1996). 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 (Yang and Waibel, 1996).

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

Ramos Sanchez et al. (1997) used normalised chromatic space to segment the mouth and presented results using the M2VTS database. The ratio of red to green intensities ($\frac{R}{G}$) (Chiou and Hwang, 1997; Lucey et al., 2000) has also been used as a chromatic feature for mouth segmentation. Lievin and Luthon (1999) used a hue logarithmic representation for their unsupervised segmentation experiments which can also be expressed as a ratio of red to green. Both the $[r, g]$ and $\frac{R}{G}$ features have been used under the pretense that they can provide distinction between redder mouth and paler skin pixels whilst being relatively independent to fluctuations in luminance.

To try and improve the class distinction between the mouth and the predominantly skin background we have tried to use some features that take into account the localised second order statistics present in adjacent pixels. To this end we have included extra features based on an image in $\frac{R}{G}$ feature space as described in Equations 3 and 4.

$$SD(i, j) = \ln \left(\sqrt{\frac{\sum_{p=i-3}^{i+3} \sum_{q=j-3}^{j+3} [\frac{R}{G}(i, j) - AI(p, q)]^2}{49}} \right) \quad (3)$$

$$DI(i, j) = \frac{R}{G}(i, j) - AI(i, j) \quad (4)$$

where

$$AI(i, j) = \frac{\sum_{p=i-3}^{i+3} \sum_{q=j-3}^{j+3} \frac{R}{G}(p, q)}{49} \quad (5)$$

Equation 3 is a measure of heterogeneity of a 7x7 region of which the pixel is in the center. Equation 4 is a measure of relative intensity of a pixel to its neighbors. These features can be vectorized and concatenated into one feature set $[\frac{R}{G}, AI, SD]$. Our work shall concentrate on comparing normalised chromatic $[r, g], \frac{R}{G}$ and $[\frac{R}{G}, AI, SD]$ features in terms of stochastic complexity and class distinction for supervised and unsupervised segmentation.

4 Test data

Of the 37 speakers in the M2VTS (Pigeon, 1996) database, 36 were used with the subject ‘pm’ being excluded due to his beard. The first 3 shots of the database were used to gain ample training data for segmentation analysis. For each speaker in the database the eye positions as well as the outer and inner labial contour were *manually* tracked from frames 1 to 100 at 10 frame intervals. This resulted in over 1000 pre-tracked frames with approximately 11 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 human trackers, who were asked to manually track the mouth and eyes of the M2VTS subjects, reported much difficulty in labelling the outer labial contour. This highlights a fundamental problem associated with mouth tracking as the uncertainty between the skin and mouth regions also suggests why approaches based on edge finding often fail. Segmentation techniques that treat the mouth much more like a texture have been more successful. This boundary uncertainty can be reduced using post-processing that place temporal and syntactic restrictions on the resultant shape so as to reduce errors from the segmentation process (Chiou and Hwang, 1997; Lucey et al., 2000; Tian et al., 2000).

This uncertainty also makes it difficult to gain an accurate quantitative measure of how effective a certain segmentation technique is. In this paper the recognition rate of correctly labelled pixels (i.e. mouth or background) in the mouth ROI is used so as to gain a rough quantitative measure of how well the segmentation process has gone. Unfortunately, this metric alone cannot be accurately used as poor error rates can be received from the subject’s mouth being open or nose being segmented along with the mouth even though the mouth has been segmented accurately. To remedy this situation, visual inspection of the segmented image was also used in order to gain a further qualitative assessment.

5 Classifier design

Theoretically a Bayesian classifier is optimal (Fukunaga, 1990) 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. Practically this is infeasible as we can only ever get estimates due to sample size limitations and analytical 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, limiting most classifiers to unimodal distributions or employing other nonparametric techniques that require less training data. For our analysis we have used a reasonable amount of pre-tracked training data allowing one to investigate classifier topologies of increasing complexity. A well known classifier design which allows for increased model complexity are Gaussian mixture models (GMM). GMMs have benefits over other parametric classifiers as they give conditional density function 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,

$$p(z) = \sum_{i=1}^Q \alpha_i N(z; M_i, \Sigma_i), \quad (6)$$

where $N(z; M, \Sigma)$ denotes a normal distribution with mean vector M , covariance matrix Σ and α denoting the mixture weight of class i . The parameters of the model (α, M, Σ) can be estimated using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). K-means clustering (Gersho and Gray, 1992) was used to provide initial estimates of these parameters.

6 Classifier topology and features results

Before investigating how to adaptively segment the mouth from subject to subject one has to find which features and classifier topologies perform best given pre-labelled data (i.e. supervised segmentation). 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 $[r, g]$, $\frac{R}{G}$ and $[\frac{R}{G}, AI, SD]$ features. The error rates were calculated by counting how many pixels were misclassified over the total number of pixels in the mouth ROI image given the known labels from the pre-tracked data. The average error rates across all 36 subjects and 3 shots are presented in Table 1.

As can be seen in Table 1 there is some benefit in using the localised chromatic second order statistics of the mouth as a discriminative feature. The best results were received using a GMM classifier with a topology of a single mode for the mouth class and a single mode for the background skin class. Further stochastic complexity degraded the over all performance. It must be noted that the use of a single mode classifier no longer warrants the term GMM as the name implies that a mixture of modes is being employed in the classification task. However, for clarity and simplicity we shall entertain this stretch in terminology for the duration of this paper. A subject by subject breakdown of the error rates averaged over all 3 shots can be seen in Figure 1 for the best GMM topologies for each feature set as indicated by the (*) in Table 1. Clearly, the ability of the classifier to segment the mouth ROI image fluctuates from subject to subject. One must realise that the error rates received in Table 1 and Figure 1 must be treated with some trepidation as they are only an approximate quantitative indication of how well the segmentation process has gone. The real results can only be seen in the final segmented image. An example of some segmented images can be seen in Figure 2 again highlighting the superior performance of the $[\frac{R}{G}, AI, SD]$ over the other feature sets and the validity of the error metric used in Table 1 and Figure 1.

Figure 2 displays some segmented images for all three feature sets being tested across four subjects from the M2VTS database. The four subjects ‘er’, ‘cg’, ‘ck’ and ‘fm’ were chosen as they demonstrated some of the benefits and problems with each chromatic representation as well as some inherent problems with chromatic segmentation in general. The benefits from using the $[\frac{R}{G}, AI, SD]$ can be clearly seen in ‘er’ with the resultant segmented image being much cleaner than other chromatic representations. Subjects ‘cg’ and ‘ck’ demonstrated some of the problems with the $[\frac{R}{G}, AI, SD]$ feature set as the physical outline of the mouth was not as clear in comparison to the purely pixel based $\frac{R}{G}$ and $[r, g]$ features. The segmented images for ‘cg’ also highlight some problems with the segmentation pixel error rates used in Table 1 and Figure 1. Superior error rates were received for $[\frac{R}{G}, AI, SD]$ due to its ability to segment the oral cavity along with the lips and not purely on how accurate the segmentation was. Subject ‘fm’ represents a major problem for chromatic mouth tracking as all feature sets failed to segment the mouth adequately. This highlights a major flaw in the philosophy behind chromatic mouth segmentation as there is often an unsubstantiated assumption that there will always be sufficient chromatic distinction between the mouth and background,

Mixtures		Recognition error (%)		
Mouth	Background	$\frac{R}{G}$	[r, g]	$[\frac{R}{G}, AI, SD]$
1	1	(*)10.91	(*)10.08	(*)8.98
1	2	12.06	10.28	9.03
2	1	11.25	10.93	9.40
2	2	12.99	12.01	9.80
4	1	12.59	11.27	-

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

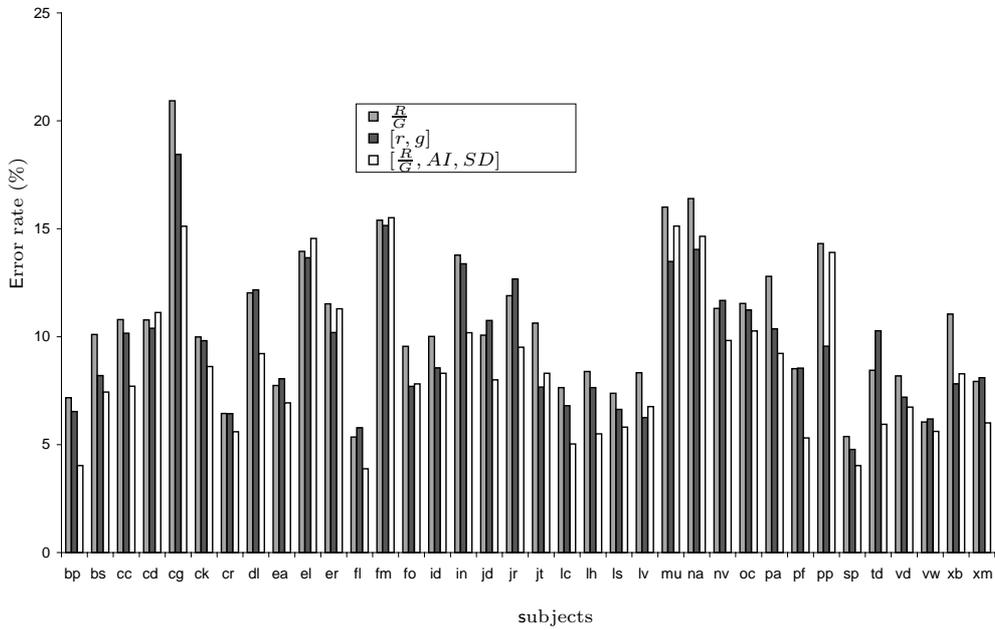


Figure 1: Supervised segmentation error rates across the 36 speakers of the M2VTS database.

		Features			
		original	$\frac{\mu}{\sigma}$	$[r, g]$	$[\frac{\mu}{\sigma}, AI, SD]$
Subjects	ck				
	cg				
	er				
	fm				

Figure 2: Segmented images across some subjects of the M2VTS database through supervised segmentation.

and all one has to do is find the correct conditional class density functions. This is not always the case. In all subjects segmentation of the nasal cavity as distinct from the lips was a problem.

7 Clustering using the EM algorithm

The problem of clustering has been well defined and investigated in pattern recognition (Fukunaga, 1990). *Clustering* is defined by Fukunaga (1990) as the classification of samples without the aid of a training set. Such a clustering approach is ideal for chromatic mouth segmentation due to problems with colour constancy and the need for subject by subject adaptivity as reported by Lievin and Luthon (1999). An important goal of finding clusters is to decompose a complex distribution into several normal-like distributions. By expressing a complex distribution through the summation of a several normal-like distributions the problem of pattern recognition especially classifier design becomes considerably easier. This is the main philosophy behind GMM classifiers. The distinction between the terms class distributions and clusters can often become ambiguous when one is dealing with clustering techniques. This is due to the assumption that a cluster or groups of clusters found during the unsupervised clustering process refers to a real world class (i.e. mouth or background).

The estimation of clusters can be a problem when real world classes overlap. Unfortunately, this is often the case with chromatic representations of the mouth and the background class distributions. Most unsupervised clustering algorithms are either *hard* or *soft* clustering. Hard clustering refers to a clustering algorithm that can only allow a given sample from a data set to belong to one particular cluster. Conversely, soft clustering allows a given sample from a data set to belong to multiple clusters. When the distinction between classes is poor, a hard clustering approach requires samples to be assigned to specific classes even though in reality there may be a large amount of uncertainty to which cluster the sample really belongs. This can result in ill formed clusters which can drastically affect the ability to segment the mouth successfully.

Maximum likelihood (ML) estimation is able to perform a soft clustering such that there are no definite boundaries between classes. The EM algorithm (Dempster et al., 1977) was chosen to perform the clustering process due to its ability to cluster via a ML criterion. In essence the problem can be thought of as estimating a Q mixture/cluster GMM as previously defined in Equation 6 such that an estimate of the class density function can be made as is required for the pixel segmentation process defined in Equation 1. However a problem using the EM algorithm (Dempster et al., 1977) and other iterative clustering algorithms is the calculation of the initial parameters (α, μ, Σ) of each cluster and how many clusters to have for each class.

This initial guess can have some serious ramifications as it can affect the convergence and final estimate of the class density function. A common problem with automatic clustering can also present itself in trying

to work out which clusters or in some cases group of clusters represent which real world classes (i.e. mouth or background) after the automatic clustering process. Fortunately, both these problems can be used to some advantage by the use of a *generic model*. Using a priori knowledge of what the generic (i.e. typical) class density functions of the mouth and background are one has a good initial guess of what the subject’s conditional class density functions are, how many clusters each class should have and what final clusters refer to what real world classes. These generic models were constructed across all the pre-tracked training data from all 36 subjects and 3 shots using a priori knowledge of what the mouth and background classes were.

7.1 Clustering results

Results are presented in Table 2 for adaptive clustering using the EM algorithm. In all cases a two mixture *generic model* was used as the initial starting point for the EM algorithm with one mixture being used for the mouth and background classes respectively. Single mixtures were used for each class in the unsupervised scenario for two reasons. Firstly, the supervised experiments suggested that unimodal classifiers performed best over more complicated GMM topologies. Secondly, preserving single mixtures for each class made the unsupervised clustering process easier, as each cluster directly related back to the mouth or background class.

The unsupervised clustering process was carried out in a similar manner to the supervised segmentation carried out in Section 6 with the first 7 frames of the pre-tracked database being used for clustering. The EM algorithm was iterated on the training data set for each subject and shot, using 10 iterations to ensure convergence. The resultant clusters were then tested on the remaining 4 frames for each subject and shot so as to gain a pixel segmentation error rate. A thorough breakdown of these error rates can be seen in Figure 3 for each subject averaged over the 3 shots used. The results show that the $[r, g]$ chromatic feature set outperforms its counterparts within an unsupervised situation.

Figure 4 demonstrates segmentation results for subjects ‘bp’, ‘ck’, ‘dl’ and ‘lv’ from the M2VTS database. These subjects were chosen to highlight some of the typical results obtained when using unsupervised segmentation techniques. The subject ‘ck’ was included again so as to act as a comparison between the supervised results shown in Figure 2 and unsupervised shown in Figure 4. The comparison is quite effective as there is little difference between supervised and unsupervised images for the $\frac{R}{G}$ and $[r, g]$ features. Subjects ‘dl’ and ‘lv’ demonstrate the superior performance of the $[r, g]$ feature, with superior mouth segmentation and shape, in comparison to the $\frac{R}{G}$ and $[\frac{R}{G}, AI, SD]$ features. The visual performance difference between the $[r, g]$ and $\frac{R}{G}$ is quite small but the $[r, g]$ tends to perform slightly better with less noisy pixels around the mouth. In all cases the $[\frac{R}{G}, AI, SD]$ segmentation was poor resulting in lots of noisy pixels and poor

Features	Recognition error (%)
$\frac{R}{G}$	12.21
$[r, g]$	10.14
$[\frac{R}{G}, AI, SD]$	12.30

Table 2: Average error rates for unsupervised clustering using single clusters for the mouth and background classes across various chromatic features.

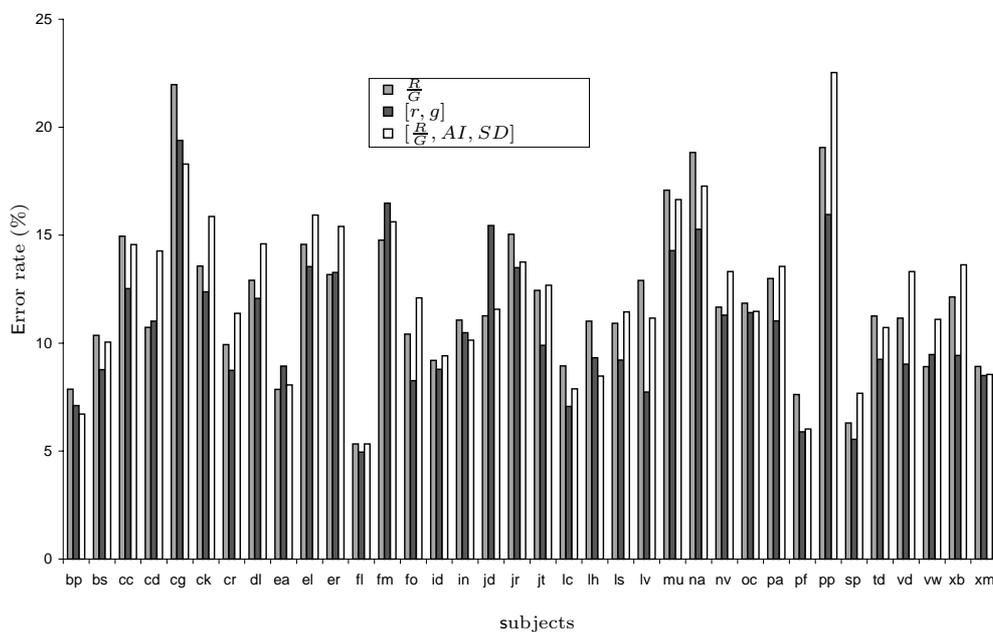


Figure 3: Unsupervised segmentation error rates across the 36 speakers of the M2VTS database.

shape extraction.

8 Conclusions and discussion

A number of chromatic features and GMM classifier topologies have been investigated for the purposes of chromatic mouth segmentation in this paper. A new chromatic feature, that used the localised second order statistics of neighboring pixels, was introduced that proved effective for supervised segmentation of the mouth. However, the feature's performance was close to the performance of the other two feature sets and in a lot of cases caused the degradation of the outer mouth contour due to the smoothing nature of the second order features. The smoothing nature of the features rather than the inclusion of the second order information was attributed to the performance improvement received and could have just as easily been achieved through a post-processing filter or morphological operation (i.e. a morphological clean and close operation on the segmented binary image) to remove any noisy or spurious pixels. There was no clear benefit in including the second order chromatic features in the segmentation stage and significant detriment in the unsupervised segmentation.

An interesting result was received in the supervised segmentation stage with unimodal distributions performing best for the mouth and background classes for all features tested. This was surprising as one would have initially thought the introduction of extra modes/classes could better segment the inner cavity of the mouth when it is open (i.e. teeth, tongue). This highlighted a problem with using normalised chromatic features as they have had their luminance component removed which would aid in the segmentation of the inner mouth but unfortunately degrades the segmentation performance of the lips. Without the luminance component the class distinction between the background skin and oral cavity is very poor in normalised chromatic space. This can be reinforced by visual inspection of the results received in Figure 2 for open mouths. The inner cavity does not get segmented at all even though inner cavity information was included in the supervised training stage.

An unsupervised mouth segmentation technique was introduced using the EM algorithm. Good performance was received for the $\frac{R}{G}$ and especially the $[r, g]$ feature. The problem of evaluating mouth segmentation performance was examined. A quantitative measure based on the number of misclassified pixels was used as well as a qualitative assessment based on the visual inspection of the segmented image. Results have been presented using the widely available M2VTS database.

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 $[r, g]$, $\frac{R}{G}$ and $[\frac{R}{G}, AI, SD]$ feature sets perform best using simple unimodal distributions for the mouth and background

		Features			
		original	$\frac{R}{G}$	$[r, g]$	$[\frac{R}{G}, AI, SD]$
Subjects	bp				
	ck				
	dl				
	lv				

Figure 4: Segmented images across some subjects of the M2VTS database through unsupervised segmentation.

classes even with ample training data. There were some cases in which chromatic mouth segmentation failed irrespective of the feature set, classifier or whether the process was supervised. This highlights a major problem in chromatic mouth segmentation as more complex approaches need to be brought to bear on subjects with low mouth/background distinction.

9 Acknowledgements

The authors would like to thank the M2VTS Project for use of their database.

References

- Chiou, G. and Hwang, J. (1997), Lipreading from Color Video, *IEEE Transactions on Image Processing* **6**(8), 1192–1195.
- Dempster, A., Laird, N. and D. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *Royal Statistical Society* **39**, 1–38.
- Fukunaga, K. (1990), *Introduction to statistical pattern recognition*, 2nd edn, Academic Press Inc., 24-28 Oval Road, London NW1 7DX.
- Gersho, A. and Gray, R. (1992), *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 3300 AH Dordrecht, THE NETHERLANDS.
- Lievin, M. and Luthon, F. (1999), Unsupervised Lip Segmentation under Natural Conditions, *ICASSP' 99*, Phoenix, Arizona, pp. 3065–3068.
- Lucey, S., Sridharan, S. and Chandran, V. (2000), Robust Lip Tracking using Active Shape Models and Gradient Vector Flow, *Australian Journal of Intelligent Information Processing Systems* **6**(3), 175–179.
- Luettin, J., Thacker, N. and Beet, S. (1996), Visual Speech Recognition using Active Shape Models and Hidden Markov Models, *ICASSP 96'*, Atlanta, GA.
- Matthews, I., Cootes, T., Cox, S., Harvey, R. and Bangham, J. (1998), Lipreading using Shape, Shading and Scale, *AVSP'98*.
- Pigeon, S. (1996), The M2VTS database, Laboratoire de Telecommunications et Teledetection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium.

- Ramos Sanchez, M. U., Matas, J. and Kittler, J. (1997), Statistical chromaticity models for lip tracking with B-splines, *Int. Conf. Audio and Video based Biometric Person Authentication*, Crans Montana, Switzerland, pp. 69–76.
- Tian, Y., Kanade, T. and J.Cohn (2000), Robust Lip Tracking by Combining Shape Color and Motion, *ACCV'2000*, pp. 1040–1045.
- Yang, J. and Waibel, A. (1996), A Real-Time Face Tracker, *Proceedings of WACV'96*, Sarasota, Florida, USA, pp. 142–147.