A Combined Approach to Feature Extraction for Mouth Characterization and Tracking

Ioana Bacivarov, Mircea C. Ionita, Peter Corcoran

Department of Electronic Engineering
National University of Ireland, Galway
email :{ ioana, mircea, pcor}@wuzwuz.nuigalway.ie

***Abstract*— A statistical Active Appearance Model (AAM) is developed to model and track the lips. A major challenge in modelling the mouth area, considered the most deformable facial feature, is the weak contrast between lips and skin colour. In order to perform accurately, AAM requires a strong initialization point. In our approach, key information is firstly extracted using chrominance analysis. Then optimal parameters of the model are determined using AAM fitting procedure. A detailed description of our model and its corresponding fitting algorithm are given and preliminary results on different databases are summarised.**

***Keywords* – statistical model of appearance, mouth detection, mouth tracking, chromatic analysis.**

I INTRODUCTION

Recent interest in facial modelling and facial features is urged by the rapid advances in technology. However, a major challenge in face modelling is analysing the mouth area, regarded as the most deformable facial feature. Moreover, lips are considered a weak feature because of their complex expression and varying pattern, but also because of their weak colour contrast and significant overlap in colour features with the face region. Factors such as differing illuminations and skin reflectance or occlusions such as teeth and tongue further complicate the task.

Mouth feature extraction represents an important stage of face image analysis for numerous application areas, like affective computing, videoconferencing, video-telephony, determining driver drowsiness, automatic bimodal speech recognition, face or visual speech synthesis. In the literature, several methods have been proposed and they can be mainly categorized [1], [2] as feature-based techniques, model-based techniques or a combination of both, i.e. hybrid approach.

The first category, the feature-based techniques, has the advantage of representing simple methods which work well under environmental constraints, but lack in robustness for real conditions. A key step in the localisation of mouth features is that of lips extraction. This is primarily based on colour-based segmentation techniques and usually the processing is done in the Hue, Saturation and Value (HSV) space [3], [4]. This colour space brings invariance to shadows, shading and highlights and permits using only the hue component for segmentation. In [2], [5], segmentation of the colour lip image is achieved by spatial fuzzy clustering. In [6] the notion of colour constancy is used. The segmentation is further improved by including features which describe chromatic second order statistics of neighbouring pixels.

Model-based techniques or holistic approaches (Deformable Templates [7], [8], Active Shape Models [9], [10], Snakes [10]) compensate for the drawbacks of the appearance methods, but they are computationally expensive and dependent on constraints learned from examples. It is common to see non-optimal local alignments when attempting to model an object that is quite different from the training set, especially when dealing with features that can fade into the background, such as the lips in our case.

In [3]. Pantic et al. combines both techniques in search for a robust mouth detector. The author develops a hybrid, knowledge-based technique for mouth feature extraction from facial images. At first coarse mouth detection is performed by applying a hue filter. Then a spatial sampling of the mouth contour is performed. Finally. the mouth movements are categorized.

In our work, we have employed the idea of using a hue filter to obtain a rough initialization of the lips features. The optimal parameters of the filter were determined by using a statistical analysis over the lips area for different conditions and subjects. Then, an estimated mouth region was determined from the filter. This was used as an initialization point for an AAM [11] of the mouth. By re-applying these techniques across several image frames of a video sequence we were then able to build a robust mouth tracking algorithm.

This paper is organised as follows. Section II details the statistical model of appearance adapted for the mouth region, while in Section III a corresponding fitting algorithm for unseen pictures is developed together with a mouth tracker. Other applications are proposed. Section IV presents some preliminary results for different databases, followed by conclusions and proposals of future work in Section V.

II MOUTH MODEL

Statistical Model of Appearance offers a form of 2D affine face model which can quickly match to the texture and shape of a detected face region. Similar statistical models can be developed for other deformable objects, sub-regions of a detected face such as eyes [12], nose or mouth. The manner in which the mouth model is constructed will permit us to extract mouth characteristics and to develop different applications (Section III).

a) Statistical Models of Appearance

Statistical Models of Appearance [11] represent both shape and texture variations and correlations between them. The desired shape to be modelled is annotated by a number of landmark points. The shape is defined by the number of landmarks chosen to best depict the contour of the object of interest, in our case the eye region. A shape vector is given by the concatenated coordinates of all landmark points and may be formally written as $s = (x1, x2, ...xL, y1, y2, ...yL)^T$, where L is the number of landmark points.

The shape model is obtained by applying Principal Component Analysis (PCA) on the set of aligned shapes:

$$s = \bar{s} + \varphi_s b_s \qquad (1)$$

where $\bar{s} = \frac{1}{N_s}\sum_{i=1}^{N_s} s_i$ is the mean shape vector, and

Ns is the number of shape observations; φ_s is the matrix whose columns are the eigenvectors; b_S defines the set of parameters of the shape model.

The texture, defined as the pixel values across the object of interest, is also statistically modelled. Using a triangulation algorithm, face patches are first warped into the mean shape. Then a texture vector $t = (t1, t2, ...tp)^T$ is built for each training image by sampling the values across the warped, i.e. shape normalised patches, with p being the number of texture samples.

The texture model is also derived by means of PCA on the texture vectors:

$$t = \bar{t} + \varphi_t b_t \qquad (2)$$

where $\bar{t} = \frac{1}{N}\sum_{i=1}^{N_t} t_i$ is the mean texture vector, with

N_t being the number of texture observations; φ_t is the matrix of eigenvectors, and b_t the parameters for the texture model.

The sets of shape and texture parameters

$$c = \begin{pmatrix} W_s b_s \\ b_t \end{pmatrix} \qquad (3)$$

are used to describe the overall appearance variability of the modelled object, where W_S is a vector of weights used to compensate the differences in units between shape and texture parameters.

After a statistical model of appearance is accomplished, the AAM algorithm can be employed to fit the model to a new image. AAM is a fast technique used to interpret unseen images using the appearance model, by finding the best match of the model to the image. Therefore, the algorithm allows us to find the parameters of the model which generates a synthetic image as similar as possible to the target image. The whole problem is treated as an optimisation problem in which we want to minimise the difference between the new image g_S and the image synthesised by the appearance model g_m:

$$\delta g = g_s - g_m \qquad (4)$$

We evaluate the current error $E = |\delta g|^2$ and minimise it by varying the shape and texture parameters c with a predicted displacement, $\delta c = A\delta g$. A is a pre-computed matrix of prediction, estimated once during the model building stage, that makes it run-time efficient. In practice, A is estimated by a set of experiments on the training set.

Then the image is resampled at the new prediction, a new error vector is calculated and if the resulting error is less than the precedent error, the new estimate is accepted. The procedure is iterative.

b) Mouth Model Design

A statistical model was built using the techniques that they are going to be described in section III.a. A detailed annotation was performed for the mouth region, using 24 landmark points. It was empirically determined that a suitable annotation would be as the one illustrated in Figure 1. If the mouth is closed, the corresponding points from the inner lip contour are merged (point number 20 overlaps number 18, the 21st point overlaps the 17th point and so on).

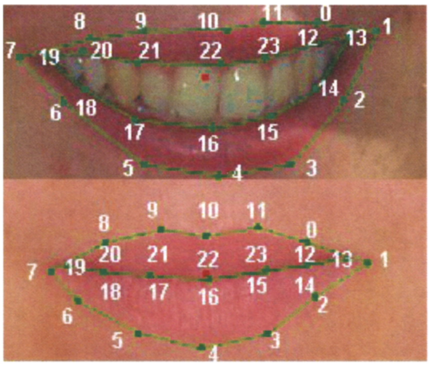

Figure 1: Mouth annotation for different expressions.

Mouth appearance is one of the most variable features and it is conditioned by many factors as skin colour, make-up, illumination, head pose, mouth expression or presence of teeth and tongue. In our training set we used various pictures, which were representative of the factors of variability mentioned above. The only condition was to annotate all images using the same number of points mapped in the same order.

Figure 2 shows the effect of varying with +/- one unit of standard deviation some representative parameters of the model: texture (first row) and shape (last three rows). The shape parameters could be differentiated in parameters describing the opening of the mouth, smiling action, head rotation, etc.. which depend on the variety described for the training set.

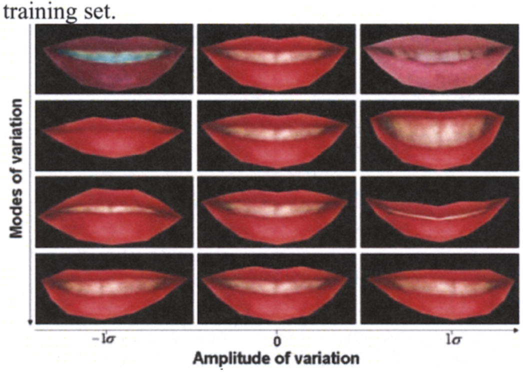

Figure 2: Modes of variation for the mouth model by +/- one standard deviation from the mean. The variation from the mean of texture parameters is represented in the first row, while the variation of different expressions and poses are exposed in the last rows.

III MOUTH FITTING ALGORITHM

a) System Overview

Firstly, before mouth features can be extracted and analysed, the face must be detected in an image and then its features must be traced. The face was inferred from the Viola-Jones face detection algorithm [13] applied on the input image. Then, our region of interest (ROI), i.e. the mouth region, was deducted from the rectangle describing the surroundings of the face. ROI was reduced then to the lower third on the y axis, while 3/5 of the face box was retained on the x axis, as shown in Figure 3.

AAM is a fast technique used to interpret unseen images using the appearance model, but it depends strongly on a good initialization. Using the information that colour of the lips is red, a hue filter was used to roughly locate mouth area (Section III.b), followed by a fine mouth detection and parameters extraction using AAM techniques (Section III.c).

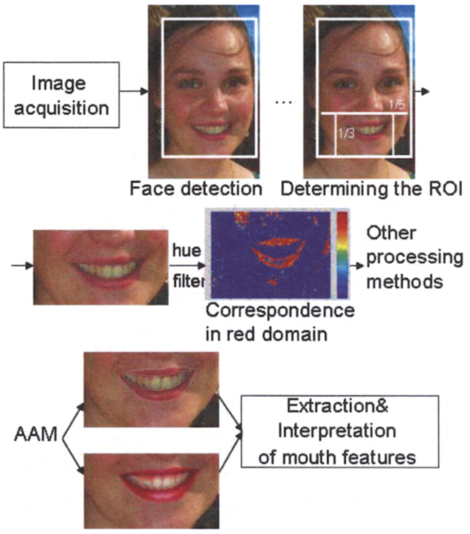

Figure 3: System Overview.

b) Chrominance analysis

The model needed a stronger initialization point in order to be fitted accurately to unseen images, consequently a pre-processing method that would provide a robust starting point. The most valuable information related to lips is their red colour, though red varies with respect to individuals, make-up, illumination etc.

Therefore, by filtering the red lip colour from the face region, we were able to identify the ROI. In a first step, the input image was transformed into the HSV colour space; hue representation is less affected by variations in illumination. Then the object of interest was filtered in the red domain, by applying the following hue filter [14]:

$$f(h) = \begin{cases} 1 - \frac{(h - h_0)^2}{w^2} & |h - h_0| \le w \\ 0 & |h - h_0| > w \end{cases} \qquad (5)$$

Where h is the shifted hue value of each pixel so that $ho = 1/3$ for red colour, ho controls the positioning of the filter and w controls the width of the filter.

As the colour for mouth region varies with respect to person identity, light conditions, make-up etc., the challenge was to find optimal parameters for the filter. Although an optimal solution would be to develop an adaptive hue filter as in [3], the simplified solution adopted in our case was to find optimal parameters for defined conditions, e.g. for a specific database.

A statistical analysis was performed on the FERET database [15], [16]. The variation of lip colour between individuals and the differences caused by varying illuminations on lips pixels for the same picture were investigated. It was noticed that standard deviation does not vary much from picture to picture, as the pictures belong to the same database, with controlled conditions. The filters coefficients were chosen after performing an average on mean and standard deviation for all pictures. The overall mean is 0.03 and it corresponds to the positioning of the filter described by h_0. The overall standard deviation approximating the filter width, w coefficient, is 0.005.

After determining the parameters of the filter and performing the actual filtering operation, each image was binarised using a predefined threshold; in our case we set the threshold value to 0.5. Morphological operations such as closing, followed by opening. can be used in order to fill in gaps and eliminate pixels that do not belong to the mouth area.

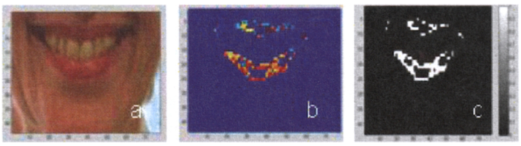

Figure 4: Mouth region pre-processing: a. original image, b. after the hue filter, c. after the binarisation.

After the lips region was determined, its centre of gravity (COG) was calculated. This point is going to be used as the initialization point for the AAM fitting algorithm.

c) Fine mouth detection and feature extraction

In a second step, AAM was applied in order to perform a fine detection and mouth features extraction. The starting point for the algorithm was the COG of the ROI. determined after applying the hue filter and binarising the image. The AAM adjusts the parameters so that a synthetic example is generated, which matches the image as close as possible (as seen in Figure 3). Optimal texture and shape parameters were determined for the image; in consequence information regarding mouth features was learned.

d) Mouth tracking & other applications

A possible application for the mouth model and its corresponding fitting algorithm was a robust mouth tracker. Robustness to mouth movements, slight head rotation and posing could be obtained in this manner.

In this section, we describe a framework for efficiently tracking the lips through an image sequence. Firstly, the mouth model was initialised using the algorithm described in Section III, applying a hue filter to find the COG of mouth pixels, used next as the initialization point for the AAM algorithm. The purpose of this stage was to determine a strong mouth position from the first frame of the video sequence. This step should be executed on the first frame and periodically repeated. Once this stage was performed, the model was able to deform and accurately track lip movements. This step was the most computationally expensive, as in next frames only an update of the shape parameters was performed.

An interesting consumer application that can be developed using our model is a smile detector. If an accurate analysis of the mouth area is accomplished, it is possible to build a smile detector based on an educated choice of geometrical distances between lips. This type of information is provided by the shape parameters determined after the fitting stage. The size of the mouth for its different states can be calculated relatively to the face size, which was determined after the face detection stage. or relatively to the distance between the eyes (distances that stay constant even if the subject is smiling) by using a smiling and non-smiling picture of the subject. The aperture of the mouth can be considered as well.

IV Preliminary Results

The proposed algorithm was tested on a specialized audiovisual speech database [17], VidTIMIT Video Dataset which is comprised of video recordings of different subjects reciting short sentences, suitable for tracking. Another standard database tested was the FERET database. Tests were also performed on our general collection of images, in which subjects presented different facial expression, various poses, make-up, illumination conditions, image resolution, etc.

Although, ideally, the hue domain should not be influenced by variations in illumination, this is not observed in practice. Therefore, changes of illumination conditions can cause important variations in the hue component for the same object. Standard face databases are normally created using a controlled environment and thus the hue component of all the images is much more consistent regarding differences between individuals. In the general case of unknown illumination conditions, the parameters of the filter mentioned above have to be adapted for each particular image, which is actually a very complex task. Yet, for a controlled environment. e.g. working with a standard database, the filter parameters, i.e. the mean and the standard deviation of the hue component of the lips. can now be estimated from within the particular database which is tested.

A statistical analysis was performed in order to investigate each database, from the point of view of the effect of the acquisition system over the texture, illumination variations, etc. of the pictures. Following this analysis, optimal parameters of the hue filter were considered.

For our tests. we have used an adaptation of the appearance modelling environment FAME [18], modifying and extending it to accommodate the techniques described in this paper. In order to reduce the computational complexity during fitting stage, we used a predefined reduced resolution of the face patch. Examples of the results of the proposed fitting algorithm obtained for different pictures from FERET database are presented in Figure 5. The shape and texture parameters after the fitting stage are depicted. The first two columns represent the results of the AAM algorithm. used as a stand-alone method, while the last two columns are the consequence of the hue filter combined with AAM, the latter used as refinement method.

The evaluation of the fitting algorithm as well as the evaluation of the tracking system was done by visual inspection. For FERET database, the system presents 77% fitting accuracy if a hue filter was used to determine a precise initialization point for AAM. If only AAM was employed and the initialization was determined in the face detection step, only 38% of the test images presented a correct fitting. The results showed the benefits of an accurate initialization point for the AAM method.

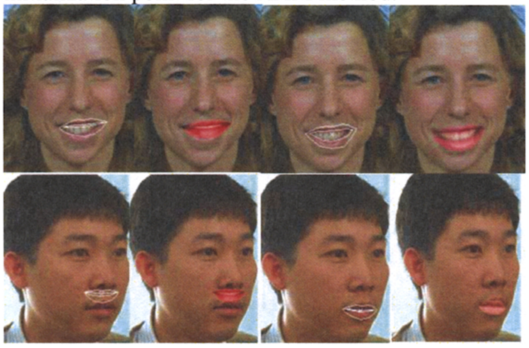

Figure 5: Mouth fitting results applying the AAM algorithm

For the tracking experiments, short movies of sequences of subjects smiling or reciting sentences were used. Our algorithms was applied for the first frame and refreshed periodically. Applying the technique on the VidTIMIT database and on our test sequences shows good tracking results for all individuals up to 30° of face orientation. Sudden movements of the head affected the tracking performances. In this situation, the fitting algorithm had to be applied more often in the video sequence. An example of the results obtained with our mouth tracker is presented in Figure 6.

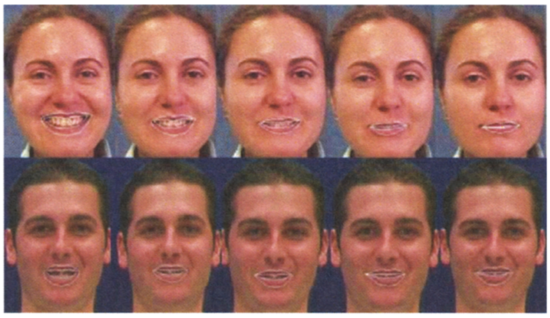

Figure 6: Example of mouth tracking sequences

Weak points could possibly be noticed when the face is strongly affected by lighting variations or the hue filter fails to find the correct initialization point. Situations like bearded people, very weak contrast between mouth and skin, illumination variation or lips colour not covered by our statistical analysis are examples in which the hue filter was suspected to fail. For these situations in which a general filter might fail, the filter parameters had to be recalculated for the specific constraints.

V CONCLUSIONS & FUTURE WORK

In this article a method for segmenting and tracking the mouth was proposed. Although still under development, the method appears to work fairly well on the available databases, as well as on video sequences taken especially for this study. The AAM, applied as a stand-alone method, performs poorly in mouth modelling, without an accurate starting point. The algorithm presented in Section III is capable of giving a robust initialization of the mouth area that facilitates the mouth alignment using AAM.

The results showed that our method works well for the images taken under controlled environment, for which the filter parameters were calculated using a statistical analysis. However, in more difficult situations, not covered by our statistical analysis, our hue filter can fail. An adaptive filter which is able to find the optimal parameters automatically is an important objective for our future work.

REFERENCES

[1] M. Pantic, L. J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, 2000, pp 1424-1445.

[2] V. Pahor, S. Carrato, "A Fuzzy Approach to Mouth Corner Detection", *Proceedings of the 1999 International Conference on Image Processing (ICIP '99),* Kobe, Japan, October 24-28, 1999, pp. 667-671.

[3] M. Pantic, M. Tomc, L.J.M. Rothkrantz, "A Hybrid approach to mouth features detection", in

Proceeding of the 2001 Systems, Man and Cybernetics Conference, 2001, pp. 1188-1193.

[4] S. Chindaro, F. Deravi, “Directional Properties of Colour Co-occurrence Features for Lip Location and Segmentation”, Proceedings of *the 3rd International Conference on Audio and Video-Based Biometric Person Authentication*, pp. 84 – 89, 2001.

[5] A. Wee-Chung Liew, S. H. Leung, and W. H. Lau, “Segmentation of Colour Lip Images by Spatial Fuzzy Clustering”, in *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 542 – 549, august 2003.

[6] S. Lucey, S. Sridharan, V. Chandran, “Adaptive mouth segmentation using chromatic features”, in *Pattern Recognition Letters*, vol. 23, pp. 1293–1302, 2002.

[7] M. Hennecke, V. Prasad, and D. Stork, “Using deformable templates to infer visual speech dynamics”, *287h Annual Asimolar Conference on Signals, Systems, and Computer*, vol. 2, IEEE Computer, Pacific Grove, pp. 576-582, 1994.

[8] A. W.C. Liew, S.H. Hung and W.H. Lau, “Lip contour extraction using a deformable model”, in *Proceedings of the International Conference on Image Processing*, vol. 2, pp. 255 – 258, 2000 .

[9] P. Gacon, P. Coulon, G. Bailly, “Non-linear Active Model for mouth inner and outer contours detection” , *In the 13th European Signal Processing Conference - EUSIPCO*, Antalya, Turquey, 2005.

[10] K. S. Jang,, “Lip Contour Extraction based on Active Shape Model and Snakes”, *IJCSNS International Journal of Computer Science and Network Security*, vol.7, no.10, October 2007.

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models”, *Lecture Notes in Computer Science*, vol. 1407, pp. 484–, 1998.

[12] P. Corcoran, I. Bacivarov, M. C. Ionita, “A Statistical Modeling based System for Blink Detection in Digital Cameras”, in Proc. *ICCE*, Las Vegas, 2008.

[13] P. Viola, M. J. Jones, “Robust Real-Time Face Detection”, *Springer Netherlands for Computer Science*, Volume 57, Number 2, May, 2004.

[14] Y. Gong and M. Sakauchi, "Detection of Regions Matching Specified Chromatic Features," *International Journal of CVGIP: Image Understanding,* Vol. 61, No. 2, 1995.

[15] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, “The FERET database and evaluation procedure for face recognition algorithms,” *Image and Vision Computing* J, Vol.16, No.5, pp. 295-306, 1998.

[16] P.J. Phillips, H. Moon, S.A. Rizvi, P. J. Rauss, “The FERET Evaluation Methodology for Face Recognition Algorithms,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 1090-1104, 2000.

[17] C. Sanderson and K.K. Paliwal, “Identity Verification Using Speech and Face Information”, *Digital Signal Processing* 14(5):449-480, 2004.

[18] M. B. Stegmann, B. K. Ersbøll, and R. Larsen, “FAME - a flexible appearance modelling environment,” IEEE *Transactions on Medical Imaging*, vol. 22, no. 10, pp. 1319–1331, 2003. [Online]. Available: http: // www2. imm.dtu.dk / pubdb/p.php?1918