# STATISTICAL CHROMATICITY-BASED LIP TRACKING WITH B-SPLINES

*M. Ulises Ramos Sánchez* *Jiři Matas* *Josef Kittler*

Department of Electronic and Electrical Engineering
University of Surrey, Guildford, Surrey GU2 5XH, United Kingdom
{M.Ramos-Sanchez,G.Matas,J.Kittler}@ee.surrey.ac.uk

## ABSTRACT

We present a statistical, colour-based technique for lip tracking intended to support personal verification. The lips are automatically localised in the original image by exploiting grey-level gradient projections as well as chromaticity models to find the mouth area in an automatically segmented region corresponding to the face area. A B-spline, initially with an elliptic shape is then generated to start up tracking. Tracking proceeds by estimating new lip contour positions according to a statistical chromaticity model for the lips. These measurements are used together with a Lagrangian formulation of contour dynamics to update the new spline control points. The method has been tested on the M2VTS database[1], where lips were accurately tracked on sequences of speaking subjects consisting of more than hundred frames. The tracker can be used to perform feature extraction from the mouth area as well as for model detection for personal verification applications.

## 1. INTRODUCTION

Tracking of lip contours has recently attracted the attention of the research community [2, 3, 4] because of the additional information conveyed to complement voice-based speech recognition techniques. Tracking of lip contours in side views against a constant background has been successfully reported [5], but frontal views without artifacts such as lipstick [6] have proved to be much more complicated and this precluded a fully automatic lip tracking in the past.

In the case of grey-level images, although some features can be quite consistent, such as the commonly referred lip intensity valley [7], standard derivative-based methods cannot be relied on and fail quite often to locate the lip boundaries in areas of particularly poor contrast, such as the lower lip. Moreover, pure intensity-based methods can be expected to be very sensitive to lighting variations and shading. Successful results have been reported [4] using grey-scale point distribution models [8], although it is unclear how to extract those models from the data in an automatic way.

To overcome the above difficulties a B-spline [9, 10, 11] tracker is presented using a statistical chromaticity model for the lips. Tests have been carried out on the M2VTS database [1]. B-splines are currently used for tracking purposes and are usually favoured with reference to other dynamic contour representations, such as *snakes* [12] because of their desirable properties like compactness, easy parameterisation and lower computational costs.

A coarse estimate of the mouth area is required for tracking the lips. In section 2. a method combining colour techniques as well as grey level image analysis is presented. The faces can be segmented automatically and the lips localised to obtain an initial fit of the lip boundary so as to start up the B-spline based [13, 14, 3] lip tracker. As far as the actual lip-tracking is concerned, a statistical colour-based algorithm has been developed for feature search. Details can be found in section 3..

## 2. LIP LOCALISATION

Lip tracking initialisation requires an approximate estimate of the mouth location. The approach developed to obtain such a coarse estimate is described in the following.

The original colour image undergoes a pixel classification step to find skin-like pixels in the image, thus allowing for the segmentation of the face area. A gradient projection analysis is then performed to identify lip-like candidates without any other prior assumption about the face pose or an approximate facial feature location. Lips are characterised by a strong projection value of the derivatives along the normal to their boundaries. This criterion, however, is not enough to determine the position of the lips since other features such as the eyebrows and the nostrils, for instance, exhibit the same property, apart from the presence of glasses in some images. As a result, the final process is aimed at evaluating the goodness of the several candidates found, and deciding which candidate is most likely to correspond to the mouth. This last step is analogous to the first one, since pixel classification is performed in the neighbouring area of the gradient projection maxima, this time using a statistical chromaticity model for the lips.

### 2.1. Pixel classification for skin segmentation

Chromaticity clustering [15] was used to segment the face area in a previously smoothed RGB image and the resulting skin cluster was assumed to correspond to a Gaussian chromaticity distribution. Such face 'skin' distribution obtained from a single speaker proved to generalise quite well for the others and thus was used to segment other speakers' faces by back-projecting the skin model on the chromaticity images and by establishing a threshold to the Mahalanobis distance between the test image pixels and the mean of the 'skin' model. A similar approach called FCC (for face colour classification) is described in [16].

To remove of *salt and pepper* noise due to the presence of spurious pixels in the image background, morphological opening is applied to the resulting image. The morphological operation removes the spurious pixels whilst still preserving the integrity of the facial area.

(a) Original image

(b) Pixel classification

**Figure 1. Face skin segmentation**

### 2.2. Generation of lip candidates based on grey-level gradient projection

Subsequent processing is restricted to the segmented face area, looking for lip-like regions inside it. The approach followed takes into account the saliency of the gradient projection along the normal direction to the lip boundaries. Similar approaches based on a vertical projection of horizontal edges are described, for instance, in [17, 18].

The original RGB image is converted into an intensity image and its gradient is computed using Sobel masks [19]. Gradient images are then obtained by replacing each pixel with the magnitude of its directional intensity derivative along the $\vec{l}$ direction, *i.e.*

$$D(i,j) = |\nabla I_{i,j} \cdot \vec{l}| \tag{1}$$

where $I_{k,l}$ is the brightness intensity of pixel $(k,l)$ and the $\vec{l}$ direction corresponds to that of the main axis of an ellipse fitting the spatial distribution of the pixels belonging to the face skin mask. The best fit ellipse is estimated by the moments method (see *e.g.* [20]).

The values of the magnitude of the computed directional derivative are projected onto the ellipse major axis:

$$P(x,y) = \sum_k D[(x,y) + k\vec{n}] \tag{2}$$

where $\vec{n}$ is the normal to the main ellipse axis, *i.e.*, the minor axis orientation.

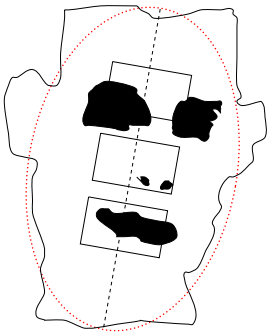
### 2.3. Selecting the best candidate

The lips will lie around a local maximum of the gradient projection occurring at a position on the main axis of the face. The mean gradient projection value is computed and all the local maxima above it are analysed to determine which one corresponds to the lip position.

Gradient projection maxima are relocated to correspond to positions along the main face axis. Around these centered positions, pixel classification is performed by comparing pixel chromaticity values by an appropriate chromaticity models for the lips. The strategy is represented in Figure 2. The rectangular boxes represent the area where pixel classification takes place. In accordance with a non-faceness assumption, pixel classification is restricted to areas that were not classified before as belonging to the face skin class, thus discouraging the presence of hits for skin pixels with similar chromaticity values to those of the lips.

As before, in the case of the face skin, a restrictive threshold was chosen for the maximum Mahalanobis distance ad-

(a) Original image



(b) Lip search strategy

(c) Lip candidate boxes

**Figure 2. Generation of lip candidates**

mitted to classify a pixel as 'lips'. If the threshold is exceeded, the lip hypothesis is rejected. Each subimage is scored according to the number of pixels classified as lips and the connectedness of the classification process. Rejects do not contribute to the final score. If a pixel is considered to belong to the 'lip' class, its neighbourhood is examined and its contribution to the score is given by the ratio of pixels recognised as lips over the size of the neighbourhood (8-connectivity is considered).

Lip chromaticity was found not to be adequately represented by a unimodal distribution and, consequently, individual chromaticity models were constructed for each of the speakers. The models were built by supervised chromaticity clustering (3 training images were used for each speaker) and tested on 80 independent images (5 for each of the 16 speakers), where the lips were always correctly localised. Typical examples of lip localisation for several speakers in the M2VTS database are shown in Figure 3.

## 3. LIP TRACKING

The lips are represented by closed quadratic B-splines [9, 11] consisting of $N$ spans and $N$ control points $\mathbf{q}_i$ $(i = 0, \ldots, N-1)$. B-splines allow for a compact representation of a contour in terms of a reduced number of **control points**. They impose certain shape selectivity that traditional *snakes* lack and their computational cost is also lower.

With reference to the dynamics model inertia, viscosity and elasticity properties have been assumed in a Lagrangian formulation [21] of the spline motion, which results in a sec-

**Figure 3. Lip localisation in the M2VTS database**

ond order differential equation governing the spline dynamics:

$$\ddot{\mathbf{Q}} + 2\beta_0\dot{\mathbf{Q}} + \omega_0^2\mathbf{Q} = \omega_0^2\mathbf{Q}_f \quad (3)$$

where $\mathbf{Q}$ is a vector containing the B-spline control points and $\mathbf{Q}_f$ is the least squares B-spline approximation to the extracted measurements of the lip contour.

The spline motion characteristics are determined by the choice made of the natural frequency $\omega_0$ and the damping coefficient $\beta_0$. In the absence of damping and for natural frequencies tending to infinity, we will be in the case where no previous shape information is retained and the spline is updated according only to the new measurement effected. Tracking will be faster but also more prone to instabilities resulting in undesirable spline shapes.

Tracking initialisation exploits chromaticity clustering in an automatically cropped image centered around the mouth location which has been generated by the lip localisation module. This clustering process allows for the segmentation of the 'skin' and 'lip' areas. The segmented lip region is then used to generate an elliptic B-spline approximation to the lip contour following the moments method (Figure 4).

Profiles along the normal to the contour [5] are used to search for the lip boundaries. Following a statistical approach, lip boundary search is considered to be a 2-class ('lip' or 'skin') classification problem along the extracted profiles.

The decision criterion follows a typical Bayesian approach [22], selecting that class (in the case of zero-one costs) for which the *a posteriori* probability is higher. If log-likelihood ratios are used instead, and each class is modelled as a bivariate Gaussian distribution in the chromaticity space, the criterion function can be expressed as follows:

$$J(\mathbf{x}) = J_{lips}(\mathbf{x}) - J_{skin}(\mathbf{x}), \quad J_i(\mathbf{x}) = -\log|\Sigma_i| - M(\mathbf{x}-\mu_i), \quad (4)$$

$M(\mathbf{x}-\mu_i)$ being the squared Mahalanobis distance of point $\mathbf{x}$ to the mean of class $i$. The decision rule would assign $\mathbf{x}$ to the 'lip' class if $J > 0$ and to the 'skin' class otherwise.

Once lip statistical models and a spline initialisation are available, lip boundary search can proceed by sampling chromaticity profiles along the spline contour and computing the decision criterion $J$.

## 4. EXPERIMENTAL RESULTS

The B-spline based tracker using the described statistical chromaticity model has been tested on the M2VTS database. In Figure 5 we can see how the B-spline follows

(a) Original image with region of interest

(b) Results of colour clustering

(c) 10-control point spline on original image

**Figure 4. Elliptic B-spline initialisation for lip tracking**

the variations in the mouth shape for one of the speakers in the database over a short run of 20 frames. Successful results are available as well for other speakers and longer sequences. The lip contour has been represented by a 10-control point B-spline. Tracking is implemented as an iterative process where the final state reached for a given frame is retained as the initial state for the next one.

## 5. APPLICATIONS

Reliable lip tracking allows for the extraction of mouth features to be used for personal authentication purposes. We can see the characteristics of those parameters as a function of time for several speakers and for a single speaker across the different shots recorded. In Figure 6(a) we have represented the extracted estimates of the lip width and height as a function of the frame number for one of the speakers. The 100-frame sequence corresponds almost entirely to the utterance of the digits 0 to 9 in French. Please note that no special care was taken to align the sequences at their starting points. Despite this fact and the different duration of the utterances, it is possible to notice a certain degree of correlation between all the instantiations.

On the other hand, as we can also see in Figure 6(b), where the same features are represented for different speakers, the degree of correlation is substantially lower. The above findings are consistent with the equivalent, well established principles exploited in acoustic-based speaker recognition and are expected to enhance their stand-alone success rates. More over, the lip shape model evolution curves can be used to detect lip shape states such as mouth open and mouth shut which can be used as control information for the selection of appropriate speaker models for a face based

(a) Frame 1 (b) Frame 4

(c) Frame 8 (d) Frame 12

(e) Frame 16 (f) Frame 20

**Figure 5. sp_01_v: tracking frames 1 to 20, in raster order**

personal identity verification technique.

## 6. CONCLUSIONS

A statistical chromaticity-based model for lip tracking has been presented. The method relies on a previous estimate of the lip position generated automatically by a lip localisation module.

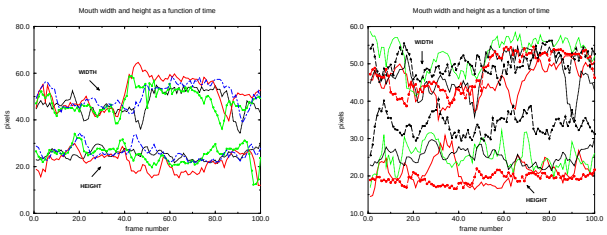
B-splines prove a suitable way to represent the lip contour in a compact way whilst still able to follow typical mouth shapes without additional geometrical constraints. Feature search implements a statistical chromaticity model to estimate the position of the lip contour.

The method has been tested on sequences of speakers of the M2VTS database, showing accurate and stable behaviour for sequences consisting of more than hundred frames. Finally, an application of lip tracking to feature extraction for image-based person authentication was presented.

## REFERENCES


[1] S. Pigeon. The M2VTS database. Technical report, UCL, Belgium, 1996 [http://www.tele.ucl.ac.be/M2VTS].

[2] C. Bregler and Y. Konig. 'Eigenlips' for robust speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2, pages 669–672. IEEE, Piscataway, NJ, USA, 1994.


(a) Intrapersonal variability (b) Interpersonal variability

**Figure 6. Mouth-based feature extraction**


[3] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Fourth European Conference on Computer Vision (Cambridge, UK, 1996)*, volume 2, pages 376–386. Cambridge, 1996.

[4] J. Luettin, N. A. Thacker, and W. Beet. Active shape models for visual speech feature extraction. Technical report, University of Sheffield, Sheffield, UK, 1995.

[5] A. Blake and M. Isard. 3d position, attitude and shape input using video tracking of hands and lips. In *SIGGRAPH '94 Conference Proceedings (Orlando, Florida, July 24–29, 1994)*, pages 185–192, 1515 Broadway, 17th floor, New York, NY 10036, USA, 1994.

[6] C. Montacié, M. J. Caraty, R. André-Obrecht, L. J. Boë, P. Deléglise, M. El-Beze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy, and H. Méloni. Applications multimodales pour interfaces et bornes evoluées (AMIBE). Technical report, Laboratoire des Formes et d'Intelligence Artificielle (LAFORIA), Université Pierre et Marie Curie, Paris, France, 1995.

[7] Y. Moses, D. Reynard, and A. Blake. Determining facial expressions in real time. In *IEEE International Conference on Computer Vision*, pages 296–301. IEEE, Piscataway, NJ, USA, 1995.

[8] T. F. Cootes and C. J. Taylor. Active shape models : A quantitative evaluation. In J. Illingworth, editor, *British Machine Vision Conference*, pages 639–648. BMVA Press, 1993.

[9] R. Bartels, J. Beatty, and B. Barsky. *An Introduction to Splines for Use in Computer Graphics and Geometry Modeling*. Morgan Kaufmann, 1987.

[10] J.D. Foley, A. Dam, S.K. Feiner, and J.F. Hughes. *Computer Graphics: Principle and Practice*. Addison-Wesley, 1990.

[11] G. Farin. *Curves and Surfaces for Computer Aided Geometric Design*. Academic Press Ltd., 1993.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *IEEE International Conference on Computer Vision*, volume 1, pages 259–268. IEEE, 1987.

[13] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *IJCV93*, 11(2):127–145, 1993.

[14] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. In *Artificial Intelligence*, page in press, 1995.

[15] J. Matas. *Colour-based Object Recognition*. PhD thesis, University of Surrey, 1995.

[16] P. Duchnowski, M. Hunke, D. Buesching, U. Meier, and A. Waibel. Toward movement-invariant automatic lip-reading and speech recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 1, pages 109–112. IEEE, Piscataway, NJ, USA, 1995.

[17] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct 1993.

[18] G. Galicia and A. Zakhor. Depth based recovery of human facial features from video sequences. In *IEEE International Conference on Image Processing*, volume 2, pages 603–606, Washington D.C., USA, October 23-26 1995. IEEE Computer Society Press, Los Alamitos, California.

[19] R. C. Gonzalez and R. E. Woods. *Digital image processing*. Addison-Wesley, 1992.

[20] K. Sobottka and I. Pitas. Localization of facial regions and features in color images. *Journal of Pattern Recognition and Image Analysis*, 1996.

[21] R. Curwen and A. Blake. *Dynamic Contours: Real-time Active Splines*, chapter 3, pages 39–57. MIT Press, Cambridge, Massachusetts, 1992.

[22] K. Fukunaga. *Introduction to Statististical Pattern Recognition*. Academic Press, 1990.