NON-LINEAR ACTIVE MODEL FOR MOUTH INNER AND OUTER CONTOURS DETECTION

Pierre Gacon[(1)], Pierre-Yves Coulon[(1)], Gérard Bailly[(2)]

LIS-INPG[(1)], ICP-INPG[(2)]
46, Avenue Félix Viallet
38031, Grenoble, France
{gacon, coulon}@lis.inpg.fr, bailly@icp.inpg.fr

ABSTRACT

Mouth segmentation is an important issue which applies in many multimedia applications as speech reading, face synthesis, recognition or audiovisual communication. Our goal is to have a robust and efficient detection of lips contour. In this paper, we focus on the detection of the inner mouth contour which is a difficult task due to the non-linear appearance variations. We propose a method based on a statistical model of shape with local appearance gaussian descriptors. Our hypothesis is that the response of the local descriptors can be predicted from the shape. This prediction is achieved by a non-linear neural network. We tested this hypothesis with a single speaker task and compared the results with previous methods. Then this approach is generalized to take care of the intra person appearance variability in a multi-speaker task.

1. INTRODUCTION

Lips segmentation can apply to various research areas such as automatic speech recognition (in human-computer interface), speaker recognition, face authentification, or to improve speech intelligibility in noisy situation for audio-video communication. Extracting the shape of lips and modeling it with a few number of parameters can allow low-bandwidth communication or to animate a clone or an avatar of a person.

Various methods have been developed to achieve lips segmentation in the last few years. They are mainly of two types: without or with a lips model.

In the first case, only information as colour or edge are used. For example, Delmas [1] proposed to use snakes and an gradient criterion to detect lips. This type of method can give convincing results if the condition of lighting and the contrast between colour of lips and skin are good. But in other cases, the segmentation might become difficult and give non-realistic results.

To have more realistic results, it is very useful to have a model for the shape of the lips.

Hennecke et al. [2] use a deformable template. The template is a model of the lips controlled by a set of parameters which are chosen by minimizing a criterion based on the edges of the lips. For this kind of approach, the lack of flexibility of the template can be a problem.

Eveno [3] proposed to use parametric curves to describe the lips and fit them to the image using gradient information based on hue and luminance. These curves are very flexible, but can still generate impossible shapes.

Cootes et al. [4] introduced active shape models. The shape of an object is learned from a training set of annotated images. After a principal component analysis (PCA) a limited number of parameters drives the model. The main interest is that the segmentation will always give a realistic result (given that the distribution of the data is effectively Gaussian). Values of the parameters are selected with an appropriate criterion. Cootes et al. [4] introduced also active appearance models in which shape and grey-level appearance are also learned. Luettin [5] also developed an active shape model method in which he learned a grey-level profile model around lips contour in the training set. This profile model provides a measure of the goodness of fit between the model and the image.

In our prior work [6] [7], we presented a method based on an active shape and sampled-appearance model. The cost function used to fit the model was based on the difference between the modelized appearance and the actual appearance of the processed image and on the computation of the flow of a gradient operator through the curves of the shape.

In this paper we try to replace this cost function by a new criteria based on the response of gaussian local descriptors. Knowing the shape, we predict the response of the descriptors with a non-linear neural network and we compare it to the response observed on the image. This is particularly adapted for the inner lip contour and mouth bottom which present a high variability and non-linearities (mouth close or open, teeth or tongue presence). The response of these filters depends on the movement induced by speech (intra person variability), on the speaker identity (inter person variability) and on the lightening conditions. We will demonstrate that this criteria is efficient in a single speaker task, and next generalize it to a multi speaker task by taking care of the inter person variability.

2. DATA SET ANNOTATION

In this paper, we will consider that the face is detected in a preprocessing step (figure 8 shows some typical processed image). As color and brightness are mixed in the RGB color space we chose to work with the YCbCr space where chromatic and luminance components are separated.

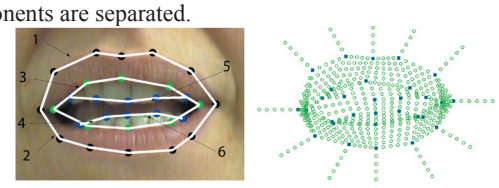

Figure 1 : Example of annotated image and mesh used for the sampling of appearance

The data-set consists in video sequences of 12 speakers. N=450 images were manually annotated to build the model (the others were used to test the algorithm). The general shape is described by 30 control-points (12 for the outer lip contour, 8 for the inner lip contour and 10 for teeth) as shown in figure 1, and the coordinates were saved in k=60 values vectors $\mathbf{s}_i$ ($1 \le i \le N$).

If the mouth is closed or if the teeth don't appear, the

corresponding points are merged with those of the inner lip contour. We also assigned a General Mouth State (GMS) to each image. The GMS is a state variable, it describes elementarily the different typical mouth position: closed, open, wide open, smiling.

3. SINGLE SPEAKER TASK AND LOCAL DESCRIPTORS

3.1 Active shape model

In this part, we only work with a single speaker, so we don't use the whole data-set but M images (M<N) of the same speaker. In order to reduce the dimensionality of the problem, we proceed to a PCA as in [4] with the s_i. The mean vector $\bar{\mathbf{s}}$ and the covariance matrix $\mathbf{S}$ and its eigenvectors $\mathbf{p}_k$ and eigenvalues λ_k with $1 \le k \le 60$ are then computed as follow:

$$\bar{\boldsymbol{s}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{s}_i \;;\; \mathbf{S} = \frac{1}{M}\sum_{i=1}^{M} \left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)\left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)^T$$

The eigenvectors of the covariance matrices correspond to the various variation modes of the data. As the eigenvectors with large eigenvalues describe the most significant part of the variance, the selection of a few modes can reduce the dimensionality of the problem. We keep 95% of the variance and the selected eigenvectors are saved in matrix $\mathbf{P}_s$. So shape will be described by a few parameter.

Finally we can generate any shape of the training set or new plausible examples by simply adjusting vector parameter $\boldsymbol{\sigma}$ with the following equation:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P}_s\boldsymbol{\sigma}$$

Segmenting mouth on an image will then consist in finding the best set of parameters that control our active mouth model, i.e the best projection in the new low-dimensional space.

We also calculated the means of $\boldsymbol{\sigma}$ parameters vectors for each 4 GMS and we saved them in vectors $\boldsymbol{\sigma}_{gms,j}$ $(1 \le j \le 4)$.

3.2 Gaussian local descriptors

In order to have a cost function to find the best set of parameters, we chose to use the first gaussian derivative filters [8] as local descriptors. These filters are convolution windows (their sizes is one tenth of the mouth width). We limit the model to the first Gaussian derivatives so we will compute the convolutions between the 3 filters G and Gx and Gy (mean and horizontal and vertical gradients, ie figure 2) and the image.

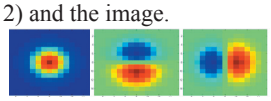

Figure 2 : Gaussian derivative filters G and Gx and Gy

As we want to predict the response of the gaussian filters from the shape, this response has to only depend of the shape. As we work on a single speaker task, there is no inter person variability. Nevertheless lightening change can modify the response of the filter for the same person and the same shape. So we use the retina filter [9] on the luminance to diminish this variability. This filter is a band pass spatial filter which enhance contours and reduce illumination variation. Figure 3 illustrates the effect of this filter. On a sequence of image of the same speaker with important illumination change from one frame to another, luminance varies a lot while the filtered luminance remains almost constant.

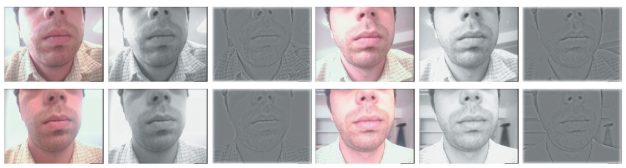

Figure 3 : Retina filter illustration (image, luminance and filtered luminance)

Knowing the shape for each image, we made a piecewise cubic interpolation and we resample the curve in order to have the filters homogeneously parceled out on the contour without recovering each other. Then we compute for each image and for Yf (filtered luminance) and CbCr components the response of the three filters in each points of the inner and outer lip contour. Figure 4 shows an example of responses for the filters on the outer contour for Yf. The representation is polar: the abscissas are the angle with the center taken as the middle of the segment linking the mouth corners.

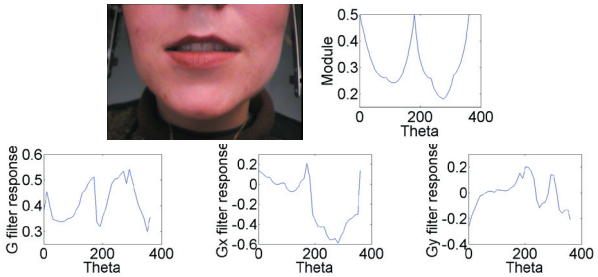

Figure 4 : Computed image, outer lip contour in polar representation, G and Gx and Gy filter responses for Yf

3.3 Descriptors response prediction and cost function

We now want to predict the 9 responses curves (3 filters x 3 components) from the shape curve with a neural network. To achieve this task, the neural network has to be able to deal with non-linear problems (as the response of filters on the inner contour vary non-linearly when the mouth is opening, or when teeth or tongue appears) so we chose to use feed forward backpropagation.

To diminish the size of the neural network, we use the active shape parameters as entry for the network and we do a PCA on the descriptors to keep 80% of the variance.

For a tested set of parameters, our cost function C_f is defined as the mean square error between the response predicted from the parameters values and the actual response of the filters computed on the processed image.

3.4 Mouth corners local model and detection

Mouth corners points are used as key-points to determine the position and scale of the mouth. But these points are quite difficult to detect as they are often not on an edge but in shadowy region as shown in figure 5. So, mouth corners are considered as the intersection point between 4 regions. Region 1 and 2 are non-homogeneous as they are characterized by an edge between lips and skin while on the other hand, region 3 and 4 are homogeneous on a chromatic point of view (figure 5). Mouth corners will then be described by a set of 4 Gaussian derivative filters and the statistical distribution of the responses of the filters are described by Gaussian mixture models.

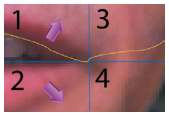

Figure 5 : Mouth corners area with characteristics regions, edge direction and line of luminance minima

As Eveno [3] proposed, the mouth corners are supposed to be on the line which links luminance minima for each column (see figure 5). So, we only have to find the index of the columns to know these mouth corners points. To do so we compute the local descriptors for each pixel of the line of interest and the most probable couple of point is selected as mouths corners. More details about this detection can be found in [7].

3.5 Lip segmentation

We want to find the set of parameters to obtain the best segmentation of mouth with the model. $C_f(I_n)(\boldsymbol{\sigma})$ is the value of our

cost function for the processed image I_n and for the PCA parameters vectors $\boldsymbol{\sigma}$.

To minimize this cost function, we have to solve a high dimension problems. To achieve this, we use a Downhill Simplex Method (DSM), which is a minimization/maximization classical method. To run the DSM, we have to define an initial guess and a search interval for the parameters which are classically the mean parameters values three time the standard deviation of the modes of our active model ($3\sqrt{\lambda_k}$). Figure 6 presents the method principle.

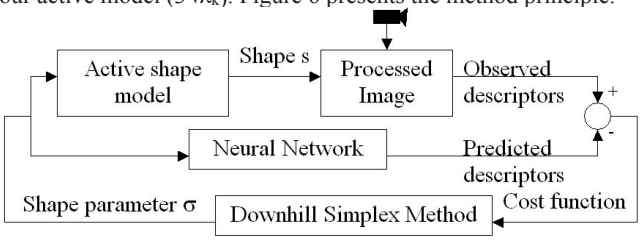

Figure 6: Single speaker method of segmentation

First image I_1

To optimize the choice of the initial guess, we test the general mouth state (GMS) by computing the $Cf(I_1)(\boldsymbol{\sigma}_{gms,j})$ for $1 \le j \le 4$, the minimum giving the initial guess ($\boldsymbol{\sigma}_{gms,j\_min}$). Then we proceed to the minimization of $Cf(I_1)(\boldsymbol{\sigma})$ by DSM (it stops when the difference between the maximum and the minimum of the simplex is lower than a threshold value) and we find a final set of parameters $\boldsymbol{\sigma}_1$.

Tracking

For image I_{n+1}, we check if:

$$\left| \frac{C_f(I_{n+1})(\boldsymbol{\sigma}_n) - C_f(I_n)(\boldsymbol{\sigma}_n)}{C_f(I_n)(\boldsymbol{\sigma}_n)} \right| \le 20\%$$

If this is verified, we assume that the mouth in the new image has practically the same shape that on the previous image. We will then minimize $Cf(I_{n+1})(\boldsymbol{\sigma})$ by DSM, with $\boldsymbol{\sigma}_n$ as initial guess and reduced search intervals $0{,}5\sqrt{\lambda_k}$ for parameters. If the condition is not verified we test again the GMS to have a new initial guess and the search intervals are $3\sqrt{\lambda_k}$.

3.6 Discussion

Here we give some results with a training set of M=50 images of one speaker. We compare it with two other methods. The first one is a previous single speaker method we presented in [6] in which both shape and appearance were modelized by a PCA with a pixel based cost function computed as the difference between modelized appearance and the observed appearance in the image. The second one is based on the computation of gradient fields. The control-points which describe the lips and the teeth can be divided in 6 curves (see figure 1). If the flow of a gradient vector through these curves is maximized, then the curves will fit with the edges of the image. Various gradient fields are used according to the curves: an hybrid edge **Grl** (based on an idea introduced by Eveno [3]) wich combined chromatic and luminance data, is used for curve 1, **Gr** (based on Cr component) is used for curves 2, 3 and 4 and **Gl** (based on luminance) is used for curves 5 and 6.

method	outer lip contour	inner lip contour	teeth	all points	number of iterations
a	1.4/0.8	1.8/0.9	1.8/0.9	1.5/0.8	9.8
b	1.4/0.9	2/1.1	2.2/1.2	1.8/1	31.1
c	2.3/1.3	6/3.5	6.2/3.5	4.4/2.5	19.2

Table 1 : Single speaker method comparison

with a : proposed method with non linear descriptors, b : active appearance and shape model ([6]), c : active shape model with cost function based on gradient flow maximization. The errors are given in percentage of the scale of the mouth : mean error/ standard deviation

If we compare in table 1 our proposed method (a) with the appearance model based method (b), we see our new cost function has the best results and converges to the final result faster. As the prediction is non-linear, it seems to deal the inner mouth area better than a method only based on linear PCA modelization of the inner mouth area, while outer mouth contour segmentation is equivalent. And as only shape is modelized, there is fewer parameters to adjust and the convergence is faster.

If we compare with the gradient flow based method (c), we see this approach fails to give robust results for the inner mouth area (and the segmentation of the outer contour is less accurate too, probably because of the bad segmentation of the inner one). When mouth is closed, the inner mouth contour is pretty hard to detect by a gradient flow maximization as it separates regions with similar YCbCr characteristics. Moreover as teeth can appear when mouth is opening, the direction of the gradient vector can change and the teeth/mouth bottom frontiers can be confounded with the lips/teeth frontiers. In comparison our descriptors prediction behaves like a "clever" gradient flow maximization as it adapts the responses of the gradient information (filters Gx and Gy) to the configuration of the mouth and the mean filter G discriminates the various frontiers.

In conclusion of this part, our cost function proved to be adapted to a single speaker task and gives accurate and robust results.

4. MULTI SPEAKER GENERALIZATION

4.1 Multi speaker model

In this part we want to adapt this cost function based on local descriptors to a multi speaker task. The basis is a shape and sampled-appearance model we presented in our previous paper [7].

This method is based on an active model approach in three steps. First, a pre-processing step allows to find a good initial guess for the model by roughly classifying the pixels in skin or lip categories with a pixel color model based on a gaussian mixture model. Next, we find mouth corners with the local appearance model (as in 3.4) which determines the position and the scale of the mouth. Finally, we optimize the parameters of a shape and sampled-appearance active model for the whole mouth area.

In this model, a distinction is made between static appearance and dynamic appearance. Static appearance is the mean appearance for each speaker and dynamic appearance corresponds to the appearance variation induced by speech and movement. As the shape is already saved is the vectors $\mathbf{s}_i$, the sampled-appearance is learned for every image by extracting the three YCbCr components at 728 features-points given by a mesh computed from the $\mathbf{s}_i$ (as shown in figure 1), which defines precisely if an appearance sample corresponds to skin, lips, teeth or inner mouth. Dynamic appearance is saved in vectors $\mathbf{a}_{d,i}$ and static appearance (or mean speaker appearance) is saved in vectors $\mathbf{a}_{m,i}$ ($1 \le i \le N$).

We then proceed to PCAs similar to 3.1 with the $\mathbf{s}_i$, $\mathbf{a}_{d,i}$ and $\mathbf{a}_{m,i}$ in order to keep 95% of the variance. Then a second step PCA is made to link shape and dynamic appearance variability (as in [4]). Finally this model is driven by this set of equations:

$$(1)\ \mathbf{c} = \begin{bmatrix} W.\boldsymbol{\sigma} \\ \boldsymbol{\alpha}_d \end{bmatrix} = \bar{\mathbf{c}} + \mathbf{P}_c \boldsymbol{\chi} \Rightarrow \begin{cases} (2)\ \mathbf{s} = \bar{\mathbf{s}} + \mathbf{P}_s \boldsymbol{\sigma} \\ (3)\ \mathbf{a}_d = \overline{\mathbf{a}_d} + \mathbf{P}_d \boldsymbol{\alpha}_d \end{cases}$$

$$(4)\ \mathbf{a}_m = \overline{\mathbf{a}_m} + \mathbf{P}_m \boldsymbol{\alpha}_m$$

Equation (2) controls the shape model, equation (3) controls dynamic appearance and equation (4) controls static appearance. Equation (1) controls the combined model for shape and dynamic appearance (with balancing coefficient W). Segmenting mouth on an image will then consist in finding the best set of 18 parameters: 9 in combined parameter vector $\boldsymbol{\chi}$ and 9 in static appearance vector parameters in $\boldsymbol{\alpha}_m$. More details about this can be found in [7].

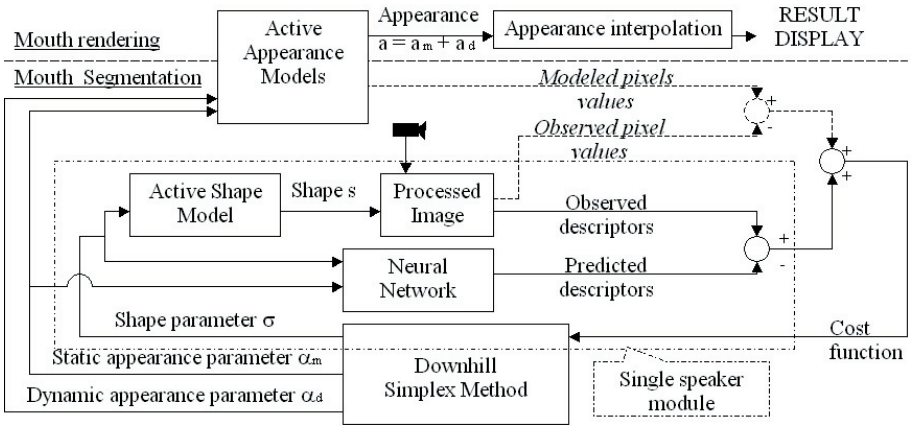

Figure 7:
Multi speaker method of segmentation
The single speaker modul corresponds to 3.5 and figure 6.
Modeled and observed pixels values (in italic on the dot-line links) are only computed while α_m has not converged on a sequence of image.
When α_m has converged, the cost function is only computed as the error of prediction of the descriptors by the neural network.
The appearance model is then only used to interpolate sampled appearance for the mouth rendering.

4.3 Gaussian derivatives filters and cost function

In [7], the cost function used to optimize the model was based on the difference between the modelized appearance and the observed appearance of the processed image and on the computation of the flow of a gradient operator through the curves of the shape. Now, we will use a generalized version of the cost function presented in 3.3. But in 3.3 the neural network was used to predict the response of gaussian derivative filters from the shape in a single speaker task. Here the speaker can vary so the response of the filters do not depend only from shape: from one speaker to another color of lips and skins and the contrast between them will fluctuate. But these characteristics are content in the static appearance.

So if not only the shape parameters but also the static appearance parameters are given in entry to a neural network it could be able to predict the response of the gaussian derivative filters in a multi speaker task. Our neural network is then trained on the data set and has finally 18 entries (parameters χ and $\boldsymbol{\alpha}_m$) 15 hidden units and 15 outputs (a PCA is made on the descriptors of the training lips in order to keep 80% of the variance). So while parameter vector $\boldsymbol{\alpha}_m$ will be converging on a sequence of images of a speaker, the cost function will be computed as the sum of the mean square error between the filters responses and its prediction and the difference between modelized and observed sampled-appearance (compared to the cost function in [7], the descriptors replace the gradient flows optimization). When $\boldsymbol{\alpha}_m$ (static appearance) has converged after a few images (typically 5 or 6), the cost function will only be computed with the filters responses. Then, the sampled appearance will only be used to generate a realistic and understandable avatar of the speaker by interpolation.

4.4 Lip segmentation

The principle is the same as in 3.5 and is summarized by figure 7: we optimize $Cf(I_n)(\chi,\boldsymbol{\alpha}_m)$ with the DSM. χ simply replaces σ but we also optimize parameter vector $\boldsymbol{\alpha}_m$ (this vector is initialized on the first image with the pixel color model, more details in [7]).

4.5 Results and conclusion

error position	outer lip contour	inner lip contour	teeth	all points
test on training set	2.7/1.3	2.8/1.4	3/ 1.4	2.8/ 1.3
leave-one-out	3/1.4	3.1/1.4	3.3/ 1.5	3.1/ 1.4

Table 2 : Mean error localization.
The errors are given in percentage of the scale of the mouth : mean error/ standard deviation

Table 2 gives result of segmentation in a multi-speaker task (tested on the training test and with a leave-one-out protocol: the tested speaker is out of the training set) and figure 8 shows some example. In 3.6, we demonstrate that our cost function was relevant for the task of segmenting mouth area and that it deals particularly well with the non-liearities of the inner mouth area. This method can be generalized to a multi speaker task and it then gives accurate and robust contour detection. Finally, the sampled appearance can be interpolated to obtain a realistic avatar of the speaker (figure 9 shows example of mouth rendering).

Figure 8 : Examples of mouths segmentation

Figure 9 : Examples of mouths rendering

5. REFERENCES

[1] P. Delmas, N.Eveno, and M. Lievin, "Towards Robust Lip Tracking", *International Conference on Pattern Recognition (ICPR'02),*Québec City, Canada, August 2002

[2] M. Hennecke, V. Prasad, and D. Stork. "Using deformable templates to infer visual speech dynamics", *28 h Annual Asimolar Conference on Signals, Systems, and Computer*, *volume 2*, IEEE Computer, Pacific Grove, pages 576-582, 1994.

[3] N.Eveno, A. Caplier, and P-Y Coulon, "Automatic and Accurate Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 706-715, May 2004

[4] T. F. Cootes. "Statistical models of appearance for computer vision", Online technical report available from http://www.isbe.man.ac.uk/bim/refs.html, 2001.

[5] J. Luettin, N.A. Thacker, S.W. Beet, "Locating and Tracking Facial Speech Features", *Proceedings of the International Conference on Pattern Recognition*, Vienna, Austria, 1996.

[6] P. Gacon, P.-Y. Coulon, G. Bailly. "Shape and Sampled-Appearance model for Mouth Components Segmentation", *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'04)*, Lisbon, Portugal, 2004.

[7] P. Gacon, P.-Y. Coulon, G. Bailly. "Statistical Active Model for Mouth Components Segmentation", *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, Philadelphia, USA, 2005.

[8] T. Lindeberg, "Feature detection with automatic scale detection", *IJVC*, vol. 30, no.2, pp. 77-116, 1998.

[9] Beaudot W.H.A., "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", PhD Thesis in Computer Science, INPG (France) december 1994