SHAPE AND SAMPLED-APPEARANCE MODEL FOR MOUTH COMPONENTS SEGMENTATION

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

LIS-INPG[(1)], ICP-INPG[(2)]
46, Avenue Félix Viallet
383031, 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 audio-video communication. In this paper, we propose a method based on a statistical model of shape and appearance to detect the lips and to create a clone of the mouth region. To create the model, the outline of the lips and teeth has to be manually annotated with 30 key-points on a few visems (400). After a step to situate mouth corners, the goal is to find the parameters to fit the model to an unknown image. The originality of this work is the automatically extracted sampled-appearance which is well adapted to describe the mouth and particularly its interior which will be also described by a state variable.

1. INTRODUCTION

Lips segmentation can apply to various research areas such as automatic speech recognition (in human-computer interface), speaker recognition, 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-bit 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: with or without a model for the lip.

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 technique 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] suggested to 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] 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 allows to give a measure of the fit between the model and the image.

We chose an active model approach in two steps for our work. First, we find mouth corners with an appearance model, which determines the position and the scale of the mouth. Secondly, we optimize the parameters of a shape and sampled-appearance model for mouth. 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), which can be dealt by a statistical model. We place a special emphasis on the teeth segmentation by using a descriptor for their presence/absence. The sampled-appearance gives a good clue for the relevance of the segmentation and allows to generate a clone which can be easily understandable by a human lip-reader.

In this paper, we present our first results which are acquired for a single speaker and with controlled lighting conditions. The images have a size of 156×214 pixels.

2. MODELS AND CRITERIA

2.1 Building of our active mouth model

The data-set is the series of images “hélène”. It consists in long video sequences (nearly 2000 images in total) of the same speaker saying phone numbers. N=400

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 figure1, and the coordinates were saved in 60 values vectors $\mathbf{s}_i$ ($1 \le i \le N$).

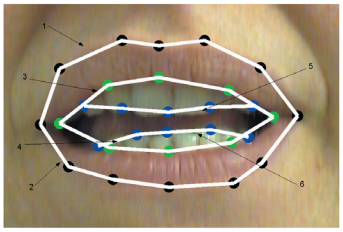

Figure 1: Example of annotated image

If the mouth is closed or if the teeth don't appear, the corresponding points will be merged with those of the inner lip contour.

Figure 2: Typical images for each IMS

While the control-points are entered, the operator also assigned an inner mouth state (IMS) for each image. The number of IMS has been fixed to 6. They describe elementarily the inner mouth area: 1) mouth closed, 2) mouth open with no teeth visible, 3) upper teeth visible, 4) upper teeth visible with inner mouth, 5) upper and lower teeth visible, 6) upper and lower teeth visible with inner mouth. Figure 2 shows one example of image assigned to each IMS.

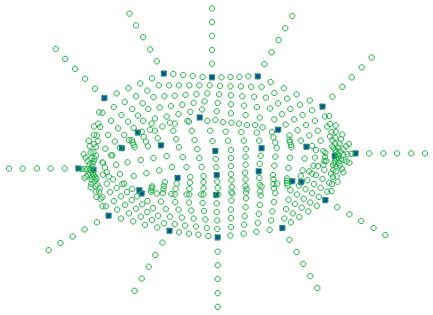

Figure 3: Mesh used for the sampling of appearance. Plain circles: control-points, empty circles: mesh

The next step is to learn the sampled-appearance for every image. The three RGB components are extracted at 728 features-points given by a mesh computed from the $\mathbf{s}_i$ (as shown in figure 3). Feature-points coordinates are saved in 1456 (728x2) values vector $\boldsymbol{\alpha}_{s,i}$ and RGB values in 2184 (728x3) values vector $\boldsymbol{\alpha}_{a,i}$ ($1 \le i \le N$). Using a mesh defines precisely if an appearance sample corresponds to skin, lips, teeth or inner mouth.

We then proceed to a two-steps PCA as in [4] with the $\boldsymbol{\alpha}_{s,i}$ (shape data) and $\boldsymbol{\alpha}_{a,i}$ (appearance data). In the first step, the mean vectors ($\overline{\boldsymbol{\alpha}}_s$, $\overline{\boldsymbol{\alpha}}_a$) and the covariance matrices ($\mathbf{S}_s$, $\mathbf{S}_a$) and their respective eigenvectors ($\mathbf{p}_{s,m}$, $\mathbf{p}_{a,n}$) and eigenvalues ($\lambda_{s,m}$, $\lambda_{a,n}$) with $1 \le m \le 1456$ and $1 \le n \le 2184$) are then calculated. For example for shape:

$$\overline{\boldsymbol{\alpha}}_s = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_{s,i}, \quad \mathbf{S}_s = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{\alpha}_{s,i} - \overline{\boldsymbol{\alpha}}_s\right)\left(\boldsymbol{\alpha}_{s,i} - \overline{\boldsymbol{\alpha}}_s\right)^T$$

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 chose to keep 95% of the variance for shape and 90% for appearance. The selected eigenvectors are saved in matrices $\mathbf{P}_s$ and $\mathbf{P}_a$. We then compute for each image the values of weight-vectors $\mathbf{b}_{s,i}$ and $\mathbf{b}_{a,i}$ ($1 \le i \le N$):

$$\mathbf{b}_{s,i} = \mathbf{P}_s^T\left(\boldsymbol{\alpha}_{s,i} - \overline{\boldsymbol{\alpha}}_s\right), \quad \mathbf{b}_{a,i} = \mathbf{P}_a^T\left(\boldsymbol{\alpha}_{a,i} - \overline{\boldsymbol{\alpha}}_a\right)$$

In the second step, we do a PCA with the values of the weight parameters to have a statistical model which links shape and sampled-appearance in order to have a coherent modelization between the two informations. We need to normalizes the units difference with the matrix N.

$\overline{\boldsymbol{\alpha}}_c$ is the mean vector, $\mathbf{S}_c$ is the covariance matrix and $\mathbf{P}_c$ is a matrix containing the eigenvectors to keep 95% of the variance (8 modes with eigenvalues λ_k, $1 \le k \le 8$). The first 2 modes of variation are shown in figure 4.

$$\overline{\boldsymbol{\alpha}}_c = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\mathbf{N}.\mathbf{b}_{s,i}\\ \mathbf{b}_{a,i}\end{pmatrix}, \quad \mathbf{S}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\begin{pmatrix}\mathbf{N}.\mathbf{b}_{s,i}\\ \mathbf{b}_{a,i}\end{pmatrix} - \overline{\boldsymbol{\alpha}}_c\right)\left(\begin{pmatrix}\mathbf{N}.\mathbf{b}_{s,i}\\ \mathbf{b}_{a,i}\end{pmatrix} - \overline{\boldsymbol{\alpha}}_c\right)^T$$

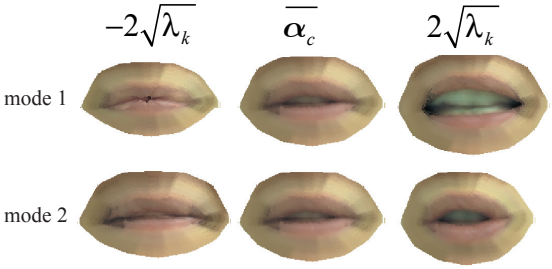

Figure 4: First 2 modes of variation (images are interpolated from sampled-appearance)

Then we can generate any shape and sampled-appearance of the training set or new plausible examples by simply adjusting $\mathbf{b}_c$ with the following equations:

$$\begin{Bmatrix}\mathbf{N}.\mathbf{b}_s\\ \mathbf{b}_a\end{Bmatrix} = \overline{\boldsymbol{\alpha}}_c + \mathbf{P}_c\mathbf{b}_c \Rightarrow \begin{cases}\boldsymbol{\alpha}_s = \overline{\boldsymbol{\alpha}}_s + \mathbf{P}_s\mathbf{b}_s\\ \boldsymbol{\alpha}_a = \overline{\boldsymbol{\alpha}}_a + \mathbf{P}_a\mathbf{b}_a\end{cases}$$

To end the training step, we have to learn the statistical properties of the weight parameters according to IMS. The values of vector $\mathbf{b}_c$ are computed for every image:

$$\mathbf{b}_{c,i} = \mathbf{P}_c^T\left(\begin{pmatrix}\mathbf{N}.\mathbf{b}_{s,i}\\ \mathbf{b}_{a,i}\end{pmatrix} - \overline{\boldsymbol{\alpha}}_c\right)$$

Then the means of each weight parameter are calculated for the 6 IMS and saved in vector $\mathbf{b}_{ims,j}$ (with $1 \le j \le 6$). The limits of variation (minima and maxima) for weight parameters are also learned for each IMS.

2.2 Appearance model for mouth corners

A second appearance model is built specifically for the mouth corners. It will be used to locate the mouth and give an initial position for the lip model. The values of luminance in 5×5 pixels squares centered on mouth corner points (previously manually annotated) are placed in vectors $\boldsymbol{\beta}_i$ for each image and the model is obtained with the same method as in 2.1. If $\boldsymbol{\beta}$ contains the modelled appearance and $\mathbf{Q}_a$ the selected modes and if $\mathbf{d}_a$ is the weight vector the equation of the model is:

$$\boldsymbol{\beta} = \mathbf{Q}_a\mathbf{d}_a + \overline{\boldsymbol{\beta}}$$

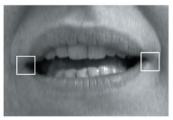

Figure 5: Appearance square for mouth corner

2.3 Criteria used for the method

2.3.1 Mouth shape-criterion

The 30 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. If **G** is a gradient field and ζ is a curve, the flow of vector **G** through ζ is calculated as:

$$\phi = \frac{\int_{\zeta} \mathbf{G}.\mathbf{dn}}{\int_{\zeta} \mathbf{ds}}$$

Various gradient fields are used according to the curves.

Poggio [6] introduced the pseudo-hue **H** for which there is a strong contrast between lips and skin (and so a strong gradient for the edge). It is computed at point (x,y) as:

$$H(x,y) = \frac{R(x,y)}{R(x,y)+V(x,y)}$$

where R and V are the RGB component.

Eveno [3] introduced the hybrid edge **Ghl,** a gradient field which combined normalized pseudo-hue $\mathbf{H_n}$ with normalized luminance $\mathbf{L_n}$ to enhance top frontier of the upper lip contour:

$$\mathbf{Ghl}(x,y) = \nabla\left[Hn(x,y) - Ln(x,y)\right]$$

The other gradient fields used are **Gh** and **Gl** respectively based on pseudo-hue and luminance.

For curve 1, **Ghl** will be used to compute the flow. **Gh** is used for curves 2, 3 and 4 and **Gl** for curves 5 and 6.

The gradient-based criterion C_g is computed as the sum of these 6 flows. It will be our shape criterion.

2.3.2. Mouth appearance-criterion

This criterion C_V simply compares the RGB values of the sampled appearance given by the model for a set of weight parameters $\mathbf{b}_c$ to the RGB values withheld in the current processed image by computating the mean square error.

We also define a global cost function to be minimized $C_C=C_V/C_g$ which combines shape and appearance criteria.

2.3.3. Mouth corners appearance criterion:

This criterion C_a is used to measure the fitting of the appearance squares for mouth corners. It was introduced by Luettin [4] to measure the fitting of his grey-level profiles to the image. It is computed as:

$$C_a(\boldsymbol{\beta}) = \left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right) - \left(\mathbf{Q}_a{}^T\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)\right)^T\left(\mathbf{Q}_a{}^T\left(\boldsymbol{\beta}-\overline{\boldsymbol{\beta}}\right)\right)$$

3. METHOD OF SEGMENTATION

3.1 Optimization method

To minimize a cost function, we have to solve 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.

3.2 Localization of mouth corners

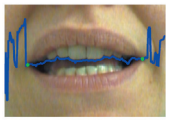

Figure 6: Luminance minima and mouth corners

We suppose that the zone of interest of the mouth is known after a pre-processing. As Eveno [3] proposed, the mouth corners are supposed to be on the line which links luminance minima for each column (see figure 6).

So, we only have to find the columns x_g and x_d to know these mouth corner points. This task is achieved by minimizing C_a by DSM. For the first processed image I, initial guess is random. For image I_{n+1} the initial guess for x_g and x_d will be the final values found for I_n.

3.3 Lip segmentation

We want to find the set of parameters for our model to obtain the best segmentation of mouth. $C(I_n)(\mathbf{b}_c)$ is the value of a criterion for the processed image I_n and for the PCA parameter vector $\mathbf{b}_c$.

3.3.1. First image

The mouth corners being known, we now have to find lips and teeth contour and appearance for the image I_n.

In literature, the DSM is classically initialized on the mean shape and the search interval is $3\sqrt{\lambda_k}$ for each parameter $\mathbf{b}_c(k)$ ($1 \le k \le 5$). To reduce the number of iterations and the risk of convergence to wrong minima, we search the most plausible IMS for the image by computing $C_V(I_n)(\mathbf{b}_{ims,j})$, for $1 \le j \le 6$.

The $\mathbf{b}_{ims,j}$ which minimizes C_V is chosen as initial guess for the minimization of $C_c(I_n)(\mathbf{b}_c)$ by DSM. The search interval for each parameter will be deduced from the limit of parameters for this IMS as learned previously in 2.2. We note $\mathbf{b}_n$ the final parameters found for image I_n.

3.3.2. Tracking

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

$$\left|\frac{C_v\left(I_{n+1}\right)\left(\mathbf{b}_n\right) - C_v\left(I_n\right)\left(\mathbf{b}_n\right)}{C_v\left(I_n\right)\left(\mathbf{b}_n\right)}\right| \le 20\%$$

If this is verified, we assume that the mouth in the new image has practically the same localization and appearance that on the previous image. We will then minimize $C_V(I_{n+1})(\mathbf{b}_c)$ by DSM, with $\mathbf{b}_n$ as initial guess and a reduced search interval $0{,}5\sqrt{\lambda_k}$ for parameters.

If this is not verified, we assume that the new image is very different (and corresponds to another IMS) and we restart as in 3.3.1..

The DSM usually converges in 30 to 50 iterations and stops when the difference between the maximum and the minimum of the simplex is under a certain threshold.

3.4 Mouth interpolation

Figure 7: From left to right: initial image, interpolated image without and with teeth-mask

When the values of parameters are known, the sampled appearance is interpolated by a triangle-based interpolation method. The result is satisfactory for lips and skin but a little too blurry for teeth. To have a more realistic aspect for teeth, the operator chooses an image of the training set on which the teeth are visible and saves the appearance in a mask which will be used to fill the teeth area defined by the shape. To take care of shadows and lighting, the mean luminance of the mask has to be adjusted to the mean luminance of the teeth appearance samples. The result is then more realistic and understandable for lip-reading as shown in figure 7.

4. RESULTS AND CONCLUSION

error position	corner lips	outer lip contour	inner lip contour	upper teeth contour	lower teeth contour
mean error in pixel	1.8	1.2	1.5	1.8	1.6

Figure 8: Mean error localization.

Our method gives accurate lip segmentation with precise contour detection and convincing appearance reconstruction. The generated clone seems to be quite understandable by a human lip-reader. Figure 8 shows the mean errors for the shape points between detected and annotated points when the method runs on the images of the training set.

Figure 9: from left to right: initial image, clone with IMS and without

Learning and modeling the teeth contour associated to the sampled-appearance which defines precisely the localization of a sample allows our method to deal with the morphological variability of the mouth. The use of IMS reduces the number of iterations by about 40 percent. It also reduces the risk of convergence to wrong minima and unfaithful clone as in figure 9. Figure 10 shows some results.

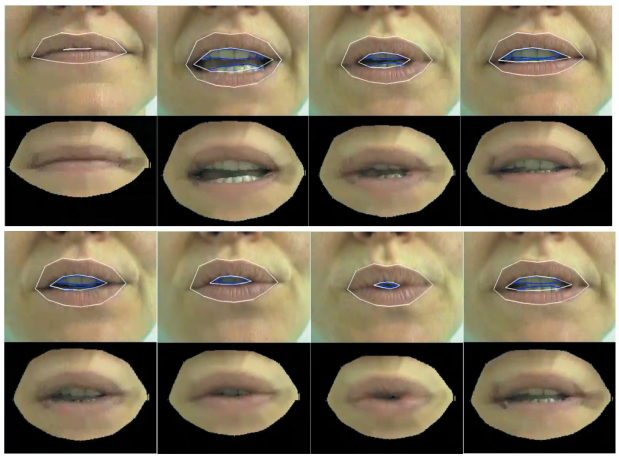

Figure 10: Mouths segmentation and interpolation, images with detected contours and generated clones

Our future works will deal with the robustness of the method for various lighting conditions and speakers. First encouraging results have already been obtained with a small training set of 6 different speakers (see figure 11). We also hope to develop an efficient semi-automatic application to annotate the images.

Finally a subjective evaluation is planned that will quantify the effective enhancement in comprehension brought by the analysis-resynthesis scheme in a telephone enquiry task.

Figure 11: Results for various speakers

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, "Jumping Snakes and Parametric Model for Lip Segmentation", *International Conference on Image Processing*, Barcelona, Spain, September 2003

[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] T. Poggio, and A. Hulbert, "Synthesizing a Color Algorithm From Examples", *Science*, Vol 239, pp 482-485, 1998.